

Modelamiento multidimensional

Carmen Gloria Wolff

Presentación

En la edición anterior de Ingeniería Informática se comenzó una serie de artículos relacionados con la tecnología Datawarehousing, el primero de ellos "[Tecnología Datawarehousing](#)" introduce al lector en esta tecnología. A continuación se presentará algunas alternativas para enfrentar el modelamiento de un datawarehousing, a través del modelamiento multidimensional.

Introducción.

La tecnología Datawarehousing debido a su orientación analítica, impone un procesamiento y pensamiento distinto, la cual se sustenta por un modelamiento de Bases de Datos propio, conocido como Modelamiento Multidimensional, el cual busca ofrecer al usuario su visión respecto de la operación del negocio.

Modelamiento Dimensional es una técnica para modelar bases de datos simples y entendibles al usuario final. La idea fundamental es que el usuario visualice fácilmente la relación que existe entre las distintas componentes del modelo.

Consideremos un punto en el espacio. El espacio se define a través de sus ejes coordenados (por ejemplo X, Y, Z). Un punto cualquiera de este espacio quedará determinado por la intersección de tres valores particulares de sus ejes.

Si se le asignan valores particulares a estos ejes. Digamos que el eje X representa Productos, el eje Y representa el Mercado y, el eje Z corresponde al Tiempo. Se podría tener por ejemplo, la siguiente combinación: producto = madera, mercado = Concepción, tiempo = diciembre-1998. La intersección de estos valores nos definirá un solo punto en nuestro espacio. Si el punto que buscamos, lo definimos como la cantidad de madera vendida, entonces se tendrá un valor específico y único para tal combinación.

En el modelo multidimensional cada eje corresponde a una dimensión particular. Entonces la dimensionalidad de nuestra base estará dada por la cantidad de ejes (o dimensiones) que le asociemos. 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 entender más el concepto, retomemos el ejemplo anterior. La descripción de una organización típica es: "Nosotros vendemos productos en varios mercados, y medimos nuestro desempeño en el tiempo": Un diseñador dimensional lo verá como: "Nosotros vendemos productos en varios mercados, y medimos nuestro desempeño en el tiempo. Donde cada palabra subrayada corresponde a una dimensión.

Esto puede visualizarse como un cubo (Figura 3), donde cada punto dentro del cubo es una intersección de coordenadas definidas por los lados de éste (dimensiones). Ejemplos de medidas son: unidades producidas, unidades vendidas, costo de unidades producidas, ganancias(\$) de unidades vendidas, etc.

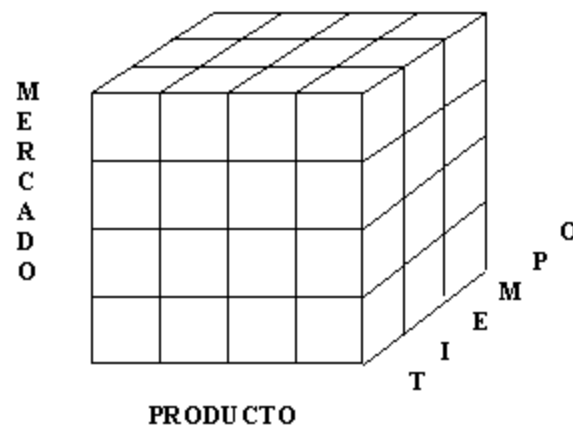


FIGURA 1: CUBO DIMENSIONAL PARA LA COMBINACIÓN DE PRODUCTO, MERCADO Y TIEMPO.

1.1 Modelos de Datos

Un factor clave presente durante todo el diseño de un DW, fue expresado por Codd en 1983: "Ustedes pueden pensar que el significado de los datos es simple...pero no es así".[Oviedo98]

Para construir un DW se debe primero tener claro que existe una diferencia entre la estructura de la información y la semántica de la información, y que esta última es mucho más difícil de abarcar y que también es precisamente con ella con la que se trabaja en la construcción de un DW.

Aquí se encuentra la principal diferencia entre los sistemas operacionales y el DW: Cada uno de ellos es sostenido por un modelo de datos diferente. Los sistemas operacionales se sustentan en el Modelo Entidad Relación (MER) y DDW trabaja con el Modelo Multidimensional.

1.1.1 Características del MER

- Maneja la redundancia fuera de los datos. Por lo tanto realizar un cambio en la base significa tocarla en un solo lugar.
- Divide los datos en entidades, las que son representadas como tablas en una base de datos.
- Los MER crecen fácilmente, haciéndose más y más complejos.
- Se puede apreciar la existencia de muchos caminos para ir de una tabla a otra. Sería natural pensar que al tener diversos caminos para llegar desde una tabla a otra, cualquiera de ellos entregaría el mismo resultado, pero lamentablemente esto no siempre sucede así.
- El diagrama se visualiza simétrico, donde todas las tablas se parecen, sin distinguir a priori la importancia de unas respecto a otras. No es fácil de entender tanto para usuarios como para los diseñadores.

1.1.2 Características del Modelo Multidimensional

En general, la estructura básica de un DW para el Modelo Multidimensional está definida por dos elementos: esquemas y tablas.

Tablas DW: como cualquier base de datos relacional, un DW se compone de tablas. Hay dos tipos básicos de tablas en el Modelo Multidimensional:

Tablas Fact : contienen los valores de las medidas de negocios, por ejemplo: ventas promedio en

dólares, número de unidades vendidas, etc.

Tablas Lock_up: contienen el detalle de los valores que se encuentran asociados a la tabla Fact.

Esquemas DW: la colección de tablas en el DW se conoce como Esquema. Los esquemas caen dentro de dos categorías básicas: esquemas estrellas y esquemas snowflake.

1.2 Conceptos asociados al DDW

1.2.1 Esquema Estrella.

En general, el modelo multidimensional también se conoce con el nombre de esquema estrella, pues su estructura base es similar: una tabla central y un conjunto de tablas que la atienden radialmente. (ver Figura 4).

El esquema estrella deriva su nombre del hecho que su diagrama forma una estrella, con puntos radiales desde el centro. El centro de la estrella consiste de una o más tablas fact, y las puntas de la estrella son las tablas lock_up.

Este modelo entonces, resulta ser asimétrico, pues hay una tabla dominante en el centro con varias conexiones a las otras tablas. Las tablas Lock-up tienen sólo la conexión a la tabla fact y ninguna más.

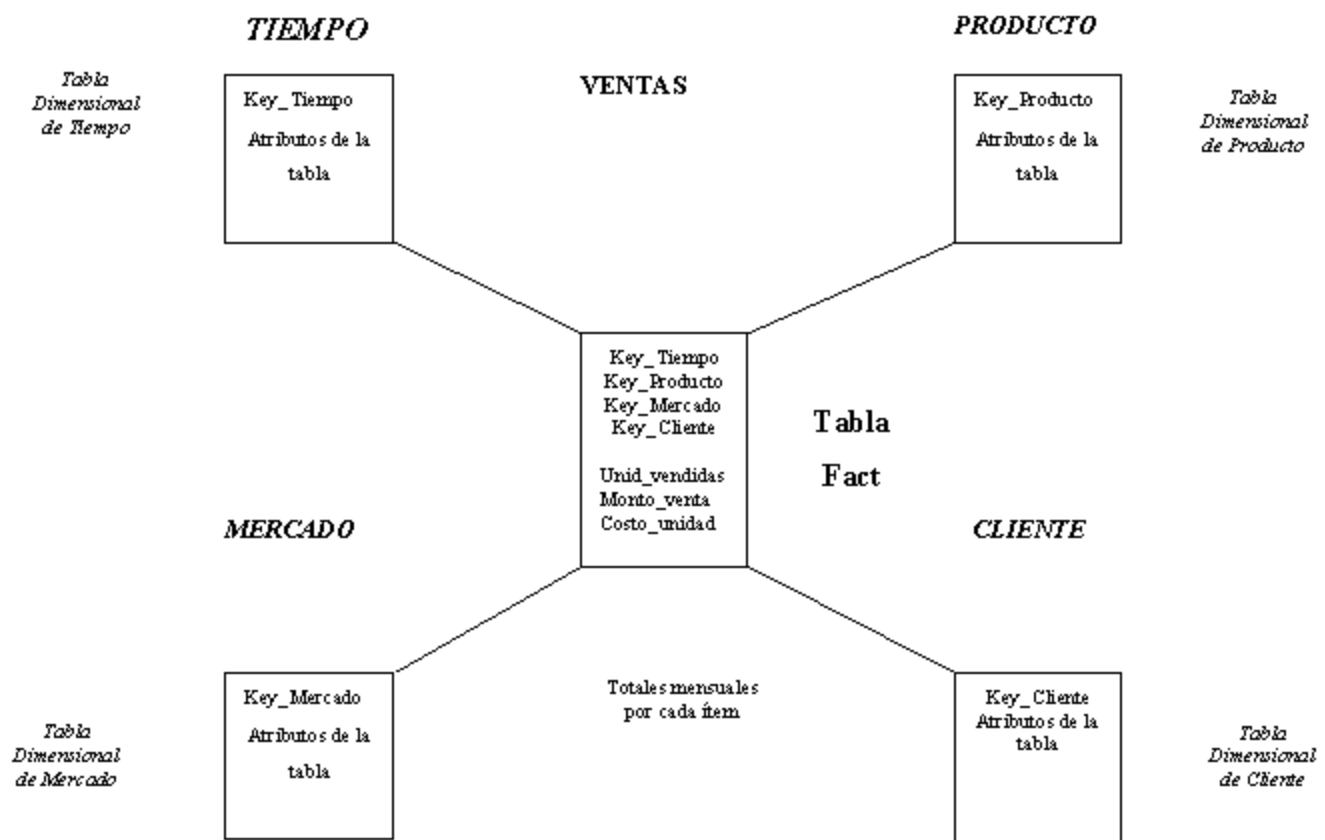


FIGURA 2: EJEMPLO DE UN ESQUEMA ESTRELLA PARA UNA BASE DE DATOS CON DIMENSIONES DE TIEMPO, PRODUCTO, MERCADO Y CLIENTE.

1.2.2 Esquema Snowflake.

La diferencia del esquema snowflake comparado con el esquema estrella, está en la estructura de las tablas lock_up: las tablas lock_up en el esquema snowflake están normalizadas. Cada tabla lock_up contiene sólo el nivel que es clave primaria en la tabla y la foreign key de su parentesco del nivel más cercano del diagrama.

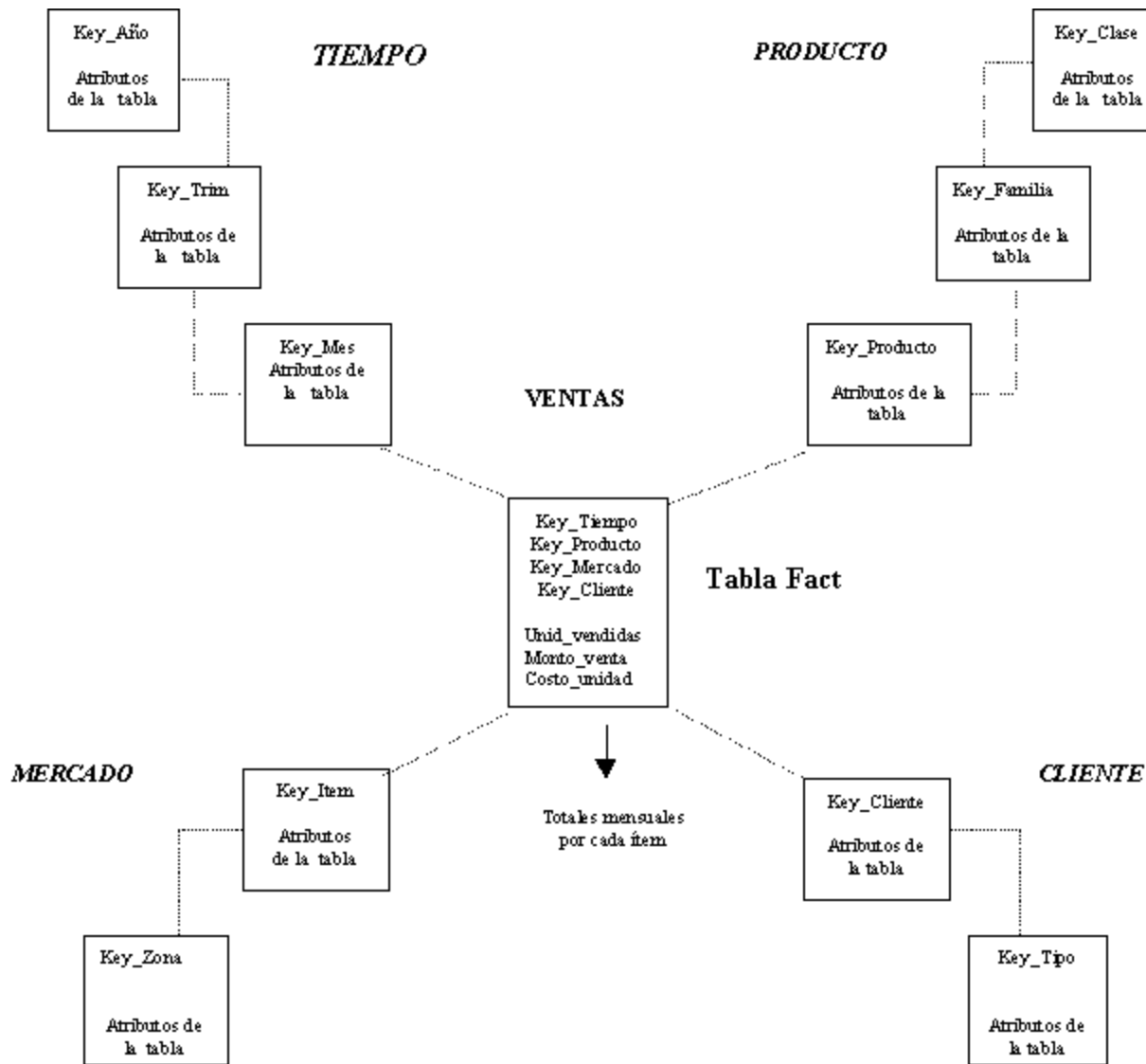


FIGURA 3: EJEMPLO DE UN ESQUEMA SNOWFLAKE PARA EL EJEMPLO ANTERIOR.

1.2.3 Tabla Fact o de Hechos.

Es la tabla central en un esquema dimensional. Es en ella donde se almacenan las mediciones numéricas del negocio. Estas medidas se hacen sobre el grano, o unidad básica de la tabla.

El grano o la granularidad de la tabla queda determinada por el nivel de detalle que se almacenará en la tabla. Por ejemplo, para el caso de producto, mercado y tiempo antes visto, el grano puede ser la cantidad de madera vendida 'mensualmente'. El grano revierte las unidades atómicas en el esquema dimensional.

Cada medida es tomada de la intersección de las dimensiones que la definen. Idealmente está compuesta por valores numéricos, continuamente evaluados y aditivos. La razón de estas

características es que así se facilita que los miles de registros que involucran una consulta sean comprimidos en unas pocas líneas en un set de respuesta.

La clave de la tabla fact recibe el nombre de clave compuesta o concatenada debido a que se forma de la composición (o concatenación) de las llaves primarias de las tablas dimensionales a las que está unida.

Así entonces, se distinguen dos tipos de columnas en una tabla fact: columnas fact y columnas key. Donde la columna fact es la que almacena alguna medida de negocio y una columna key forma parte de la clave compuesta de la tabla.

1.2.4 Tablas Lock-up o Dimensionales

Estas tablas son las que se conectan a la tabla fact, son las que alimentan a la tabla fact. Una tabla lock_up almacena un conjunto de valores que están relacionados a una dimensión particular. Tablas lock_up no contienen hechos, en su lugar los valores en las tablas lock_up son los elementos que determinan la estructura de las dimensiones. Así entonces, en ellas existe el detalle de los valores de la dimensión respectiva.

Una tabla lock_up está compuesta de una primary key que identifica unívocamente una fila en la tabla junto con un conjunto de atributos, y dependiendo del diseño del modelo multidimensional puede existir una foreign key que determina su relación con otra tabla lock_up.

Para decidir si un campo de datos es un atributo o un hecho se analiza la variación de la medida a través del tiempo. Si varía continuamente implicaría tomarlo como un hecho, caso contrario será un atributo.

Los atributos dimensionales son un rol determinante en un DDW. Ellos son la fuente de todas las necesidades que debieran cubrirse. Esto significa que la base de datos será tan buena como lo sean los atributos dimensionales, mientras más descriptivos, manejables y de buena calidad, mejor será el DDW.

1.3 Pasos básicos del Modelamiento Multidimensional

1. Decidir cuáles serán los procesos de negocios a modelar, basándose en el conocimiento de éstos y de los datos disponibles. Ejemplo: Gastos realizados por cada mercado para cada ítem a nivel mensual. Productos vendidos por cada mercado según el precio en cada mes.
2. Decidir el Grano de la tabla Fact de cada proceso de negocio.
Ejemplo : Producto x mercado x tiempo. En este punto se debe tener especial cuidado con la magnitud de la base de datos, con la información que se tiene y con las preguntas que se quiere responder. El grano decidirá las dimensiones del DDW. Cada dimensión debe tener el grano más pequeño que se pueda puesto que las preguntas que se realicen necesitan cortar la base en caminos precisos (aunque las preguntas no lo pidan explícitamente).
3. Decidir las dimensiones a través del grano. Las dimensiones presentes en la mayoría de los DDW son: tiempo, mercado, producto, cliente. Un grano bien elegido determina la dimensionalidad primaria de la tabla fact. Es posible usualmente agregar dimensiones adicionales al grano básico de la tabla fact, donde estas dimensiones adicionales toman un solo valor para cada combinación de las dimensiones primarias. Si se reconoce que una dimensión adicional deseada viola el grano por causar registros adicionales a los generados, entonces el grano debe ser revisado para acomodar esta dimensión adicional.
4. Elegir las mediciones del negocio para la tabla fact. Se deben establecer los ítems que quedarán determinados por la clave compuesta de la tabla fact.

1.4 Profundizaciones de Diseño

1.4.1 La Dimensión Tiempo

Virtualmente se garantiza que cada DDW tendrá una tabla dimensional de tiempo, debido a la perspectiva de almacenamiento histórica de la información. Usualmente es la primera dimensión en definirse, con el objeto de establecer un orden, ya que la inserción de datos en la base de datos multidimensional se hace por intervalos de tiempo, lo cual asegura un orden implícito.

1.4.2 Dimensiones que varían lentamente en el tiempo

Son aquellas dimensiones que se mantienen “casi” constantes en el tiempo y que pueden preservar la estructura dimensional independiente del tiempo, con sólo agregados menores relativos para capturar la naturaleza cambiante del tiempo.

Cuando se encuentra una de estas dimensiones se está haciendo una de las siguientes fundamentales tres elecciones. Cada elección resulta en un diferente grado de seguimiento sobre el tiempo:

Tipo 1: Sobreescribir el viejo valor en el registro dimensional y por lo tanto perder la capacidad de seguir la vieja historia.

Tipo 2: Crear un registro dimensional adicional (con una nueva llave) que permita registrar el cambio presentado por el valor del atributo. De esta forma permanecerían en la base tanto el antiguo como el nuevo valor del registro con lo cual es posible segmentar la historia de la ocurrencia.

Tipo 3: Crear un campo “actual” nuevo en el registro dimensional original el cual almacene el valor del nuevo atributo, manteniendo el atributo original también. Cada vez que haya un nuevo cambio en el atributo, se modifica el campo “actual” solamente. No se mantiene un registro histórico de los cambios intermedios.

1.4.3 Niveles

Un nivel representa un nivel particular de agregación dentro de una dimensión; cada nivel sobre el nivel base representa la sumarización total de los datos desde el nivel inferior. Para un mejor entendimiento, veamos el siguiente ejemplo: consideremos una dimensión Tiempo con tres niveles: Mes, Semestre, Año. El nivel Mes representa el nivel base, el nivel Semestre representa la sumarización de los totales por Mes y el nivel Año representa la sumarización de los totales para los Semestres.

Agregar niveles de sumarización otorga flexibilidad adicional a usuarios finales de aplicaciones EIS/ DSS para analizar los datos.

1.4.4 Sobre Jerarquías

A nivel de dimensiones es posible definir jerarquías, las cuales son grupos de atributos que siguen un orden preestablecido.

Una jerarquía implica una organización de niveles dentro de una dimensión, con cada nivel representando el total agregado de los datos del nivel inferior. Las jerarquías definen cómo los datos son sumarizados desde los niveles más bajos hacia los más altos. Una dimensión típica soporta una o más jerarquías naturales. Una jerarquía puede pero no exige contener todos los valores existentes en la dimensión .

Se debe evitar caer en la tentación de convertir en tablas dimensionales separadas cada una de las relaciones muchos-a-uno presentes en las jerarquías. Esta descomposición es irrelevante en el

planeamiento del espacio ocupado en disco y sólo dificulta el entendimiento de la estructura para el usuario final, además de destruir el desempeño del browsing.

Bibliografía y Referencias.

- [Carrasc97] "Bases de Datos Multidimensionales", Universidad de Concepción, Preparado por Jorge Carrasco-Alumno Magister en Ciencias de la Computación, Diciembre 1997.
- [Communi98] Communications of the ACM, September 1998, Volume41 - Number9 : Datawarehousing.
- [Compute96A] Computers World N°115 , Septiembre 1996.
- [Compute96B] Computer World N°119, Noviembre 1996.
- [Compute97A] Computer World 11 Junio 1997.
- [Compute97B] Computer World N°135, 25 junio 1997.
- [Compute97C] Computer World N°136, 7 Julio 1997.
- [Compute97D] Computer World N°139, 20 Agosto 1997.
- [Compute97E] Computer World N°145, 12 Noviembre 1997.
- [Compute98] Computer World N°151 18 Febrero 1998.
- [Corey93] "Oracle Data Warehousing", Michael J. Corey & Michael Abbey, Computer World 1993 - pág. 218.
- [Informa96] Revista Informática Volumen 18, 8 Septiembre 1996.
- [Jiménez98] "Introducción Al Datawarehouse", Departamento de Informática de la Universidad De Concepción, Preparado por Claudia Jiménez – Docente Departamento.
- [Mcguff1] "Designing The Perfect Datawarehouse", Frank McGuff, <http://www.techguide.com/>
- [Mcguff2] "Datawarehouse Modeling", Frank McGuff, <http://www.techguide.com/>
- [MicroSt96] "Data Warehousing, Data Modeling and Design", MicroStrategy Education, Nov 96
- [Oviedo98] "Diseñando Un Datawarehouse", Taller De Computación - Tópico: Bases De Datos, Depto. Ingeniería Informática Y Ciencias De La Computación - Facultad De Ingeniería - Universidad De Concepción, Preparado por Rodrigo Oviedo
- [Paper1] "Managing The Warehouse Throughout Its Lifecycle", <http://www.techguide.com/>
- [Paper2] "Building A Decision Support Architecture For Datawarehousing", <http://www.techguide.com/>
- [Paper3] "Putting Metadata To Work In The Warehouse" , <http://www.techguide.com/>
- [Paper4] "Enterprise Storage: Delivering Data Warehousing Business Results", <http://www.techguide.com/>
- [Paper5] "A Practical Guide To Getting Started With Data Warehousing", <http://www.techguide.com/>
- [Santoro97] "Metodología de Desarrollo para la Construcción de un Datawarehouse bajo un enfoque Bottom Up", .Memoria para optar al título de: Ingeniero Ejecución en Informática,. Universidad Técnica Federico Sta. María, Departamento de Informática, Autores: Andrea Santoro Rojas, Luis Humberto Ponce Trujillo, Enero 1997.
- [Solucio96] Especial de Datawarehouse, Revista Soluciones Avanzadas, Edición Junio 1996 - Pág. 40-72
- [Softwar96] "Executive Briefing for the Datawarehouse", Software AG Education Services, January 1996
- [Wolff98A] "Datawarehousing", Taller De Computación - Tópico: Bases De Datos, Depto. Ingeniería Informática Y Ciencias De La Computación - Facultad De Ingeniería - Universidad De Concepción, Desarrollado por Carmen Gloria Wolff..
- [Wolff98B] "Consideraciones para Enfrentar el Desarrollo de un DW", Asignatura Gestión Proyectos de Ingeniería de Software (GPIS), Depto. Ingeniería Informática Y Ciencias De La Computación - Facultad De Ingeniería -, Preparado por Carmen Gloria Wolff.

Direcciones Internet Recomendadas

- <http://www.techguide.com/>
- <http://pwp.starnetinc.com/savmony.html>
- <http://www.people.memphis.edu/~tsakagch/dw-web.htm>
- <http://www.datamation.com>
- <http://www.dmreview.com>
- <http://www.dbdp.com>
- <http://www.dbmsmag.com>