Propuesta de un Modelo Multidimensional para un Datawarehouse sobre pacientes diabéticos

M. E. Llorente[1], A. Sigura[1], J. Besso[1], E. Mangia[1], A. J. Hadad[1,2], M. Mancini[1], N. Quijada[1], 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

Se sabe que la tecnología Datawarehousing debido a su orientación analítica, impone un procesamiento y pensamiento distinto, la cual se sustenta por un modelado de Bases de Datos propio, conocido como Modelado Multidimensional. Su objetivo es ofrecer al usuario información integral que le permita visualizar la operación del negocio.
Se propone en este trabajo formalizar un modelo lógico de la solución, planteando una primera definición de las posibles tablas del modelo Multidimensional con sus atributos y dimensiones, definiendo una primera versión del modelo de un DW para pacientes diabéticos. Para su elaboración se definen cuales son los elementos necesarios para el logro del objetivo planteado, y para el tema en estudio, el criterio de selección de los mismos. Los resultados obtenidos están basados en el análisis de los requerimientos a satisfacer y la fuente de información seleccionada.

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

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 el modelo multidimensional cada eje corresponde a una dimensión particular.
Entonces la dimensionalidad de la base a proponer está dada por la cantidad de ejes (o dimensiones) que se le asocie. Cuando una base puede ser visualizada como un cubo de tres o más dimensiones, es más fácil para el usuario organizar la información e imaginarse en ella cortando y rebanando el cubo a través de cada una de sus dimensiones, para buscar la información deseada.
Para construir un DW se debe primero tener claro que existe una diferencia entre la estructura de la información y su semántica y que esta última es mucho más difícil de abarcar. En esto se encuentra la principal diferencia entre los sistemas operacionales y el DW.
En lo que refiere al modelado de los datos para la problemática dentro de ámbito médico, requiere el preprocesamiento de datos a fin de contextualizar su contenido informativo y facilitar su interpretación para un proceso de toma de decisiones. Dicho preprocesamiento puede involucrar la aplicación de técnicas para abstracción temporal, generación de índices,

extracción de características relevantes, identificación de estados, etc. [1-6].

Líneas de investigación y desarrollo

De las líneas de investigación descriptas para este proyecto en el trabajo presentado en el WICC 2012 [7], en esta propuesta se analizaron las siguientes líneas:
1. Estructuras de datos representativas del dominio de análisis.
2. Granularidad y dimensiones del modelo.

Resultados y Objetivos

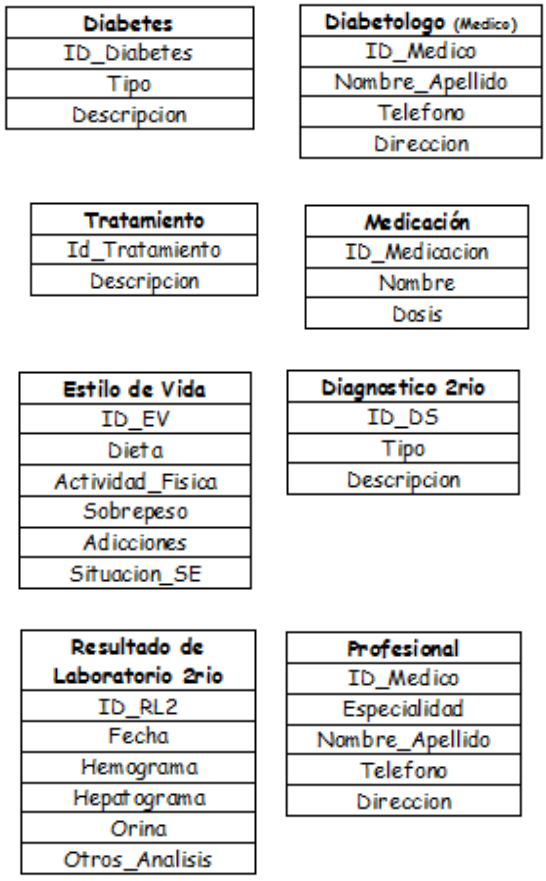
Para el desarrollo del modelo se tuvo en cuenta la evolución en las metodologías de diseño de los Modelos Multidimensionales, desde el año 1998 hasta la actualidad [8]. Analizadas las propuestas de diferentes autores de referencia, es posible definir que las metodologías están basadas en dos aspectos fundamentales:

- Requerimientos: basados en Casos de Uso, dentro de un proceso iterativo, su desarrollo se describe en trabajos presentados en congresos anteriores [9-10].

- Fuentes de Información existentes: se decidió utilizar el modelo de datos de Historias Clínicas del caso de estudio presentado en [11].

La elección de la fuente de información desde una publicación se basa en que la misma tiene planteos similares, en este aspecto, al proyecto propuesto y que aún no resultó posible lograr una información equivalente de ámbitos médicos locales.

La creación de un modelo lógico de DW incluye estructuras temporales vinculadas a los niveles de las jerarquías dimensionales, que posibilitan registrar los datos y recuperar la información variante en el tiempo, incorporando nociones de almacenamiento, performance y estructuración de los datos. Además se debe tener en cuenta un componente adicional que son las bases de datos fuentes. Durante el diseño se analizó la correspondencia entre el modelo lógico con las tablas y atributos de las bases fuentes y los elementos del esquema conceptual
A continuación se analizó la estructura del modelo elegido y se extrajo la información para poder plantear diferentes ideas de atributos para el Modelo Multidimensional. En la Fig. 1 se muestra la primera definición de las posibles entidades relevantes representadas en tablas, con sus atributos y su clave primaria.



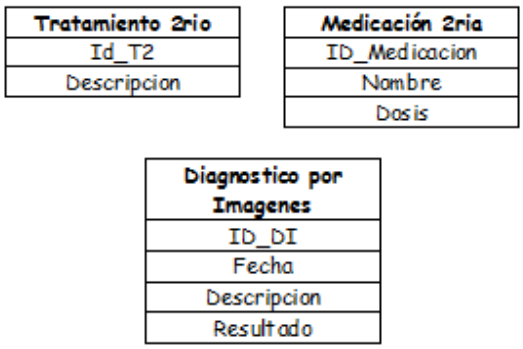


Figura 1 Entidades y atributos propuestos

A partir de dicha propuesta, se analizaron las relaciones existentes entre las entidades definidas y la relevancia de cada una de ellas. Luego se estudió el aporte de cada una a la resolución de los requerimientos definidos, con el objetivo de proponer un Modelo Multidimensional y establecer la correspondencia entre los requerimientos y los datos fuentes. Del análisis realizado, se definieron las dimensiones que muestra la Fig. 2.

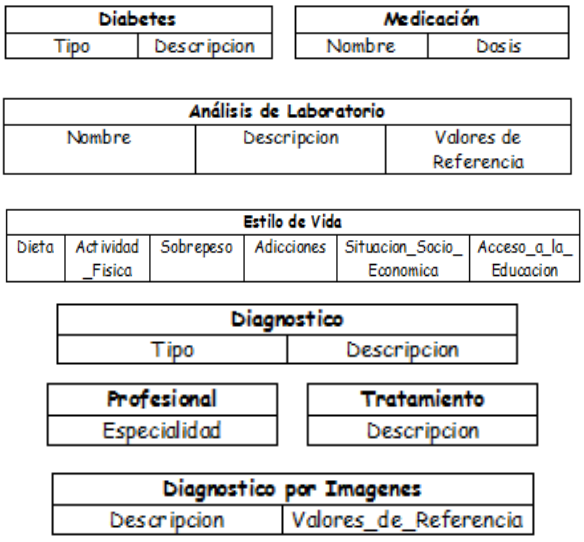


Figura 2. Dimensiones del Modelo.

Por otro lado, se definieron las siguientes dos tablas de hechos, centrándose en la problemática propuesta:

- Diagnóstico base: orientado a registrar la evolución de la enfermedad en estudio (diabetes).

- Diagnóstico secundario: se basa en registrar eventos clínicos relacionados o no con la enfermedad base.

Esta conclusión está fundamentada en el hecho de entender que existen eventos que originan diagnósticos secundarios, que pueden o no estar relacionados con el diagnóstico base que es el que nos interesa. Por otro lado, el diagnóstico base tiene evolución propia, por lo que interesa realizar esta discriminación diagnóstica.
Por último se realizaron las uniones pertinentes entre las tablas, definiendo las claves primarias y foráneas. De esta forma se obtuvo el primer Modelo Multidimensional propuesto por el grupo de investigación, Fig. 3.

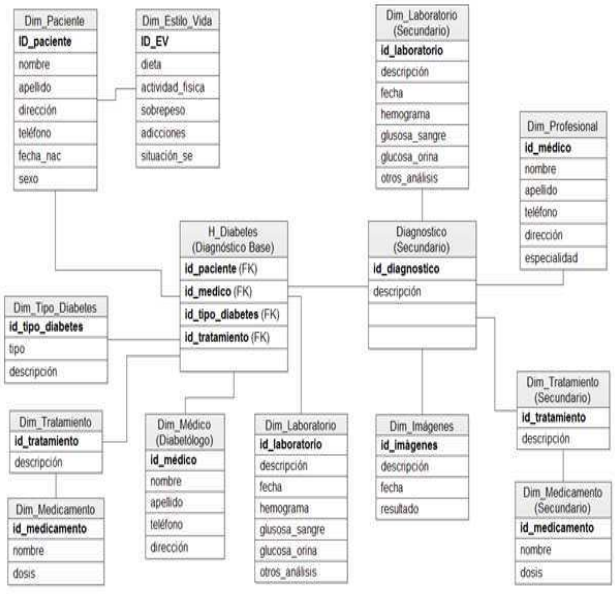


Figura 3. Modelo Multidimensional propuesto.

El modelo planteado podrá sufrir modificaciones de acuerdo a los nuevos casos de uso, que se definan en las próximas iteraciones. Sin embargo, a partir de la fuente

de información propuesta y el modelo planteado, se está en condiciones de definir los procesos ETL que permitan cargar la información transformada al DW.
Para llevar a cabo la implementación de la DW, se consideró importante contar con una herramienta que permita analizar resultados concretos, en el proceso de diseño del modelo de datos y procesos ETL.
Luego de haber analizado y comparado diferentes herramientas, se decide utilizar InfoSphere Platform de IBM. Uno de los paquetes que brinda esta herramienta es Information Server, el cual da **soporte al proceso ETL**. Se pueden crear, diseñar y administrar las ETL necesarias. 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.
Por otro lado, para soportar el diseño del modelo de datos multidimensional, la herramienta brinda la aplicación **InfoSphere Data Architect** [12]. Es decir, con ella se puede crear el modelo multidimensional planteado en el presente trabajo.

Formación de Recursos Humanos

El equipo de trabajo está conformado por especialistas del área informática y de 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] “Temporal Abstraction for the Analysis of Intensive Care Information” A. Hadad, D. Evin, B. Drozdowicz, O. Chiotti. Journal of Physics: Conference Series. Volume 90, 2007. ISSN: 1742-6596

[2] “A Comparative Analysis of Preprocessing Techniques in Colour Retinal Images”,Adrián Salvatelli, Gustavo Bizai, Gisela Barbosa, Bartolomé Drozdowicz, Claudio Delrieux. Journal of Physics-Conferences Series (JPCS). Noviembre de 2007.

[3] “Implementación y aplicación de algoritmos Retinex al preprocesamiento de imágenes de retinografía color”. Autores: N. Londoño, G. Bizai, B. Drozdowicz. Revista Ingeniería Biomédica. Volumen 3. Paginas 36-43. ISSN 1909 – 9762.

[4] “Modelos de seguimiento para la supervisión de procesos complejos en aplicaciones biomédicas”. Autor: Alejandro Hadad. Encuentro Internacional de Investigación en Ingeniería de Sistemas e Informática. Tunja, Colombia, 6 al 8 de Octubre de 2010.

[5] “Predicción de estados de hipotensión empleando Modelos Ocultos de Markov". Autores: Diego Evin, Alejandro Hadad, Mauro Martina, Bartolomé Drozdowicz. Encuentro Internacional de Investigación en Ingeniería de Sistemas e Informática. Tunja, Colombia, 6 al 8 de Octubre de 2010.

[6] “Prototipo para la comparación de patrones temporales secuenciales de arritmias cardíacas”. Autores Hadad, Alejandro Javier; Solano, Agustin Ezequiel y Drozdowicz, Bartolomé En: Ventana Informática No. 26 (ene.-jun., 2012). Manizales (Colombia): Facultad de Ciencias e Ingeniería, Universidad de Manizales. pp 29-43 ISSN: 0123-9678

[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] “A survey of Multidimensional Modeling Methodologies”. Oscar Romero, Alberto Abelló. International Journal of Data Warehousing & Mining, 5(2), 1-23, April-June 2009

[9] “Proceso de Diseño basado en Casos de Uso para un Datawarehouse Clínico”. M. E. Llorente, Aldo Daniel Sigura, Javier Besso, Alejandro Hadad, Bartolomé Drozdowicz. CACIC 2012

[10] “Análisis de fuentes de información para el proceso de diseño de un datawarehouse sobre pacientes diabéticos”, M.E.Llorente, Aldo Daniel Sigura, Alejandro Hadad, Bartolomé Drozdowicz. WICC 2013

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

[12] Sitio Oficial de IBM Infosphere DataStage: http://www-01.ibm.com/software/data/infosphere/datastage/