Implementación de un Modelo Multidimensional para un Datawarehouse sobre pacientes diabéticos

E. Mangia[1], D. Omar[1], M. E. Llorente[1], J. Besso[1], A. Sigura[1,2], A. J. Hadad[1,2], B. Drozdowicz[1,2]
[1]Facultad Ciencia y Tecnología, Universidad Autónoma de Entre Ríos
[2]Facultad Ingeniería, Universidad Nacional de Entre Ríos
Ruta 11, Oro Verde, Entre Ríos, Argentina
mellorente@arnet.com.ar, bdrozdo@santafe-conicet.gov.ar

Resumen

En el presente trabajo se propone la implementación inicial de un Datawarehouse a partir del Modelo Multidimensional propuesto en etapas previas del proyecto de investigación que da contexto al mismo.
Se incluye la implementación de las estructuras de datos, correspondientes a las fuentes de información seleccionadas como entrada al proceso de transformación de información, que debe ser incluida en el Datawarehouse. También se describe la implementación de las estructuras de datos que forman parte del Modelo Multidimensional definido y finalmente la definición y aplicación de los proceso de Extracción, Transformación y Carga (ETL), necesario para la registración de información en el DW para pacientes diabéticos.
Para la elaboración del prototipo, se utiliza la herramienta IBM InfoSphere que permite la definición de estructuras de datos de entrada y salida, y la definición de procesos de transformación (ETL) necesarios.

Palabras clave: Datawarehouse, Pacientes Diabéticos, Estructuras de Datos, Modelo Multidimensional, ETL, Prototipo.

Contexto

El presente trabajo se inserta en un Proyecto de Investigación Plurianual (PIDP) denominado "*Sistema de Soporte a la Toma de Decisiones basado en datawarehouse para pacientes diabéticos*". Dicho proyecto es desarrollado en la Facultad de Ciencia y Tecnología de la Universidad Autónoma de Entre Ríos (FCYT - UADER).

Introducción

En la última década se han reportado diferentes enfoques en lo que refiere a metodologías de diseño de datawarehouse en el ámbito de la salud. Por citar algunos ejemplos en [1] se destacan las dificultades en los pasos para la captura de datos a nivel organizacional en relación al problema del monitoreo de enfermedades infecciosas en un ambiente clásico. En [2] se hace hincapié • en el aspecto temporal de largo plazo en relación a los requerimientos y en el rol de los sistemas de información clásicos operacionales con los módulos ETL en el proceso de diseño. En [3] se remarca el rol de las etapas de extracción de requerimiento, diseño del modelo de datos, implementación y despliegue, profundizando finalmente en los aspectos de arquitectura, al que le da un rol más relevante para contextos clásicos. En [4] se destaca del rol del proceso de integración de la historia clásica electrónica presente en los sistemas operacionales con un sistema de información basado en datawarehouse. En lo que refiere a experiencias más especifica relacionada con el diseño de

orientado a la gestión de información relacionada con pacientes diabéticos. En [5] se presenta una interesante caracterización del contexto para diabetes y remarca la dinámica con que se presentan los requerimientos, lo cual tiene un impacto importante en el proceso de diseño. Por otro lado, recientemente, se ha reportado [6] el desarrollo de un sistema basado en datawarehouse para diabetes en el cual se enfoca el diseño centrado en los datos del paciente en lugar de abordarlos desde la perspectiva de las enfermedades y tratamientos.

Este último enfoque tiene un impacto marcado en la implementación de este tipo de sistemas generando modelos cuya granularidad mínima se define a nivel de paciente

En este trabajo, la implementación de un DW, incluye la definición de procesos que transformen la información proveniente de las fuentes de entrada y la carga de la misma en las estructuras de información definidas de acuerdo al modelo multidimensional propuesto.

Esta transformación requiere de un trabajo de análisis para la depuración, limpieza y codificación, que permita una correcta carga de información.

Líneas de investigación y desarrollo

De las líneas de investigación descriptas para este proyecto en el trabajo presentado en el WICC 2012 [1], en esta propuesta se analizaron las siguientes líneas:

1. Estructuras de datos representativas del dominio de análisis.
2. Métodos ETL para fuentes de información de referencia.

Resultados y Objetivos

Para la implementación de un prototipo del DW para pacientes diabéticos, se ha seleccionado la herramienta IBM InfoSphere la cual facilita la implementación de un prototipo que refleje este proceso. Esta plataforma provee un gran conjunto de herramientas para brindar soporte a la extracción, limpieza, transformación, carga, almacén y análisis de datos, entre otros. También, permite extraer datos de diferentes fuentes u orígenes: base de datos relacionales, archivos de texto, extracción desde planillas de cálculo, archivos separados por punto y coma (.csv), servicios web, archivos XML.

El Modelo Multidimensional del DW para enfermos diabéticos que es utilizado para la implementación del prototipo es el representado en la Fig. 1.

Como fuente de información, para comenzar a definir el modelo multidimensional, se decide utilizar el modelo de datos de Historias Clínicas propuesto en un caso de estudio [2]. La elección de esta fuente de información desde una publicación se basa en que la misma tiene planteos similares, en este aspecto, a los planteados en el proyecto de investigación y que aún no resultó posible lograr una información equivalente de ámbitos médicos locales.

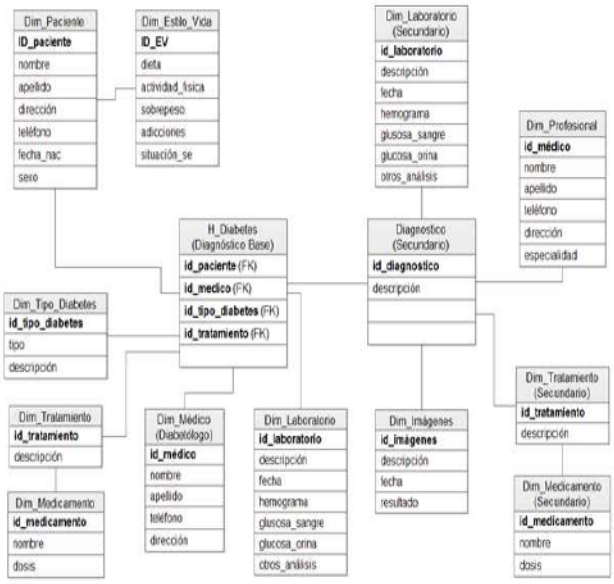

Figura 1. Primer Modelo Multidimensional propuesto.

En primera instancia se realizó el análisis del

modelo de Historia Clínica de Pacientes Diabéticos como Fuente de Información (FI) y del Modelo Multidimensional como DW destino.
Del anterior análisis se determinó la correspondencia entre las entidades del primer modelo con las del segundo. A partir de ello se especificó el mapeo de campos entre dichas entidades, detallando de cada uno la transformación necesaria a aplicar para cargar el DW.

Teniendo en cuenta ambos modelos y el análisis realizado, se propusieron dos casos de procesos ETL:

- Cargar datos de pacientes. Este proceso es simple porque los datos no requieren de ninguna transformación compleja. Se puede decir entonces, que la carga desde la FI hacia la DW se realiza de forma directa.

- Cargar estudios de pacientes. A diferencia del anterior, este proceso requiere de una transformación, ya que debe calcularse la variación del valor de un estudio de un paciente con respecto al valor anterior.

La FI propuesta se encuentra soportada en el motor de base de datos DB2, también proporcionado por IBM. El modelo multidimensional actualmente es simulado utilizando DB2 como una BD relacional. Será trabajo futuro implementarlo como DW en la herramienta específica para tal fin, provista por InfoSphere Platform.

- Origen: BD Historia Clínica (HC)
- Destino: DW

A continuación se especifica el mapeo para el caso de Pacientes.

Mapeo: Dim_Paciente y Patient

La dimensión Dim_Paciente se relaciona con la entidad Patient. El mapeo de sus campos se grafica en la Tabla 1.

Patient	Dim_Paciente	Comentario
P_Name	nombre apellido	El campo P_Name debe desdoblarse en el destino en 2 atributos. Esto será posible si dicho campo almacena texto con una estructura definida. (Ej. Nombre, Apellido)
P_Address	dirección	
P_Sex	sexo	
P_Birthdate	fecha_nac	
	teléfono	

Tabla 1 – Mapeo de la dimensión Paciente

A continuación se detalla el proceso realizado para la implementación de la ETL de Pacientes:

Pre-requisitos:

- Se dispone de la herramienta instalada en un sistema operativo compatible.
- Se dispone del hardware mínimo requerido por la herramienta.
- Se realizaron las configuraciones pertinentes tanto del sistema operativo como de la herramienta.

Se creó un proyecto de tipo ANALIZERPROJECT en "DataStage Administrator". Este proyecto contendrá todas las definiciones de las fuentes de información, trabajos (procesos ETL), metadatos, entre otros.
Se creó un Nombre de Origen de Datos (DSN) para la base de datos de Historia Clínica y otro para el DW (utilizando DB2 como motor de base de datos).
A continuación se crearon en "DataStage Designer" los Data Connection Object (DCO) correspondientes, utilizados para vincular los DSN definidos anteriormente.
Se importó la definición de la FI (metadatos), es decir las tablas de la BD relacional de HC a través del DCO creado. Además se importaron los metadatos correspondientes a las tablas del

DW.
Se diseñó la ETL, utilizando el módulo "DataStage Designer". Se crearon dos objetos "ODBC Connector" para la BD de HC y para la DW respectivamente. Estos se unieron a través de un proceso "Transformer" utilizado para transformar los datos de origen y luego almacenarlos en el destino. Dicho proceso utiliza la función Split para dividir el campo **P_NAME** del origen en **NOMBRE** y **APELLIDO** en el destino.
Por último, se ejecutó la ETL diseñada, con un total de 100 registros de la tabla **PATIENT**, obteniendo como resultado 100 registros en el destino, con la transformación mencionada anteriormente.
Los resultados obtenidos se muestran a través de las imágenes capturadas con el uso de la herramienta propuesta (Fig. 2 y 3).

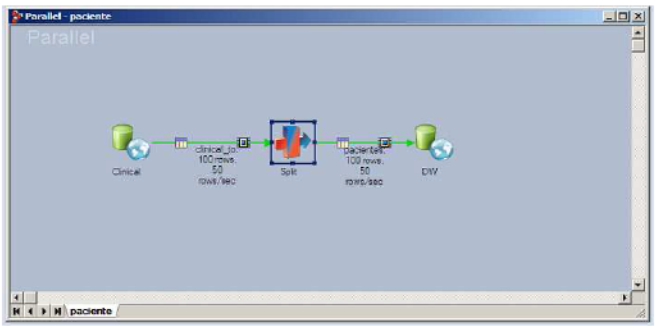

Figura 2 - ETL Paciente ejecutada exitosamente

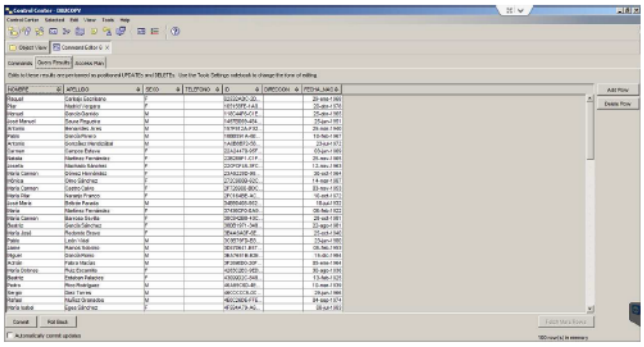

Figura 3 - Tabla Dim_Paciente

Formación de Recursos Humanos

El equipo de trabajo está conformado por especialistas del área de sistemas de información, inteligencia artificial y bioingeniería. Integrantes del equipo tienen formación de postgrado tanto en el área de sistemas de información como en el área biomédica, así como también experiencia en el ámbito profesional en lo que refiere al desarrollo de sistemas.
Se incorporan tres alumnos becarios de la carrera Licenciatura en Sistema de la Facultad de Ciencia y Tecnología de la UADER

Referencias

[1]Wisniewski et al. Development of a Clinical Data Warehouse for Hospital Infection Control. Journal of the American Medical Informatics Association Volume 10 Number 5 Sep / Oct 2003. 10:454–462. DOI 10.1197/jamia.M1299.

[2] Xuezhong Zhouemail, Shibo Chenemail, Baoyan Liu, Runsun Zhang, Yinghui Wang, Ping Li, Yufeng Guo, Hua Zhang, Zhuye Gao, Xiufeng Yan. Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support. Artificial Intelligence in Medicine February–March, 2010, Volume 48, Issues 2-3, Pages 139–152 DOI: http://dx.doi.org/10.1016/j.artmed.2009.07.012

[3] Sahama, Tony R. and Croll, Peter R. (2007) A Data Warehouse Architecture for Clinical Data Warehousing . In Roddick, J. F. and Warren, J. R., Eds. Proceedings Australasian

Workshop on Health Knowledge Management and Discovery (HKMD 2007) CRPIT, 68, pages pp. 227-232, Ballarat, Victoria.

[4] Craig S. Ledbetter, Matthew W. Morgan. Toward Best Practice: Leveraging the Electronic Patient Record as a Clinical Data Warehouse JOURNAL OF HEALTHCARE INFORMATION MANAGEMENT®, vol. 15, no. 2, Summer 2001
© Healthcare Information Management Systems Society and Jossey-Bass, A Publishing Unit of John Wiley & Sons, Inc.

[5] Torben Bach Pedersen,Christian S. Jensen. Research Issues in Clinical Data Warehousing. 8 IEEE. Published in the Proceedings of SSDBM'98, July 1-3 1998 in Capri, Italy

[6] Lichin Chen, Chiou-Shiang Wang, I-Ching Wang, Hui-Chu Yu, Hui-Yu Peng, Hui-Chuen Chen, Chia-Hsiun Chang, Tien-Jyun Chang, Yi-Der Jiang, Lee-Ming Chuang, Feipei Lai. Patient-centric Data Warehouse Design An Empirical Study Applied in Diabetes care. BIOTECHNO 2014 : The Sixth International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies. Copyright (c) IARIA, 2014. ISBN: 978-1-61208-335-3

[7] "Sistema de soporte a la toma de decisiones basado en datawarehouse para pacientes diabéticos." M. E. Llorente, Aldo Daniel Sigura, Alejandro Hadad, Bartolomé Drozdowicz. WICC 2012

[8] "Design and development of a web-based application for diabetes patient data management". Deo SS, Deobagkar DN, Deobagkar DD. Informatics in Primary Care (2005) 13(1):35-41