Proyecto "Técnicas y herramientas para diseño lógico y mantenimiento de data warehouses relacionales"

Informe Técnico

Noviembre 2002

Toda la documentación del proyecto se encuentra accesible en forma electrónica usando la dirección:

http://www.fing.edu.uy/inco/grupos/csi/Proyectos/csic2000

En ella se encuentran todos los informes presentados así como los reportes técnicos y artículos generados en el marco del proyecto.

1 Resumen

Un elemento importante (aceptado desde mediados de la década del 90) en la infraestructura de un sistema de información para la toma decisiones de una organización de gran porte es lo que en ingles se conoce como *data warehouse* (DW). Un DW contiene datos e indicadores obtenidos de las bases de datos operacionales mediante procesos tales como la integración y el control de la calidad de datos y almacenados en estructuras que permiten, en particular, un acceso eficiente a las consultas OLAP (Online Analytical Processing) que satisfacen una parte de los requerimientos de toma de decisiones. Ejemplos típicos de requerimientos OLAP en el área de negocios lo constituyen la generación dinámica de reportes que calculan y hacen ranking de las ventas totalizadas por diferentes variables como pueden ser el cliente, el país y/o el año.

Considerando a un DW como una jerarquía de bases de datos [1], en este proyecto nos concentramos en la base de datos conocida como *DW corporativo*. Los DW corporativos son utilizados como depósitos intermedios para facilitar las tareas y separar el procesamiento entre las bases operacionales y las bases finales usadas por las herramientas orientadas al usuario final. Los DW corporativos habitualmente incluyen indicadores ya calculados para acelerar y simplificar la carga del próximo nivel en la jerarquía de bases de datos y sus datos se obtienen periódicamente a partir de las bases operacionales. Los DW corporativos relacionales (basados en el modelo relacional como modelo de datos) constituyen la opción más difundida.

Las *técnicas de gestión de bases de datos para DW corporativos* han evolucionado en forma significativa y rápida. Estas técnicas brindan principalmente funcionalidades de almacenamiento apropiado y procesamiento eficiente de consultas. Sin embargo, pocas han sido las *técnicas de diseño de la estructura de los DW corporativos* que han acompañado a estas técnicas de gestión. Las características de un DW corporativo relacional hacen que las técnicas de diseño a utilizar sean diferentes de aquellas utilizadas en el diseño de bases de datos relacionales. En particular deben permitir trabajar con redundancia en los datos y definir y mantener el DW corporativo a partir de bases de datos operacionales ya existentes.

El proyecto aborda problemas de diseño y mantenimiento de data warehouses corporativos relacionales. El objetivo es definir técnicas e incorporarlas en una herramienta de software para resolver principalmente los problemas de definición del esquema del DW corporativo y la gestión de

su evolución. El enfoque de la solución se basa en el uso de primitivas de transformación de esquemas. Estas operaciones permiten por un lado, representar conocimiento de criterios de diseño para llegar al esquema final a partir de los esquemas fuentes y por otro lado, proveer la traza del diseño realizado para ser utilizada en la gestión de la evolución.

En el resto del informe se usará *DW* como abreviación a *DW corporativo*. Asimismo se usará el término *fuentes* como abreviación de *bases de datos operacionales*.

2 Principales actividades desarrolladas

2.1 Primer año

Las principales actividades desarrolladas durante el primer año del proyecto se clasifican en especificas y de difusión.

Dentro de las actividades específicas se encuentran:

- Revisión bibliográfica de las técnicas y estrategias de diseño de data warehouses.
- Definición de las primitivas de transformación de esquemas para diseño de data warehouses relacionales.
- Estudio de características de las primitivas y clasificación de las mismas. Definición de la traza de primitivas.
- Definición de guías de uso de las primitivas definidas.
- Desarrollo de un primer prototipo de una herramienta de ayuda al diseño de DW basada en las primitivas incluyendo la traza de aplicación de las mismas.

Dentro de las actividades de difusión se destacan:

- Dictado del curso “Sistemas de data warehousing” de 60 hs. para el diploma de actualización profesional ofrecido por el Instituto de Computación de la Facultad de Ingenieria. Participaron en esta actividad los docentes Alejandro Gutiérrez, Adriana Marotta, Verónika Peralta y Raúl Ruggia.
- Dictado del curso “Sistemas de Data Warehousing y OLAP” de 14 hs. en la 8va Escuela de Verano de Cs. Informáticas (RIO 2001) organizada por la Universidad Nacional de Río Cuarto. Área de Computación. Córdoba, Argentina. en Río IV, Argentina. Participaron en esta actividad los docentes Adriana Marotta y Verónika Peralta.
- Trabajo de tres días con la profesora Ana Moura del Instituto Militar de Ingenieria (IME), Río de Janeiro, Brasil. Ana Moura fue invitada por el grupo Concepción de Sistemas de Información del Instituto de Computación para establecer lazos de cooperación. En dicha oportunidad pudimos intercambiar en particular el trabajo en torno al diseño de data warehouses.
- Puesta en marcha y participación de un seminario interno dentro del Instituto de Computación de la Facultad de Ingenieria sobre temas relacionados a la representación de los procesos de carga y refresque de data warehouses usando conceptos del área conocida como *workflow*.

2.2 Segundo año

Las actividades realizadas en el segundo año también se pueden clasificar en específicas y de difusión.

Dentro de las actividades específicas se encuentran:

- Estudio y definición de formas de incorporar técnicas de integración de esquemas en el mecanismo de diseño de DW basado en operaciones de transformación de esquemas.
- Estudio del problema de la repercusión de cambios en el DW frente a cambios en el esquema de las bases de datos operacionales.
- Dirección de un proyecto de grado de la carrera Ingeniero en Computación por Adriana Marotta y Verónika Peralta cuyo objetivo es la incorporación en el prototipo desarrollado en el primer año del mecanismo de repercusión de cambios en el esquema del DW frente a cambios en el esquema de las base de datos operacionales.
- Resolución de casos de estudio de mediano porte enfocados a la construcción de sistemas de data warehousing completos para identificar las etapas en donde los mecanismos de diseño y de mantenimiento de esquemas definidos en el proyecto pueden aplicarse. Se realizaron dos experiencias en esta línea. Un caso de estudio fue resuelto como parte del curso de Sistemas de Data Warehousing dictado en octubre 2001 [9]. El otro caso de estudio formó parte de un proyecto de grado de la carrera Ingeniero en Computación dirigido por Alejandro Gutiérrez y Regina Motz [8].
- Estudio del problema del diseño de la carga y actualización del DW y su vinculación con las operaciones de transformación de esquemas.

Dentro de las actividades de difusión se destacan:

- Dictado del curso “Sistemas de data warehousing” de 60 hs. para el diploma de actualización profesional ofrecido por el Instituto de Computación de la Facultad de Ingenieria. Participaron en esta actividad los docentes Alejandro Gutiérrez, Adriana Marotta, Verónika Peralta y Raúl Ruggia.
- Participación en el seminario interno dentro del Instituto de Computación de la Facultad de Ingenieria iniciado en el primer año sobre temas relacionados a la representación de los procesos de carga y refresque de data warehouses usando conceptos del área conocida como *workflow*.
- Presentación de Adriana Marotta e Ignacio Larrañaga en las VII Jornadas de Informática e Investigación Operativa organizadas por el InCo – Pedeciba Informática en diciembre de 2001.
- Mantenimiento de las páginas web del proyecto y del seminario interno. (Accesible con la dirección http://www.fing.edu.uy/inco/grupos/csi/esp/Proyectos/csic2000).

3 Principales resultados obtenidos

A lo largo de todo el proyecto se realizaron actividades que tenían como cometido difundir los trabajos relacionados con los temas del proyecto. Estas actividades permitieron realizar un relevamiento de trabajos existentes sobre técnicas de diseño de data warehouses [2] y presentaciones de trabajos existentes en el tema de carga y actualización de DW en el marco del seminario interno cuyo material se encuentra accesible desde la página del proyecto.

Con respecto al trabajo en el tema de diseño de esquemas de DW a partir de los esquemas de las fuentes se destacan los siguientes resultados.

- La definición de un conjunto de primitivas de transformación de esquemas que permiten representar estrategias para construir data warehouses relacionales. Se estudiaron características de las primitivas definidas y se realizó una clasificación de las primitivas según diferentes criterios que permiten facilitar su uso por parte de un diseñador de DW [3], [10].
- El desarrollo de un primer prototipo consistente en un ambiente gráfico de ayuda al diseño de DW relacionales que incluye la implementación de las primitivas y provee mecanismos para aplicar reglas de consistencia que aseguran invariantes sobre los esquemas de DW [4].
- La propuesta de un mecanismo para resolver el diseño de DW a partir de múltiples bases fuentes, el cual adapta la propuesta de [3]. En este mecanismo se parte de un esquema de DW objetivo, se establecen correspondencias semánticas entre este esquema y los esquemas fuentes, y luego se aplican las primitivas de transformación de esquemas. Se adapta el conjunto de primitivas de [3] de forma de poder resolver los problemas de integración [6].

Con respecto al trabajo en el tema de repercusión de cambios en el DW frente a cambios en las fuentes se destacan los siguientes resultados.

- Una propuesta para resolver la propagación de los cambios ocurridos en el esquema fuente hacia el esquema del DW, y gestionar la evolución en el DW. En ella se especifican: 1- dependencias entre elementos del esquema fuente y elementos del esquema de DW, 2- una taxonomía de cambios sobre el esquema fuente, 3- reglas de propagación de los cambios. También se trabaja sobre la aplicación de la evolución al DW analizando los posibles modelos a seguir (versiones o adaptativo), y teniendo en cuenta la adaptación de las instancias [11].
- La construcción de un prototipo de una herramienta para la evolución de DW de acuerdo a las especificaciones de [11], e implementado como una extensión de [4]. Esta herramienta permite a un diseñador de DW aplicar cambios sobre el esquema fuente y propagarlos al esquema del DW, actualizándose automáticamente la traza del proceso de diseño. [7]
- La proposición de una arquitectura de sistema de información que permita incorporar 2 categorías de técnicas de propagación de cambios a un data warehouse con información extraída de la web: (1) propagación de cambios en esquemas fuentes y (2) propagación de agregado o eliminación de fuentes [5].

En lo referente al tema de diseño de la carga y actualización de los datos del DW y su vinculación con las operaciones de transformación de esquemas se destacan como resultados interesantes la experiencia en el uso de una herramienta de paralelización aplicada para resolver la carga asociada a una de las operaciones de transformación de esquemas y un trabajo aún en curso que propone generar un proceso de carga inicial basado en información del esquema del DW generado por las operaciones de transformación de esquemas.

- La aplicación de paralelismo en los procesos de carga es conocido que puede dar beneficios interesantes, como por ejemplo la reducción de tiempos. El trabajo descripto en [12] analizó la dificultad y beneficios de este tipo de propuestas, utilizando una herramienta genérica y conocida en el área (PVM) para implementar la carga de datos sobre un esquema de DW obtenido por la apliacación reiterada de la operación de agregación. Si bien el trabajo fue muy puntual permitió introducirse en esta clase de técnicas y plantearse líneas de trabajo.
- El diseño de procesos de carga y actualización es una tarea tediosa y complicada, generalmente con muchos aspectos comunes o deducibles a partir de información existente. El trabajo iniciado en [13] presenta un ejemplo concreto en el cual se toma la información existente del esquema, y se elabora un proceso de carga que puede ser tomado como base por el diseñador. La intención es seguir trabajando en esta área, aplicando otros conceptos como el presentado e incorporando ideas

presentes en trabajos relacionados con la retoma de un proceso de carga suspendido y limpieza de datos.

El proyecto obtuvo como resultados importantes la formación de sus participantes y contactos regionales, de los cuales se destacan.

- La finalización por parte de Adriana Marotta de su tesis de maestria del Pedeciba titulada “Diseño y mantenimiento de Data Warehouses a través de transformaciones de esquemas”. Su trabajo fue fundamental en el proyecto ofreciendo las principales especificaciones en el tema de diseño y mantenimiento de esquemas de DWs relacionales que sirvieron para llevar a cabo los prototipos.
- La presentación del trabajo [5] en el Workshop realizado en Brasil permitió establecer el contacto con la profesora Ana Moura del Instituto Militar de Ingenieria (IME), Río de Janeiro, Brasil quien trabaja en temas relacionados al presente proyecto y en general a aquellos temas del grupo al que trabajan los participantes de este proyecto (Concepción de Sistemas de Información del Instituto de Computación).

4 Evaluación de los resultados obtenidos

Con respecto al tema de diseño de esquemas de DW y la repercusión de cambios en el esquema frente a cambios en los esquemas fuentes, nuestra evaluación es muy positiva. Se lograron resultados que están muy en la dirección de los objetivos planteados. Cabe señalar que para ello fue necesario tomar una decisión cuando trabajamos en la construcción de una especificación formal de las primitivas de diseño. El objetivo principal de esta actividad en el contexto del proyecto era contar con un tiempo para encontrar una forma de expresar con precisión la funcionalidad de las primitivas. Estudiamos dos posibilidades en ese sentido. Por un lado, el uso de un lenguaje de especificación como *B* y por otro lado el uso de un prototipo. La primera posibilidad la evaluamos demasiado larga para este proyecto previendo retrasos en el resto de las actividades. Por esta razón se siguió a lo largo de todo el proyecto el camino de precisar nuestras proposiciones mediante el desarrollo de prototipos.

Con respecto al tema del diseño de la carga y actualización de datos del DW vinculados con las operaciones de transformación de esquemas nuestra evaluación es también positiva si bien los resultados obtenidos son parciales con respecto a los objetivos planteados. Observamos que se trata de un tema muy vasto en cuanto a las técnicas que pueden aplicarse. Los trabajos realizados en el proyecto apuntaron principalmente a realizar presentaciones en el marco del seminario para comprender las técnicas actualmente utilizadas y experimentar con algunas técnicas de paralelización aplicada a casos particulares de diseño de DW usando las operaciones de transformación de esquemas. Consideramos que el proyecto ha sido fundamental para comenzar a tratar el tema y permitió obtener una primera formación en el tema que continuaremos.

En lo que se refiere al desarrollo del proyecto cabe señalar que uno de los principales obstáculos enfrentados durante el segundo año fue la dificultad en disponer de tiempo suficiente para dedicarse a tareas de investigación. En este sentido, se destaca la importancia que tuvo la puesta en marcha del seminario interno que permitió dinamizar las actividades y comunicar los trabajos en curso.

Por último, evaluamos como muy positivo las formas logradas de difusión de los temas manejados en el marco del proyecto variando desde presentaciones en un seminario interno, la escritura de reportes técnicos, escritura y presentación de artículos en jornadas locales y en conferencias regionales e internacionales, dictado de cursos locales de actualización, dictado de un curso en una escuela de verano regional e información pública disponible via internet.

5 Perspectivas

El presente proyecto se concentró principalmente en ofrecer un mecanismo de diseño y mantenimiento del esquema del DW corporativo relacional a partir de los esquemas de las bases operacionales. Si bien se proveen guías para el uso de las operaciones de transformación de esquemas, la elección de las operaciones a aplicar es realizada manualmente por el diseñador. El uso de las mismas en los casos de estudio nos permitió observar que un diseño primario del esquema relacional del DW podría ser automatizado partiendo de un modelo conceptual multidimensional originado a partir de los requerimientos del usuario complementado con directivas declarativas sobre el diseño esperado. Esta línea de trabajo permitió la proposición de un nuevo proyecto que fue presentado ante la CSIC y de un tema de tesis de maestría para Verónika Peralta.

Por otro lado, el proyecto permitió a Ignacio Larrañaga interiorizarse en los temas referentes a la carga y actualización del DW brindándole un área posible para su tesis de maestría.

6 Referencias

[1] M. Bouzeghoub, F. Fabret, M. Matulovic-Broqué. *Modeling Data Warehouse Refreshment Process as a Workflow Application*. Proc. DMDW 1999. (Disponible: http://www.dbnet.ece.ntua.gr/~dwq/publications.html)

[2] A. Gutiérrez, A. Marotta, *An Overview of Data Warehouse Design Approaches and Techniques.* Reporte Técnico INCO-01-09. InCo - Pedeciba, Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay. Octubre 2000. ISSN 0797-6410.

[3] A. Marotta, *Designing Relational Data Warehouses through Schema-Transformation Primitives.* Reporte Técnico INCO-01-10. InCo - Pedeciba, Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay. Diciembre 2000. ISSN 0797-6410.

[4] A. Gutiérrez, A. Marotta, *Designing Relational Data Warehouses through Schema-Transformation Primitives - A Prototype.* Reporte Técnico INCO-01-11. InCo - Pedeciba, Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay. Junio 2001. ISSN 0797-6410.

[5] A. Marotta, R. Motz, R. Ruggia, *Managing Source Schema Evolution in Web Warehouses.* International Workshop on Information Integration on the Web, WIIW '2001. Brazil. Abril 2001.

[6] A. Marotta. *Resolución de la integración en el diseño del Data Warehouse.* Reporte Técnico INCO-02-07. InCo - Pedeciba, Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay. Junio 2001. ISSN 0797-6410.

[7] A. Alcarraz, M. Ayala, P. Gatto. *Diseño e implementación de una herramienta para la Evolución de un Data Warehouse Relacional*. Informe final del grado Ingeniero en Computación. Supervisores: Adriana Marotta, Verónika Peralta. In.Co., Facultad de Ingeniería. Universidad de la República. Montevideo, Uruguay. Junio 2001.

[8] A. Gutiérrez, R. Motz, B. Revello, L. Silva. *Construcción de un sistema de apoyo a la toma de decisiones para el área gerencial del Hospital de Clínicas*. Anales 30o. JAIIO, Subserie: Simposio Argentino de Informática y Salud (SIS), Vol. 4., páginas 232 - 242, Setiembre 2001, Buenos Aires, Argentina. También como reporte técnico INCO-01-07. InCo - Pedeciba, Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay. Julio 2001. ISSN 0797-6410.

[9] A. Caorsi, H. Paggi, G. Perez. *Caso de Estudio: Arte Espectacular*. Informe para la evaluación del curso “Sistemas de Data Warehousing” del Diploma de Actualización Profesional. Mayo 2002.

[10] A. Marotta, R.Ruggia. *Data Warehouse Design: A Schema Transformation Approach.* XXII Conferencia Internacional de la Sociedad Chilena de Ciencia de la Computación. Chile 2002.

[11] A. Marotta. *Managing source schema evolution in relational data warehouses.* Reporte técnico en elaboración.

[12] I. Larrañaga. *Aplicación de PVM a la carga de datos para el análisis OLAP*. Reporte técnico en elaboración.

[13] I. Larrañaga. *Automatic initial load of data warehouses as a workflow process.* Reporte interno. Julio 2002.