UNIVERSIDAD NACIONAL DE COLOMBIA
SEDE MEDELLÍN
FACULTAD DE MINAS

DYNA
http://dyna.medellin.unal.edu.co/

The conceptual modeling in the process of computer-assisted generation of data warehouse models

Lindsay Alonso Gómez-Beltrán [a], Rosendo Moreno-Rodríguez [b] & Ramiro Pérez-Vázquez [c]

[a] *Dirección de Informatización, Universidad Camagüey, Cuba . lindsay.gomez@reduc.edu.cu*
[b] *Departamento de Física, Matemática y Computación, Universidad Central "Marta Abreu" de las Villas, Cuba. rosendo@uclv.edu.cu*
[c] *Departamento de Física, Matemática y Computación, Universidad Central "Marta Abreu" de las Villas, Cuba. rperez@uclv.edu.cu*

Received: November 23th, de 2013. Received in revised form: May 5th, 2014. Accepted: May 22th, 2014

Abstract
This paper introduces the methodological guidelines for the data warehouse model computer assisted generation. These guidelines are divided into four different stages: information analysis, conceptual model and logical design are the first ones and the last one occurs within them and it is known as the traceability stage. These stages describe a data warehouse design proposal can be obtained from the inherited operational systems (E/R). One of the main stages we considered to be important is the data warehouse conceptual model.
This paper goes deeper into the different ways to obtain the conceptual model from the logical structure of the institutional inherited systems, taking into account that these systems generally use a relational model in its structure. In order to accomplish this, it is proposed to use the interrelation among entities to generate a graph of interrelation of attribute and then apply a set of design rules to obtain the conceptual model.

Keywords: Data warehouse design, Data warehouse model, conceptual model.

El modelado conceptual en el proceso de generación asistida por computadoras de modelos de almacenes de datos

Resumen
En este trabajo se presentan las pautas metodológicas para la generación asistida por computadoras de modelos de almacenes de datos (AD), estas pautas metodológicas se dividen en 4 etapas, las 3 etapas primeras son: análisis de la información, modelo conceptual y diseño lógico y una última que ocurre dentro de cada una de las etapas anteriores que es denominada etapa de trazabilidad. Estas etapas describen cómo podemos obtener a partir de los sistemas operacionales heredados (E/R) una propuesta de modelado de almacenes de datos. Una de las etapas que se considera de mayor importancia es el modelo conceptual del almacén de dato, en este trabajo se profundiza en la obtención del modelo conceptual a partir de la estructura lógica de los sistemas heredados de las instituciones, teniendo en cuenta que estos generalmente utilizan un modelo relacional en su estructura; para ello se propone utilizar la interrelación entre entidades para generar un grafo de interrelación de atributo y luego aplicar un grupo de reglas de diseños para obtener el modelo conceptual.

Palabras clave: Diseño de almacenes de datos; Modelo de almacenes de datos; diseño conceptual.

1. Introducción

Una de las temáticas que más desarrollo demanda en los momentos actuales en cuanto a las tecnologías software es sin dudas la Ingeniería del Software, y dentro de esta, el desarrollo de métodos y algoritmos apropiados que conlleven a la creación y explotación de herramientas que asistan por medios computacionales al desarrollo de otros sistemas (herramientas CASE).

Dentro del campo de los Sistemas de Información, basados en la explotación de Sistemas de Bases de Datos, en los últimos tiempos se ha desarrollado la creación y explotación de Almacenes de Datos como uno de los típicos Sistemas de Ayuda a la Toma de Decisiones. Estos sistemas que por concepto, manipulan información histórica recopilada en sistemas de información tradicionales de gestión empresarial, tienen el objetivo de descubrir nuevas informaciones que permitan avanzar en productividad y logros de todo tipo; se caracterizan entre otras cosas -según

varios autores entre los que se pueden mencionar a [1] y [2] por tener una estructura de tablas bastante diferente a la heredada de los sistemas tradicionales (casi siempre relacionales).

Desde la introducción del modelo de datos multidimensional como formalismo de modelado para Almacenes de Datos, han aparecido en la literatura sobre el tema, distintas propuestas metodológicas para capturar la estructura del almacén de datos. Las soluciones siguen diferentes aproximaciones al diseño: las soluciones que tienen en cuenta solamente las necesidades de los usuarios (dirigidos por los requisitos), los que analizan la fuente de los datos (dirigido por las fuente de los datos) y algunos proponen un híbrido de estos dos paradigmas.[3]

Para obtener un buen diseño del almacén de los datos es necesario utilizar una metodología que tenga en cuenta las necesidades de los usuarios y también las fuentes de los datos.

El problema básico del diseño de un AD consiste en obtener un conjunto de esquemas multidimensionales que permitan satisfacer los requisitos de análisis de los usuarios y que puedan ser mantenidos por las bases de datos operacionales existentes en la organización. Las etapas del diseño de un AD pasa por el modelo conceptual, lógico y Físico.

Algunas metodologías describen como llegar al diseño conceptual de un almacén de datos, dentro de estas podemos citar a: [4] que propone un diseño del almacén de los datos en donde se pasa por cuatro fases secuenciales: Análisis y especificación de los requerimientos, diseño conceptual, lógico y físico. Proponen obtener el modelo conceptual a partir de la entrada del esquema E/R de los sistemas operacionales, dividen este proceso en tres etapas secuenciales y posteriormente analizan cómo este cumple con la forma normal multidimensional.

En [5] se propone una metodología híbrida en donde primero se aplica un paradigma suministrado por los datos para determinar los esquemas candidatos y posteriormente ser validado por una fase dirigida por los requerimientos. Esta es una de las primeras metodologías que se acercan al automatizado del proceso. En este trabajo se maneja la filosofía que las tablas que contienen más campos numéricos en el modelo relacional son promisorias a constituir una tabla de hecho en el modelo multidimensional, además plantean que cualquier tabla que tenga relación de uno a mucho con estas jueguen un papel de dimensiones; sin embargo, este acercamiento genera muchos resultados y no trabaja con la posibilidad de obtener un modelo que contenga una tabla de hecho sin hecho y tampoco se describe cómo se obtendría un modelado de constelación de hechos.

En [6] se presenta un acercamiento híbrido para obtener un esquema multidimensional conceptual. Ellos proponen recoger los requisitos multidimensionales y después trazarlos hacia las fuentes de los datos en un proceso de conciliación. Sin embargo, ellos sugieren que su acercamiento también pudiera ser considerado como demanda-manejado si el usuario no quiere tener en cuenta las fuentes de los datos. El autor introduce una metodología orientada a metas basándose en una estructura i*.

En [7] se presenta una metodología para obtener el esquema conceptual multidimensional realizando preguntas SQL a los modelos relacionales. Este acercamiento es totalmente automático y sigue un paradigma híbrido. Aunque no lleva a cabo dos fases bien definidas (manejado por los datos o manejado por los requisitos), este es la primera que trata de automatizar la fase de manejado por los requisitos.

Como podemos observar existen un gran número de metodologías que tratan de obtener el modelo conceptual. Existen otros trabajos que obtienen el modelado directamente a nivel lógico y algunos llegan a la obtención del modelo físico.

En [8] se describe una herramienta que permite diseñar esquemas de AD a través de transformaciones, permitiendo obtener un modelo lógico de AD y una traza del diseño. Ellos desarrollan el modelo lógico del AD utilizando un grupo de transformaciones de esquemas a las cuales le llaman primitivas, las mismas abstraen y materializan técnicas de diseño, pero ellos no tienen en cuenta el modelo conceptual ni tampoco realizan una propuesta de las tablas de hechos, ni trabajan con base de datos heterogéneas. En [9] el trabajo parte de que el diseñador representa universo del discurso utilizando notación UML y con esto obtiene un esquema UML, al cual luego lo enriquecen, pero al igual que el anterior no tiene en cuenta las bases de datos heterogéneas y tiene como dificultad que el diseñador tiene que analizar la base de dato fuente, y solo trabaja para la generalización y agregación.

En [3] se propone una metodología del diseño de almacenes de datos a nivel conceptual utilizando un contexto de MDA. En [10] se observar un trabajo profundo sobre MDA para la generación automática de código a partir de modelos UML.

En [11] se define un armazón metodológica general para el diseño de almacenes de datos. Basado en un modelo de hechos dimensional (DFM), el autor habla de una metodología de forma general, pero no realiza un trabajo profundo en el proceso de semiautomatizar el análisis del esquema de las base de datos ni cómo se trabaja con las bases de datos heterogéneas, solo las menciona, dejando todo este trabajo en manos del diseñador.

En este estudio se presenta un grupo de pautas metodológicas para la generación asistida por computadoras de modelos de almacenes de datos (AD), estas pautas metodológicas se dividen en 4 etapas, las 3 etapas primeras son: análisis de la información, modelo conceptual y diseño lógico; es necesario señalar que las etapas deben de aplicarse en el orden en que se mencionan y una cuarta etapa que ocurre dentro de cada una de las anteriores que es denominada etapa de trazabilidad. Estas etapas describen cómo podemos obtener a partir de los sistemas operacionales heredados utilizando el modelo lógico (relacional) una propuesta de modelado de almacenes de datos. En [12] se propone el desarrollo de la primera etapa (Análisis de requisito de información), este es un trabajo desarrollado por los mismos autores de este trabajo, por lo que en este trabajo solo se tocara algunos puntos importantes y el trabajo se enfocara en todos los detalles para la obtención del modelo conceptual. Se desarrollara un caso de estudio en donde se detalla cada uno de los elementos fundamentales de esta etapa.

Tabla 1.
Pautas Metodológicas.

Etapas	Entradas	Salida	Involucra
Análisis de requisitos de información	1. Requisitos de Usuario 2. Esquema OLTP	Esquema refinado	Diseñador, gerente de sistema de información y usuario final.
Modelo conceptual.	Esquema refinado	Modelo conceptual	Sistema, diseñador.
Diseño lógico	Modelo conceptual	Esquema lógico del AD	Sistema, diseñador.
Trazabilidad	Todas las etapas.	Traza del modelo.	Sistema

Fuente: Beltrán, L. A. G., et al., 2013

2. El desarrollo de las etapas

En este apartado se presenta mediante la Tabla 1 la relación de cada una de las etapas, cuáles son las entradas de información necesarias y cuáles son sus salidas, además de los actores involucrados en cada una de las etapas.

2.1. Análisis de la información. (Análisis de requisitos de información)

En los últimos años se han consagrado muchos esfuerzos por lograr metodologías de modelado multidimensional, en este sentido, se han logrado varios acercamientos y se ha presentado en la literatura, de forma que apoyen al diseño multidimensional de los almacenes de datos.

En [12] se trabajó el análisis de la información utilizando el paradigma híbrido secuencial, en donde se utilizó primero un análisis de los requisitos de los usuarios para determinar las necesidades del usuario final y posteriormente analizar los sistemas OLTP de las bases de datos fuentes, buscando que las necesidades del usuario puedan ser mantenidas por las bases de datos operacionales existentes en la organización.

En esta primera etapa se distinguen dos fases a desarrollar de forma secuencial:

Fase 1 - Especificación de requisitos de usuario: consiste en identificar las necesidades de análisis de los usuarios.

Fase 2 – Analizar los modelos de las base de datos operacionales: con el objetivo de buscar la información que mantenga las necesidades del usuario final dando como resultado un esquema refinado.

En la literatura se puede encontrar que esta primera fase cuyo objetivo es obtener los requisitos de información que tienen los usuarios para el apoyo a la toma de decisiones, i.e. medidas interesantes y el contexto para su análisis. Puede realizarse utilizando diferentes métodos, se señalan dos trabajos que utilizan diferentes métodos y que se consideran los más importantes. En [3] se adapta un método de elicitación de requisitos basado en metas y en [13] se utiliza una aproximación basada en objetivos (mediante el uso de i*).

En la primera fase se considera trabajar con la propuesta utilizada en [3] adoptando un método de elicitación de requisitos basado en metas, donde los requisitos de usuario son recogidos por medio de entrevistas. El propósito de las entrevistas es obtener información acerca de las necesidades de análisis de la organización. El aporte fundamental que se realiza en esta primera etapa se encuentra en la restructuración de la elicitación de requisito, para contribuir a una óptima documentación del proyecto de creación de almacenes de datos.

2.1.1. Fase 1 - Especificación de requisitos de usuario

Teniendo en cuenta el trabajo de [3] y [14] el esquema del diseño propuesto es dividido en tres etapas: a) definición de la misión, b) identificación de las metas de análisis y c) especificación de los requisitos.

En la Tabla 2 se muestra un resumen de cómo quedaría la documentación del proyecto relacionado con el caso de estudio que se utiliza perteneciente al análisis de la actividades de intervenciones quirúrgicas y de urgencia de un hospital [15], [12].

2.1.2. Fase 2 – Analizar las base de datos operacionales

Para la segunda fase de la primera etapa se propone un grupo de pasos que pueden ser automatizados y que cumpla con los requisitos necesarios para desarrollar las siguientes etapas de la metodología propuesta.

1. Análisis de los esquemas de la(s) base(s) de dato fuente(s).
2. Selección de la información de interés.
3. Validación de la información seleccionada.

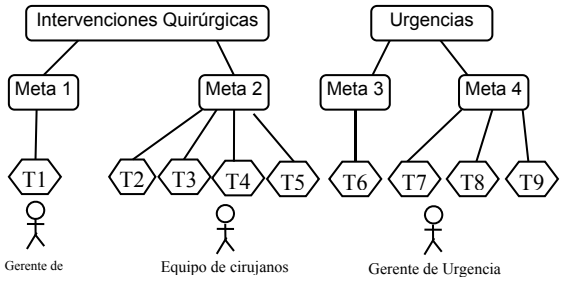
El objetivo general de esta segunda fase es que el diseñador y el usuario final puedan determinar la información que describa los eventos que ocurren dinámicamente en las empresas y que presentan la forma estructural de las empresas.

El primer paso, análisis de los esquemas de la(s) base(s) de dato fuente(s), es utilizado para determinar que OLTP es necesario para poder soportar los requisitos de información detectados en la primera fase.

Se propone una herramienta que es capaz de ayudar al diseñador en esta labor. Para ello la herramienta es capaz de conectarse a las bases de datos de los OLTP y extraer el esquema y guardarlo en un fichero XML. La herramienta logra la conexión con un gestor de base de dato relacional.

El segundo paso: selección de la información de interés, es necesario para deslindar entre los datos que aportan al almacén de dato y que se corresponde con lo determinado en la primera fase. En la Fig. 1 se muestra el resultado de la implementación de la segunda fase de la primera etapa.

Tabla 2.
Resumen de las etapas a y b.

Misión	El sistema proporcionará los medios necesarios para analizar el Hospital Clínico "San Cecilio" de Granada, España. Con el objetivo de proporcionar información sobre la conducta de las actividades de Intervenciones Quirúrgicas y Urgencias.		
P.N.	**Metas**	**Tarea**	**Usuario**
Intervenciones Quirúrgicas	**Meta 1:** Calidad de la planificación de las intervenciones	T1: Analizar las intervenciones suspendidas con la intención de determinar las causas.	Gerente de intervenciones
	Meta 2: Efectividad de los tratamientos en las intervenciones	T2: Analizar las intervenciones realizadas con el fin de ver el comportamiento de este con el tratamiento.	Equipo de cirujanos
		T3: Analizar las intervenciones realizadas con el fin de ver el comportamiento de las anestesias utilizadas.	
		T4: Analizar las intervenciones realizadas con el fin de ver la relación de esta con el diagnostico.	
		T5: Analizar las intervenciones realizadas con el fin de ver el comportamiento de los implantes.	
Urgencias	**Meta 3:** Calidad de los servicios de urgencia	T6: Analizar el Tiempo que demora la atención de las urgencias con el fin de determinar las causas.	Gerente de Urgencia
	Meta 4: Efectividad y eficiencia de las urgencias.	T7: Analizar las urgencias con el fin de determinar el número de enfermos por las diferentes localidades.	
		T8: Analizar las urgencias para determinar las causas.	
		T9: Analizar las urgencias para determinar las áreas más afectadas.	
Árbol de Refinamiento de Metas			

Fuente: Los autores

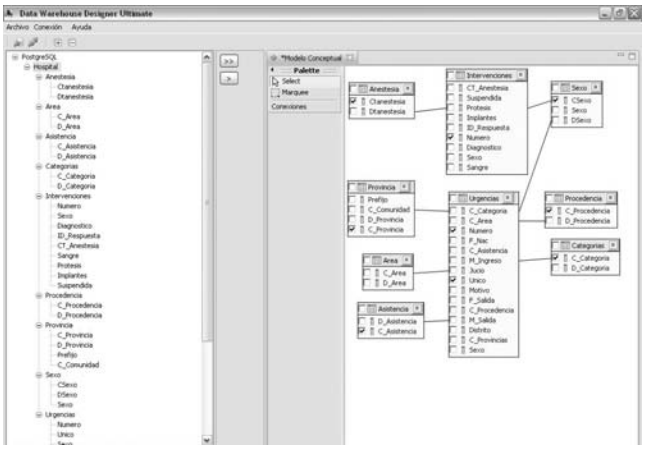


Figura 1: Imagen de la herramienta en la primera etapa.
Fuente: Los autores

El tercer paso se propone para validar que la selección sea correcta y tenga querencia y para ello se deben de verificar las siguientes reglas.

Regla 1: Todas las tablas que se seleccionaron deben de estar relacionadas entre sí.

Descripción de la Regla: Para la validación de los datos seleccionados es necesario comprobar que todas las tablas que se seleccionaron tienen relación entre sí. Esto garantiza que no existan tablas aisladas que no representa ninguna información.

Es importante señalar que se debe de tener presente el trabajo con las base de datos heterogéneas, para ello definimos la siguiente regla.

Regla 2: Las base de datos heterogéneas deben de relacionarse por algún atributo que se encuentre en las base de datos a relacionar y que represente al mismo dominio.

Descripción de la Regla: La relación entre base de datos está dado por la posible existencia de diferentes gestores de base de datos y la necesidad de la integración entre estas. Es importante señalar que nosotros no realizamos transformación de los tipos de datos entre los gestores, solo dejamos la propuesta para que se realice a la hora de implementar el proceso de transformación, limpieza y carga.

Al terminar este proceso queda un esquema refinado relacional de las base de datos que intervienen en la construcción del almacén de dato.

2.2. Modelo conceptual

Después de obtener el esquema refinado que satisface las necesidades del usuario y quedando soportado por los OLTP se propone trabajar en la obtención del modelo conceptual.

Para la obtención del modelo conceptual se propone que se ejecuten una serie de pasos:

1. Seleccionar los elementos que se convertirán en hechos.
2. Obtener un grafo de interrelación de atributo.
3. Determinar el(los) hecho(s).

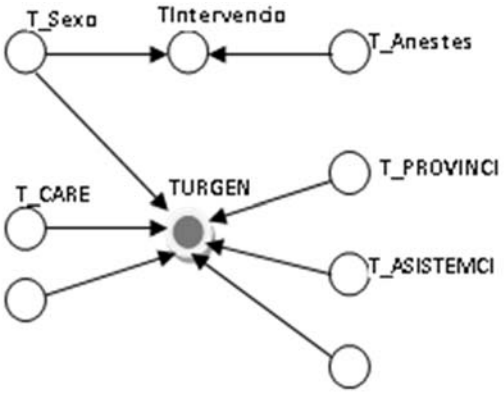


Figura 2: Diagrama de interrelación.
Fuente: Los autores

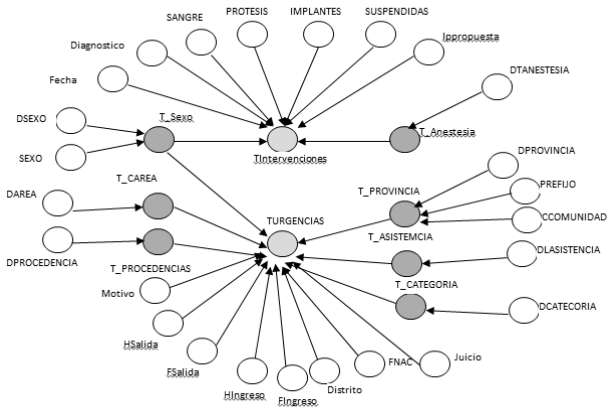


Figura 3: Grafo de interrelación de Atributo.
Fuente: Los autores

4. Determinar las dimensiones.
   4.1. Seleccionar los niveles.
   4.2. Seleccionar los descriptores.
5. Determinar las medidas.
   5.1. Mediadas directas.
   5.2. Medidas Indirectas.
6. Obtención del diagrama del modelo conceptual.
7. Refinamiento del modelo conceptual.

Es importante señalar que para que estos pasos se puedan automatizar se definen un grupo de reglas. En lo adelante comenzaremos a explicar cada uno de estos pasos.

El primer paso de esta segunda etapa es la selección del hecho o de los hechos, en los últimos tiempos en las empresas se ha comenzado a trabajar con diferentes hechos dentro del mismo modelo, aspecto el cual no se ha tenido presente en ninguna de las metodologías que se analizaron.

En la herramienta se realiza una propuesta al diseñador de que elemento se puede convertir en hecho, pero queda de manos del diseñador elegir la propuesta o seleccionar otro elemento, esta propuesta se realiza teniendo en cuenta la interrelación de las tablas. En la Fig. 2 podemos ver un ejemplo de la selección de la propuesta.

Como se puede observar en la Fig. 2 y utilizando la teoría de grafo, podemos notar que se puede obtener un grafo dirigido y con este determinar el grado de un nodo, para seleccionar la propuesta se propone tener en cuenta los siguientes criterios.

1. Los nodos que mayor grado tienen.
2. Los nodos que mayor grado de salida tienen.

Se le muestra una propuesta al diseñador para que él escoja cuál de los elementos puede convertirse en su hechos o hechos, primero se analiza qué nodo tiene el grado más alto y si existen más de uno con el mismo grado, entonces se selecciona cuál de estos tiene el mayor grado de salida, esto último se tiene en cuenta ya que ellos representan a las relaciones de mucho a mucho y son muy probables a convertirse en hechos.

En el ejemplo al diseñador se le muestra la propuesta del nodo TURGENCIA que es el nodo que mayor grado posee. El diseñador, en su selección, teniendo presente lo seleccionado en la primera etapa y escoge además de TURGENCIA al elemento TINTERVENCIÓN.

Posteriormente de seleccionar los hechos se pasa al siguiente paso que es la obtención de un grafo de interrelación de atributo, dando lugar al grafo que se muestra en la Fig. 3:

A partir de este momento se comenzarán a ejecutar una serie de pasos y en cada uno de ellos se aplican algunas reglas para transformas este grafo de interrelación de atributo al modelo multidimensional.

Es importante destacar los elementos que se tienen en cuenta para la construcción del modelo conceptual.

Este modelo está compuesto por un hecho o por varios hechos y cada uno tiene o no mediadas asociadas, las medidas se pueden clasificar en directas o indirectas. Las medidas directas son aquellas que se obtienen directamente de la fuente de los datos y las indirectas las que se obtienen por la transformación o el cálculo de algún atributo. Los hechos tienen relacionado a él un grupo de dimensiones que pueden estar compartidas para diferentes hechos, las dimensiones están compuestas por lo menos de un nivel y cada nivel tiene asociado al menos un descriptor.

Para la construcción de este modelo el primer paso es crear el hecho, para ello se definen la siguiente regla.

Regla 3: Cada nodo seleccionado como candidato a convertirse en hecho genera una tabla de hecho.

Descripción de la Regla: Cada uno de los nodos que se seleccionó como candidato a convertirse en hecho se transforma como una tabla de hecho en donde el nombre del nodo pasa a ser el nombre de la tabla de hecho. Si este nodo tiene un identificador propio este pasa a ser parte de los identificadores del hecho.

El segundo paso es la determinación de las dimensiones:

Regla 4: Cada nodo que tiene relación con el nodo convertido en hecho y es un nodo entidad y además tiene asociado a él al menos un nodo atributo, se convertirá en dimensión del hecho.

Descripción de la Regla: Los nodos que cumplen con estas condiciones se convierten en tablas de dimensión, en donde el nombre del nodo se convierte en el nombre de la dimensión, esta dimensión será completada cuando se analicen las otras reglas.

Si el nodo entidad no posee un nodo atributo este se elimina y se sigue el análisis en el nivel siguiente del grafo de interrelación de atributo.

Es importante resaltar un concepto introducido, los

nodos entidad son clasificados así cuando se genera el grafo de interrelación de atributo en donde se diferencia entre los nodos que provienen de atributos y los nodos que provienen del nombre de la entidad.

En muchas ocasiones es importante tener los datos en varios niveles de granularidad, es decir, es importante para los negocios poder consultar los datos a distintos niveles.

En las dimensiones es importante identificar cuáles son sus niveles de detalle, para ello se identificó la siguiente regla.

Regla 5: Los niveles de una dimensión son aquellos nodos que se relacionan con el nodo convertido en dimensión incluyendo él mismo.

5.1: Nivel 0: es la propia dimensión.

En el nivel 0 de la dimensión consideramos que siempre debe de existir al menos un descriptor, si esto no ocurre se elimina este nivel y la dimensión tomaría el nombre del próximo nivel. Esto se tiene en cuenta para no obtener una representación de una relación de mucho a mucho entre el hecho y la dimensión que verdaderamente representa el contexto del modelo.

5.2: Nivel > 0: representa las entidades que tienen relación con el nodo convertido en dimensión y se considera la jerarquía dentro de la dimensión.

En el resto de los niveles se considera que hay que eliminar aquellos niveles que no contienen descriptores y que interrelacionan dos niveles. Esto es necesario para eliminar las posibles relaciones de mucho a mucho y que no representan ninguna información para el modelo multidimensional.

Otro elemento que se considera que hay que tener presente en el modelo conceptual son los descriptores. De aquí se deriva la siguiente regla.

Regla 6: Los nodos interrelacionados a un nodo convertido en nivel de una dimensión y que estos nodos sean nodos atributos son convertidos en descriptores del nivel en cuestión.

Descripción de la Regla: Los descriptores se encuentran asociados a los niveles que poseen una dimensión y describen cierta característica del nivel y de la dimensión.

Para ver de forma clara cada uno de estos elementos se muestra el ejemplo de la Fig. 4 en donde se puede ver de forma clara estas dos situaciones.

Este ejemplo muestra un caso de estudio que es muy utilizado por la comunidad científica [3] , donde se representa las ventas de una cadena de tiendas.

En la Fig. 5 podemos observar el grafo de interrelación de atributo obtenido. En donde el hecho a analizar es

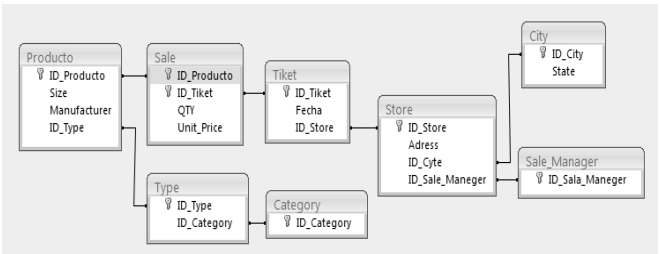


Figura 4: Base de dato relacional de ventas.
Fuente: Los autores

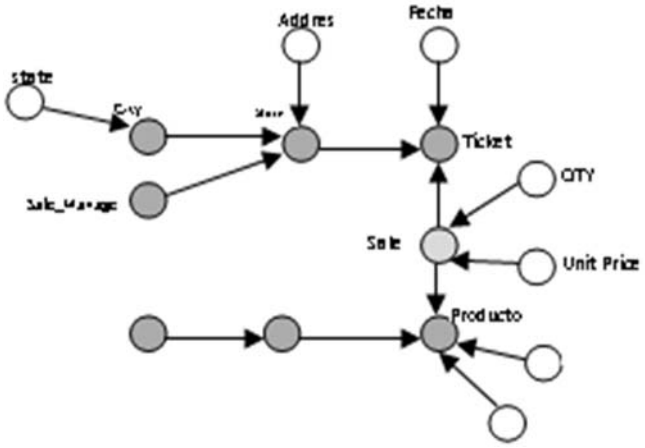


Figura 5: Grafo de interrelación de atributo.
Fuente: Los autores

Tabla 3.
Niveles y descriptores.

Dimensión	Nivel	Nombre	Descriptores
TICKET	Nivel 0	Ticket	Fecha
	Nivel 1	Store	Address
	Nivel 2	City	State
		Sale Manager	Manager
PRODUC	Nivel 0	Product	Size
			Manufacture
	Nivel 1	Type	Type
	Nivel 2	Category	Category

Fuente: Los autores

"SALE", aplicando la regla 4 las dimensiones son "TICKET" y "PRODUC" y los niveles y descriptores se muestran en la Tabla 3.

Una dimensión que tiene mucha importancia en el diseño de los almacenes de datos es la dimensión tiempo, por tal motivo se deja para analizar su obtención de forma separada a las restantes dimensiones.

Nuestra propuesta tiene en cuenta la obtención de esta dimensión en tres condiciones que a nuestro entender pueden suceder.

1. Determinar atributos que su tipo de dato sea de tiempo o fecha.
2. Pueden existir atributos que sean de tipo de dato numérico o de texto que representen la fecha.
3. En los OLTP no existe diseñado ningún atributo que represente la fecha.

Regla 7: Buscar algún nodo que sean de tipo de dato tiempo o fecha, crear una dimensión fecha y poblarlo con este.

En el primero de los casos la dimensión tiempo se poblará a partir de la transformación de un atributo de los OLTP de tipo de dato fecha, en ocasiones en los OLTP pueden existir muchos atributos que cumplan con esta condición, en la propuesta se selecciona aquellos que se encuentran relacionados directamente con el hecho o se encuentran más cercano a éste, en el caso que no se pueda

determinar con precisión el diseñador debe de seleccionar cuál será este atributo.

Regla 8: Si no existiera un atributo de tipo fecha pero existe algún atributo que se pueda extraer esta información el diseñador debe de especificarlo y tenerlo en cuenta para cuando se realice el proceso de transformación.

Otra de las condiciones que en ocasiones suceden es cuando no existe ningún atributo de tipo de dato fecha, pero en el OLTP existe algún atributo que puede transformarse y poblar a esta dimensión, en este caso permitimos que el diseñador los seleccione y le recomendamos que debe de tenerlo en cuenta para el proceso de limpieza y transformación.

Regla 9: Si el diseñador no pudo seleccionar un atributo para poblar la dimensión tiempo se propone poblar ésta cuando se realicen cargas de los OLTP al modelo multidimensional.

El último de los casos es muy difícil que suceda pero hay que tenerlo presente por si ocurriese; este puede manifestarse cuando no existe ningún atributo que pueda ser utilizado para poblar la dimensión tiempo, en este caso se propone al diseñador que determine los niveles y que obtenga el poblado de la dimensión tiempo a través de la fecha de la PC en el momento de carga, este último caso no es recomendado.

Luego de obtener el hecho y las dimensiones pasamos a definir las medidas del modelo, se considera al igual que [3, 16, 17] que se puede clasificar las medidas en dos categorías teniendo como referencia el origen de esta.

Medidas Directas: Estas se obtienen directamente de la fuente sin realizar ningún tipo de transformación o cálculo.

Medidas Indirectas: Son aquellas medidas que se obtienen transformando o realizando un cálculo de algunos de los datos de la fuente.

Regla 10: Todos los atributos que tienen relación con el nodo raíz que no son identificadores y no son entidades se convertirán en medidas directas de la tabla de hecho en cuestión.

Para la determinación de las medidas directas se analiza en la tabla de hecho todos los atributos relacionados con éste y se propone que se conviertan en medidas directas. La selección de las medidas de este tipo queda por parte del diseñador. Después de proponer cuáles atributos pueden convertirse en medidas el diseñador debe de seleccionar cuáles considera son de interés y cuáles tienen sentido semánticamente.

En el ejemplo de ventas que se trabaja existen dos medidas que provienen de los sistemas operacionales directamente que son QTY y Unit Prece.

Regla 11: Se pueden agregar atributos a la tabla de hecho que provengan de una operación de cálculo o de las funciones de agregación SUM, AVG, MAX, MIN y COUNT.

Las medidas indirectas o también llamadas en la literatura como funciones de agregación o contadores de instancias de la tabla de hecho, pueden ser de distintos tipos como SUM, AVG, MAX, MIN y COUNT.

En la propuesta se obtiene un diagrama del modelo conceptual, este tiene en cuenta todos los elementos mencionados anteriormente. En las Figs. 6 y 7 se puede ver el diagrama de los dos ejemplos que se trabajan.

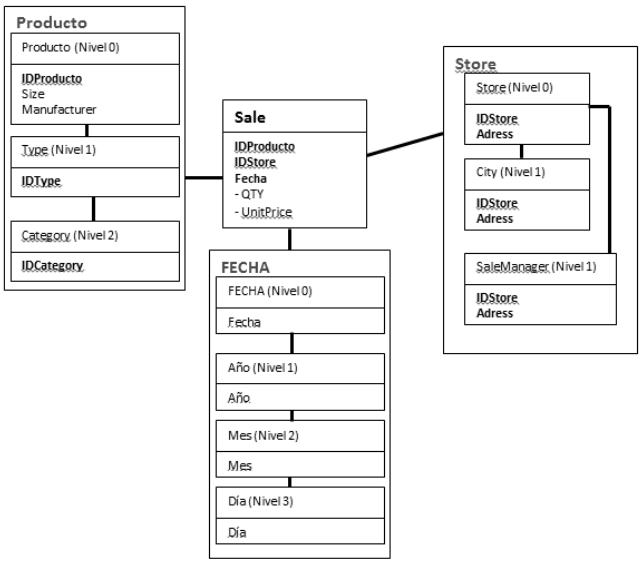


Figura 6: Diagrama del modelo conceptual de ventas
Fuente: Los autores

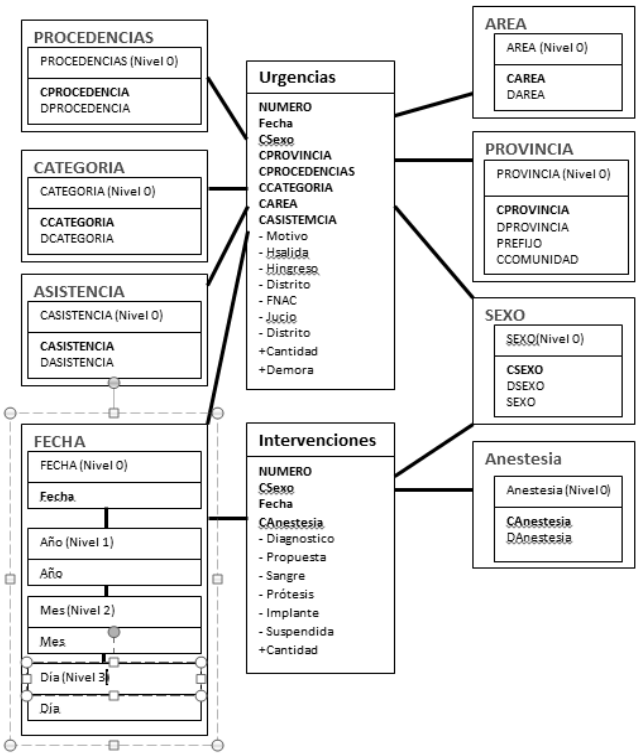


Figura 7: Diagrama del modelo conceptual de intervenciones y urgencia de Hospital.
Fuente: Los autores

Como se puede observar el la Fig. 6 la dimensión Ticket fue sustituida por el próximo nivel de la jerarquía debido a que el único atributo que contenía este nivel está relacionado con el atributo convertido en la dimensión tiempo y por quedar sin sentido el nivel ticket este fue eliminado. También es importante señalar que en la

dimensión tiempo se consideró los niveles de año, mes y día, es importante señalar que esta dimensión tiene que tener al menos el nivel de año.

2.3. Refinamiento del modelo conceptual.

En la etapa de refinamiento del modelo conceptual es importante tener presente los siguientes aspectos:

1. Medidas directas que no son numéricas que pueden convertirse en dimensiones.
2. Niveles dentro de una jerarquía que se pueden convertir en dimensiones.
3. Agregar niveles de detalle en la dimensión tiempo.
4. Eliminar medidas directas que no tienen sentido.
5. Cambiar nombre de atributos y de tablas que provienen de los OLTP.
6. Eliminar atributos, jerarquías y tablas que no tienen sentido.

Todos estos elementos son importantes para poder tener un modelo conceptual refinado que no contenga elementos indeseados en la modelación del almacén de dato.

3. Conclusiones

Este trabajo es una continuidad a un anterior estudio realizado [12] en donde se presentó la etapa de análisis de la información. En el presente trabajo se profundiza en la etapa de obtención del modelo conceptual, en donde se exponen cómo obtener el modelo conceptual de forma semiautomática mediante la aplicación de 11 reglas de diseño.

Se presenta cuáles son los pasos que sigue la metodología para la obtención de este modelo, los cuales son: seleccionar los elementos que se convertirán en hechos, obtener un grafo de interrelación de atributo, determinar el(los) hecho(s), determinar las dimensiones donde se tiene que seleccionar los niveles y los descriptores, determinación de las medidas directas o indirectas, obtención del diagrama del modelo conceptual y por último el refinamiento del modelo conceptual. En trabajos futuros se abordara la transformación de este modelo conceptual al modelo lógico.

Bibliografía

[1] Inmon, W.H., Building the Data Warehouse - Fourth Edition, Wiley Publishing, 2005.

[2] Kimball, R., Ross, M., Thornthwaite, W. et al., The data warehouse lifecycle toolkit: Expert methods for designing, developing and deploying data warehouses, John Wiley & Sons, 1998.

[3] Sánchez, L.Z., Metodología para el diseño conceptual de almacenes de datos, Dr. Tesis, Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, Valencia, España, 2008.

[4] Hüsemann, B., Lechtenbörger, J., and Vossen, G., Conceptual datawarehouse design, en: The International Workshop on Design and Management of Data Warehouses (DMDW 2000), 2000.

[5] Phipps, C. and Davis, K.C., Automating data warehouse conceptual schema design and evaluation, en: 4th International Workshop on Design and Management of Data Warehouses, Toronto, Canada, pp. 23-32, 2002.

[6] Giorgini, P., Rizzi, S. and Garzetti, M., Goal-oriented requirement analysis for data warehouse design, en: DOLAP'05, Bremen, Germany, 2005.

[7] Romero, O. and Abelló, A., Multidimensional design by examples, en: 8th International Conference on Data Warehousing and Knowledge Discovery, Krakow, Poland, pp. 85-94, 2006.

[8] Marotta, A., Data warehouse design and maintenance through schema transformations, MSc. Thesis, Instituto de Computación - Facultad de Ingeniería, Universidad de la República, Uruguay, 2000.

[9] Akoka, J., Wattiau, I.C. and Prat, N., Dimension hierarchies design from UML generalizations and aggregations, en: ER 2001, Verlag Berlin Heidelberg, 2001.

[10] [10]Arango, F., Gómez, M.C. and Jaramillo, C.M.Z., Transformación del modelo de clases UML A. Oracle9i® bajo la directiva mda: Un caso de estudio. DYNA [Online], 73 (149), 2006. Available at: http://www.bdigital.unal.edu.co/10976/1/fernandoarangoisaza.2006.pdf.

[11] Golfarelli, M. and Rizzi, S., A methodological framework for data warehouse design, in: ACM First International Workshop on Data Warehousing and OLAP, Washington, D.C., USA.,1998.

[12] Beltrán, L.A.G., Rodríguez, R.M. and Vázquez, R.P., Generación asistida por computadoras de modelos de almacenes de datos: Análisis de la Información. DYNA [Online], 80 (177), 2013. Available at: http://www.revistas.unal.edu.co/index.php/dyna/article/view/28017.

[13] Mazón, J.N. and Trujillo, J., Desarrollo de modelos multidimensionales de almacenes de datos basado en MDA: Del análisis de requisitos al modelo lógico, en: Actas del IV Taller sobre Desarrollo de Software Dirigido por Modelos, MDA y Aplicaciones, 2007, pp. 41-50.

[14] Pelozo, C.E.I., A requirements engineering approach for object-oriented conceptual modeling, Dr. Thesis, Information Systems and Computation Department, Valencia Polytechnic University, Valencia, Spain, 2003.

[15] Ávila, E.T., Data warehousing con procesamiento de datos textuales, Dr. Tesis, Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, Granada, España, 2010.

[16] Kamble, A.S., A conceptual model for multidimensional data, en: Proceedings of the fifth Asia-Pacific conference on Conceptual Modelling (APCCM '08), Australia, 2008.

[17] Gerardo, C.G., Un sistema para el mantenimiento de almacenes de datos, Dr. Tesis, Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, Valencia, España, 2008.

L.A. Gómez-Beltrán, recibió su título de Ingeniero en Informática en 2005 del Instituto Superior Politécnico José Antonio Echeverría (CUJAE), la Habana Cuba, el MSc. en Informática Aplicada en 2010, de la Universidad de Camagüey, Cuba. Desde el 2005 trabaja para la Universidad de Camagüey, Cuba. Actualmente es Profesor Auxiliar del Departamento de Computación de la Facultad de Informática y ocupa el cargo de Director de Informatización de la Universidad de Camagüey Cuba. Sus intereses de investigación incluyen: bases de datos, almacenes de datos, diseño conceptual de almacenes de datos.

R. Moreno Rodríguez, recibió su título de Ingeniería en Sistemas Automatizados de Dirección en 1980, el MSc en Ingeniería en 1981, ambos del Instituto Superior Electrotécnico de Leningrado "V.I. Ulianov (Lenin)", Rusia. Es MSc. en Computación Aplicada en 1998 y Dr en Ciencias Técnicas en 2006 de la Universidad Central de Las Villas – UCLV, Cuba. Actualmente es profesor Titular de Facultad de Matemática, Física y Computación de la Universidad Central de Las Villas, Cuba, labora en Centro de Estudios de Informática (CEI) y ocupa el cargo de Asesor metodológico de Postgrado en el Departamento de Postgrado de la misma universidad. Sus intereses de investigación incluyen: bases de datos.

R. Pérez-Vázquez, recibió su título de Licenciatura en Computación en 1980 de Universidad Central de Las Villas. Es Dr en Ciencias Técnicas en 1991, de la Universidad Politécnica de Kiev, Rusia. Es profesor Titular, de la Facultad de Matemática, Física y Computación, Centro de Estudios de Informática (CEI) de la Universidad Central de Las Villas – UCLV, Cuba. Actualmente ocupa el cargo de secretario general de la UCLV. Sus intereses de investigación incluyen: bases de datos, Sistemas de Información Geográfico, Almacenes de Datos.