Sistemas de Información Orientados a la Toma de Decisiones: el enfoque multidimensional

En el presente artículo se describe el modelo de datos multidimensional y las tecnologías OLAP, surgidos en el ámbito de los Sistemas Orientados a la Toma de Decisiones, como consecuencia de la necesidad de analizar los enormes volúmenes de información almacenados actualmente en las empresas. Asimismo, se realiza una comparativa con los modelos relacionales, indicando cuál es la aportación de este nuevo modelo.

Isabel Dapena Bosquet
Ingeniera Informática del ICAI (Promoción 2001). En 2002 ingresó en el Instituto de Investigación Tecnológica como Investigadora en Formación, donde desarrolla su actividad en el Área de Sistemas Inteligentes.

Antonio Muñoz San Roque
Dr. Ingeniero Industrial del ICAI (Promoción 1991). Profesor Propio Agregado de la ETS de Ingeniería (ICAI) de la Universidad Pontificia Comillas de Madrid, donde actualmente es Director del Departamento de Electrónica y Automática.

Álvaro Sánchez Miralles
Dr. Ingeniero Industrial del ICAI (Promoción 1998). Profesor de la ETS de Ingeniería (ICAI) de la Universidad Pontificia Comillas de Madrid.

Las bases de datos surgieron con el objetivo de organizar la información de acuerdo a una estructura de datos coherente, que evitara las inconsistencias y problemas que surgen al actualizar o borrar datos en los sistemas basados en ficheros. Dicha estructura de datos, junto con el resto de los objetos, relaciones y restricciones existentes entre ellos, forman el llamado *modelo de datos* [9], el cual ha ido evolucionando y adaptándose a la complejidad del mundo real que representa.

La recolección y almacenamiento de datos relativos a productos, clientes, operaciones, transacciones on-line, etc. ha experimentado un gran desarrollo gracias a los avances en hardware, permitiendo el almacenamiento masivo (terabytes) y estable de los mismos. En consecuencia, las empresas disponen de una gran cantidad de datos que han de ser transformados en información sobre la que basar su estrategia de actuación.

Las técnicas y herramientas de los *Sistemas de Ayuda a la Toma de Decisiones* (DSS) que se han desarrollado para disponer de una información completa y de las herramientas de análisis necesarias son, por un lado, lo que se ha llamado Almacén de Datos (en inglés, *DataWarehouse*) y por otro, las herramientas de Minería de Datos (*Data Mining*) y herramientas OLAP (*On-Line Analytical Processing*).

Tipos de aplicaciones de los Sistemas de Ayuda a la Toma de Decisiones

Se pueden establecer tres categorías de aplicaciones orientadas al análisis del negocio:

- Generadores de informes: suelen ser las que tienen menos requerimientos de análisis. Se basan en modelos relacionales y utilizan SQL.
- Consultas *ad-hoc*: tienen mayor interacción con el usuario al proporcionar diferentes técnicas de navegación y selección de los datos. Están diseñadas sobre modelos relacionales.
- Herramientas analíticas: generan nuevos resultados a partir de complejos cálculos sobre varias dimensiones. A este grupo pertenecen los sistemas de predicción y de planificación mediante el análisis de escenarios.

En este artículo se presenta en primer lugar el concepto de *DataWarehouse* y posteriormente las características de los dos modelos de datos más utilizados en la actualidad: el modelo relacional, que es el más ampliamente extendido en los sistemas de gestión de bases de datos (SGBD) y el modelo multidimensional, que facilita el desarrollo de las aplicaciones de análisis de información.

DataWarehouse

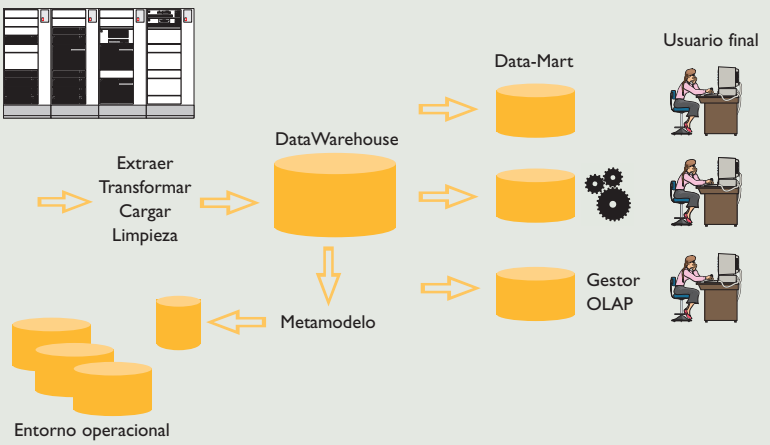
Los almacenes de datos o *DataWarehouse* recogen los datos de los distintos entornos operacionales de la empresa, los filtran y procesan para su almacenamiento, proporcionando una plataforma sólida de datos consolidados e históricos para el posterior análisis del negocio, tal y como se muestra en la Figura 1.

El *DataWarehouse* responde así a la compleja necesidad de obtención de información útil sin el sacrificio del rendimiento de las aplicaciones operacionales, debido a lo cual se ha convertido actualmente en una de las tendencias tecnológicas más significativas en los niveles de gestión dentro de la organización.

Por su parte el *DataMart* es una visión parcial del DataWarehouse enfocada a un departamento o área específica, como por ejemplo los departamentos de Finanzas o Marketing, permitiendo así un mejor control de la información que se está abarcando.

Una definición muy extendida de *DataWarehouse* es la realizada en [6], donde se define como una colección de datos no volátil, integrada, temporal y orientada al tema, usada principalmente para la toma de

Figura 1. Arquitectura de un *DataWarehouse*



decisiones. De esta definición se desprenden las siguientes características:

- Orientación al tema: la información se clasifica en base a los aspectos que son de interés para el analista o usuario final (por ejemplo: producto, cliente, proveedor).
- La integración de los datos se manifiesta en convenciones de nombres consistentes, en la medida uniforme de variables o en la confluencia de múltiples fuentes de datos.
- Temporal: la mayoría de los datos almacenados están referidos a un instante o período de tiempo, quedando almacenado un histórico.
- No volátil: Una vez almacenados los datos, éstos no son modificados.

El modelo de datos relacional

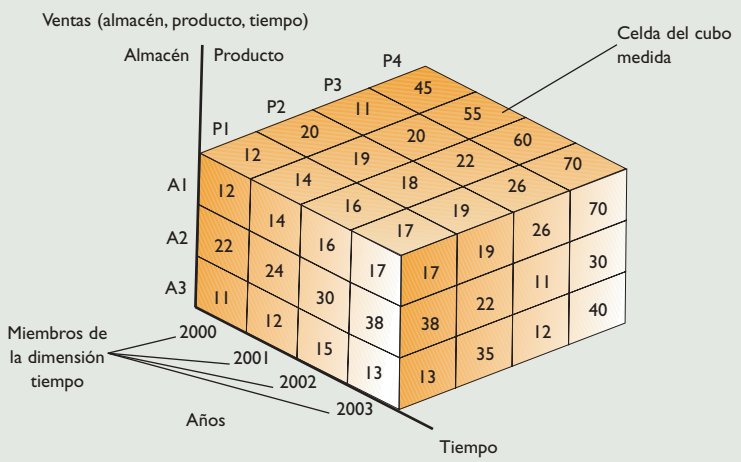
En la década de los 70, E. F. Codd [4] enunció las doce reglas que debía cumplir un modelo de datos relacional, estableciendo las bases para la estandarización del mismo. El objetivo inicial de este modelo consistía en aislar al usuario (o aplicación) de la estructura física de los datos. Textualmente en [4]: "…*se propone un modelo relacional de datos como una base para proteger a los usuarios de los cambios que potencialmente pueden alterar la representación de los datos, causados por el crecimiento del banco de datos y por los cambios en los caminos de acceso.*"

Las características del modelo relacional han hecho que prácticamente todos los SGBD comerciales implementen este modelo (como por ejemplo Oracle, DB2, Sysbase, Informix, …) y por extensión que se haya establecido en los entornos operacionales y transaccionales de todas las empresas, donde se aplica el paradigma *On-Line Transaction*

Tabla 1. Diferencias entre Entornos Operacionales y Soporte a Decisiones

Concepto	Sistemas Operacionales (OLTP)	Sistemas de Ayuda a la Toma de Decisiones (OLAP)
Datos	Valores actuales	Datos históricos y/o calculados
Organización	Por aplicación	Por áreas de la empresa
Acceso	Muy frecuente (lectura/escritura)	Baja frecuencia
Actualizaciones	Actualizaciones de campos	No se actualiza. Se manipula
Tiempo de respuesta	Medido por el tiempo de la transacción (del orden de segundos)	Medido por el tiempo de la consulta (del orden de minutos)
Tamaño de la BD	100 MB-GB	100 GB-TB
Usuarios	Miles	Cientos
Unidad de trabajo	Transacciones	Consultas complejas

Figura 2. Definición de los elementos del hipercubo



Processing (OLTP) diseñado para una eficiente selección, almacenamiento y consulta de datos en procesos en los que el usuario requiere una respuesta inmediata. Ejemplos de transacciones son: transferir dinero entre cuentas, un cargo o abono, una devolución de inventario, etc.

Este modelo de datos se basa en la teoría de conjuntos y la estructura de datos que maneja es la *relación* o tabla bidimensional. Cada elemento de la relación se denomina tupla, y está constituido por un conjunto de n valores ordenados. Cada uno de estos n valores se corresponde con el atributo Ai de la relación, el cual toma valores de un dominio Di. De tal manera que una relación R se representa por $R(A_1, A_2, \ldots, A_N)$, tal que $R \subset D_1 x D_2 x \ldots x D_N$.

Modelo multidimensional

A pesar de los buenos resultados obtenidos con el modelo relacional en los sistemas operacionales, la utilización de este modelo en aplicaciones orientadas a la toma de decisiones presenta varias carencias, como se cita en [13].

Una de las principales carencias es el bajo rendimiento de las consultas: el modelo relacional está orientado a transacciones que manejan pocos registros simultáneamente, mientras que los sistemas de ayuda a la toma de decisiones (DSS) tienden a procesar grandes volúmenes de datos. Otra de las limitaciones es la propia estructura de la base de datos: las consultas realizadas en los DSS son muy complejas y su definición no está fijada de antemano. Como las consultas dependen de lo que necesite el usuario en cada momento, con un modelo relacional se debería generar un índice por cada posible consulta que desee el usuario, lo que dificulta la gestión y mantenimiento de la base de datos.

En la Tabla 1 se presenta una comparación de los requisitos de los sistemas operacionales y de los Sistemas Orientados a la Toma de Decisiones.

Dados los requisitos propios de los DSS es necesario un nuevo modelo de datos: el modelo multidimensional. Éste ofrece grandes ventajas sobre el modelo relacional siempre que se trabaje sobre datos agregados, totales, subtotales, series temporales y diversos grados de detalle de los datos. En resumen, es un modelo adecuado para el estudio a alto nivel de los datos, al ofrecer una mayor flexibilidad y rapidez de acceso para el análisis de los mismos [11] y [12].

Por otra parte, si lo que se quiere es acceder a un dato individual básico como puede ser el importe de una operación concreta, la ventaja del modelo multidimensional desaparece en favor del relacional. Éstos son capaces de recuperar un dato individual con mayor eficiencia que las multidimensionales y, dada su utilización masiva en sistemas OLTP, están optimizados para la inserción de registros y el control concurrente de usuarios.

En el modelo de datos multidimensional, los datos se organizan en torno a los conceptos de la empresa y la estructura de datos manejada en este modelo son matrices multidimensionales o *hipercubos*. Un *hipercubo* consiste en un conjunto de celdas, de tal manera que cada una está identificada por la combinación de los miembros de las diferentes dimensiones y contiene el valor de la medida analizada para dicha combinación de dimensiones.

Las variables o medidas son aquellas características del negocio que pueden ser cuanti-

ficadas y son seleccionadas para el análisis. Por ejemplo: ventas, compras, costes,… Se corresponden con los datos numéricos. Los valores que toman las variables son el resultado de las diferentes combinaciones posibles de los miembros de las dimensiones sobre las que se definen.

Las dimensiones se definen como los atributos categóricos que caracterizan a una variable. Por ejemplo: producto, cliente, vendedor, geografía, tiempo,… El tiempo se considera una dimensión más del modelo, puesto que los análisis de las cifras de un negocio se especifican para un periodo de tiempo concreto. Las dimensiones se utilizan para seleccionar y agrupar los datos al nivel de detalle deseado. Los miembros de la dimensión se caracterizan por organizarse en una o varias jerarquías, las cuales presentan distintos niveles que permiten realizar operaciones de agregación y desagregación (*roll-up* y *drill-down*).

En la Figura 2 se puede observar la estructura de un *hipercubo*. El cubo representa la variable "Ventas" sobre las dimensiones "Almacén", "Producto" y "Tiempo".

Destacan dos tendencias en cuanto a modelos de datos multidimensionales como se indica en [8] y [10]: modelos puramente multidimensionales o hipercubos [1] y [2] y modelos basados en relaciones [7].

Los modelos basados en relaciones se aproximan a la solución adoptada en bases de datos estadísticas, donde se diferencian los atributos categóricos (dimensiones) de los numéricos.

Uno de los modelos relacionales más característico es el *Star-Schema* o modelo en estrella [7], en el que los atributos dimensionales o categóricos se agrupan en las tablas de dimensiones, cada una con una clave primaria y los atributos numéricos o medidas dan lugar a la tabla de hechos o *Fact Table* (ver Figura 3). Los atributos dimensionales suelen ser de tipo texto y se utilizan para calificar las posibles consultas que se realicen. Por otro lado, cada fila de la tabla de hechos es una medida tomada para la intersección de todas las dimensiones que la definen, estableciéndose una relación muchos-a-muchos entre la tabla de hechos y las tablas de dimensiones. Para el ejemplo de *hipercubo* de la Figura 2, su modelo en estrella sería el siguiente: cada eje del cubo se corresponde con una tabla dimensión (tablas Producto, Cliente, Tiempo y Almacén) y el contenido del cubo con la tabla de hechos (tabla Ventas).

Figura 3. Ejemplo de "Star-Schema"

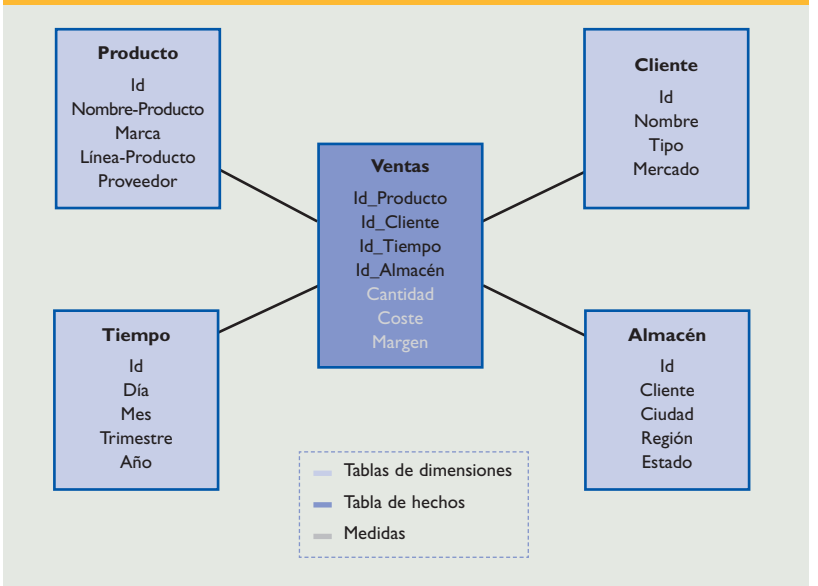


Figura 4. Ejemplo de Snowflacking

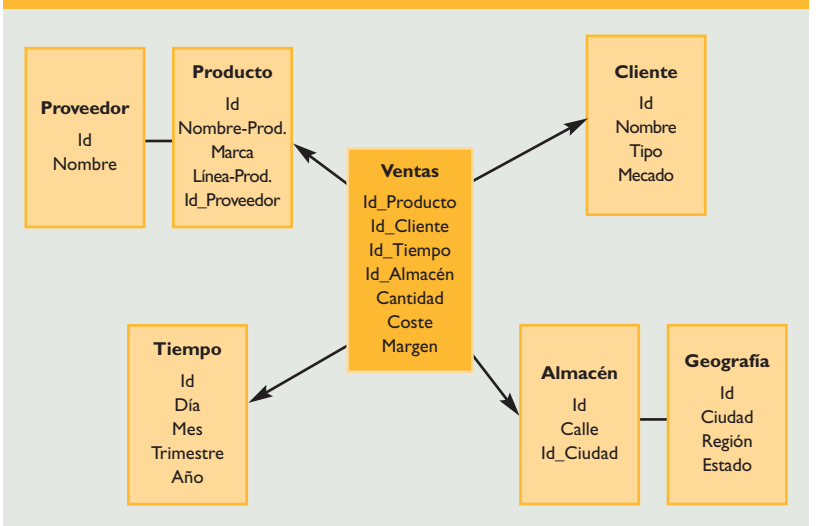


Figura 5. Operaciones roll-up/drill-down a lo largo de la dimensión

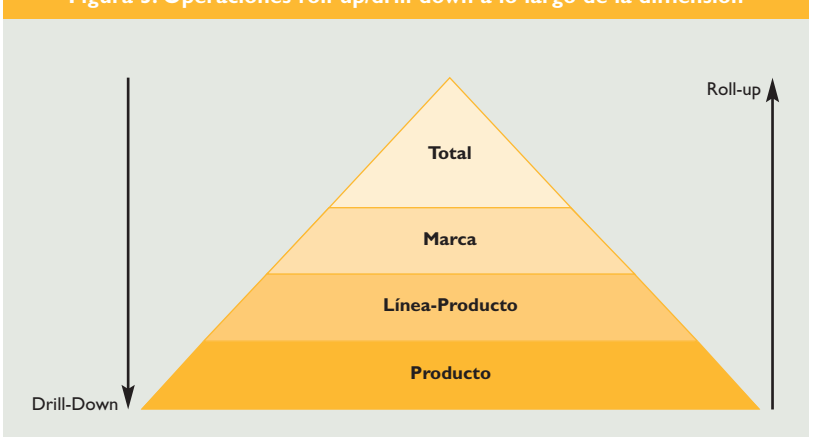


Tabla 2. Ejemplo del operador CUBE

Prod	Alm	Año	Total Ventas
P6	A1	2000	18
P6	A1	Todos los años	18
P6	A2	2001	31
P6	A2	2002	40
P6	A2	Todos los años	71
P6	Todos los almacenes	2000	18
P6	Todos los almacenes	2001	31
P6	Todos los almacenes	2002	40
P6	**Todos los almacenes**	**Todos los años**	**89**
P7	A1	2000	46
P7	A1	Todos los años	46
P7	A2	2000	36
P7	A2	2001	29
P7	A2	Todos los años	65
P7	A3	2000	48
P7	A3	Todos los años	48
P7	Todos los almacenes	2000	130
P7	Todos los almacenes	2001	29
P7	**Todos los almacenes**	**Todos los años**	**159**
P8	A1	2000	11
P8	A1	Todos los años	11
P8	A2	2000	12
P8	A2	Todos los años	12
P8	A3	2000	18
P8	A3	2001	19
P8	A3	2002	24
P8	A3	Todos los años	61
P8	Todos los almacenes	2000	41
P8	Todos los almacenes	2001	19
P8	Todos los almacenes	2002	24
P8	**Todos los almacenes**	**Todos los años**	**84**
Todos los productos	A1	2000	75
Todos los productos	A1	Todos los años	75
Todos los productos	A2	2000	48
Todos los productos	A2	2001	60
Todos los productos	A2	2002	40
Todos los productos	A2	Todos los años	148
Todos los productos	A3	2000	66
Todos los productos	A3	2001	19
Todos los productos	A3	2002	24
Todos los productos	A3	Todos los años	109
Todos los productos	Todos los almacenes	2000	189
Todos los productos	Todos los almacenes	2001	79
Todos los productos	Todos los almacenes	2002	64
Todos los productos	Todos los almacenes	Todos los años	332

Por cuestión de rendimiento y por simplificación de las consultas es preferible que las tablas de dimensiones no se normalicen. El proceso de normalización recibe el nombre de *snowflacking* por el que una tabla de dimensiones se descompone en una o varias. Sobre el modelo de la figura anterior se ha desnormalizado la tabla "Producto" y la tabla "Almacén".

Consultas

Las operaciones más comunes realizadas sobre los datos multidimensionales son: *cube* y *roll-up* definidas en [5]. *Cube* calcula todas las posibles agregaciones que resultan de las combinaciones de atributos incluidos en la cláusula de la consulta, generando totales y subtotales para dichas combinaciones de los atributos. Así en el ejemplo de la Tabla 2 se ha aplicado el operador *cube* a las dimensiones "Producto", "Almacén" y "Tiempo" sobre la tabla "Ventas" del ejemplo anterior, obteniéndose subtotales y totales para todos las combinaciones de todos los productos con todos los almacenes y todos los años. Se muestra a continuación el resultado de aplicar este operador por producto, almacén y tiempo. Se ha destacado en negrita lo que vendría a ser el total de ventas por producto. La última fila representa el total de ventas para todos los productos, almacenes y años.

Por su parte, la operación de *roll-up* consiste en la agregación a lo largo de los distintos niveles de la jerarquía de una de las dimensiones. De ahí, la importancia que ha recibido el tratamiento de las jerarquías en todos los modelos multidimensionales, ya que determina la posibilidad de realizar las diferentes agregaciones (ver Figura 5).

En el siguiente ejemplo de la Tabla 3, se muestra el resultado de aplicar el operador *roll-up* sobre la dimensión "Producto": totales y subtotales para la jerarquía definida en la dimensión "Producto".

Implementación física del modelo multidimensional

Existen dos tendencias para la implementación física del modelo multidimensional: *Multidimensional OLAP* (MOLAP) y *Relational OLAP* (ROLAP). En la primera, (MOLAP) la información se almacena directamente en matrices multidimensionales. Generalmente son bases de datos propietarias, esto es, cada empresa de software propone su mejor solución OLAP. Bajo esta arquitectura, el

Tabla 3. Ejemplo de "roll-up" sobre la jerarquía de la dimensión "Producto"

Línea de producto	Marca	Suma-Ventas	Línea de producto	Marca	Suma-Ventas
CD	Sony	113	Móvil	Panasonic	159
CD	Todas las marcas	113	Móvil	Siemens	84
Dect	Panasonic	113	Móvil	Todas las marcas	332
Dect	Siemens	63	TV	Loewe	218
Dect	Todas las marcas	176	TV	Sony	240
DVD-Recorder	Loewe	104	TV	Todas las marcas	458
DVD-Recorder	Panasonic	22	TV-Plasma	Loewe	131
DVD-Recorder	Todas las marcas	126	TV-Plasma	Todas las marcas	131
Monitor	Sony	205	Video-recorder	Loewe	190
Monitor	Todas las marcas	205	Video-recorder	Panasonic	55
Móvil	Motorola	89	Video-recorder	Todas las marcas	245
Todas las líneas de producto			**Todas las marcas**		**1786**

diseñador de la base de datos debe especificar las agregaciones necesarias, las cuales son calculadas y almacenadas, reduciéndose el tiempo de respuesta para las consultas on-line. La arquitectura se presenta en dos capas: capa de base de datos, que realiza la manipulación de los datos e incluye parte de la lógica de la aplicación para ejecutar las consultas OLAP y capa de presentación, que muestra los datos al usuario. Un ejemplo de esta implementación es Arbor Essbase [14].

La arquitectura ROLAP propone acceder directamente al *DataWarehouse* implementado sobre una base de datos relacional. Las operaciones sobre los datos se traducen en consultas SQL, manejadas por un gestor OLAP propio. Con el fin de mejorar el rendimiento se utilizan índices y rutinas de cálculo de agregaciones. En este caso la arquitectura es de tres capas: capa de base de datos; gestor OLAP, que se encarga de la ejecución de las consultas OLAP e interfaz de aplicación. Como ejemplo de soluciones ROLAP cabría citar a MicroStrategy [15] y Oracle 9i [16].

Conclusiones

En el presente artículo se han presentado las dos tendencias de sistemas de información imperantes actualmente en la empresa: el enfoque transaccional de los entornos de operación donde ha triunfado el modelo relacional y los sistemas de *Bussiness Intelligence* orientados a la toma de decisiones, en los que dada la complejidad de las consultas y las operaciones realizadas sobre los datos se utiliza el modelo multidimensional. ▮

Bibliografía

[1] R. Agrawal, A. Gupta, S. Sarawagi. "Modeling Multidimensional Databases". Proc. 13th Int. Conf. Data Engineering, ICDE. 1997.

[2] L. Cabibbo, R. Torlone. "A logical Approach to Multidimensional". Proceedings of the 6th International Conference on Extending Database Technology (EDBT-98), Valencia, Spain.

[3] E.F. Codd, S.B. Codd and C.T. Salley. "Providing OLAP to User-Analysts: An IT Mandate". 1993.

[4] E. F. Codd. "A relational model of data for large shared data banks." Communications of the ACM, v.13 n.6, p.377-387, June 1970.

[5] J. Gray, S. Chaudhuri, A. Bosworth, et al. "Data Cube: A Relational Aggregation Operator Generalizing Group By, Cross-Tab, and Sub-Totals". Data Mining and Knowledge Discovery (1997).

[6] W. H. Inmon. John Wiley & Sons. "Building the DataWarehouse." Second edition, 1996.

[7] R. Kimball. "The DataWarehouse Toolkit." Wiley Computer Publishing, 1996.

[8] P. Marcel. "Modeling and querying multidimensional databases: an overview." Networking and Information Systems Journal. Vol 2, nº5, 1999.

[9] Ullman, Jeffrey y Widom, Jennifer. "Introducción a los Sistemas de Bases de Datos". Editorial Prentice Hall, México 1999, ISBN: 970-17-0256-5.

[10] Vassiliadis, Panos. "DataWarehouse Modeling and Quality Issues". Ph D thesis, Universidad de Atenas (Grecia). Junio 2000.

[11] "The Case for Relational OLAP. A White Paper Prepared by MicroStrategy, Inc." 1995.

[12] OLAP Council White Paper. 1997.

[13] "Designing the Datawarehouse on Relational Databases." Stanford Technology Group Informix Inc. Articles. Pages 1-14.

[14] Recursos de la dirección web: http://www.arborsoft.com

[15] web: http://www.microstrategy.com

[16] Recursos de la dirección web: http://www.oracle.com