

El modelo multidimensional

Data Warehousing

OLAP vs. OLTP

Data Warehousing

El modelo multidimensional

Dimensiones, medidas y hechos

Arquitectura de un data warehouse

Funcionamiento detallado de un data warehouse

Implementación de un data warehouse

MOLAP

ROLAP: Esquema en estrella, copo de nieve y constelaciones de hechos

Bibliografía

- Ralph Kimball:
“The Data Warehouse Toolkit:
Practical Techniques for Building Dimensional Data Warehouses”
John Wiley & Sons, 1996. ISBN 0-471-15337-0
- Jiawei Han & Micheline Kamber
“Data Mining: Concepts and Techniques”
Morgan Kaufmann, 2000. ISBN 1-55860-489-8
- C.J. Date:
“Introducción a los sistemas de bases de datos”
Pearson Educación, 2001. ISBN 968-444-419.
- Ramez A. Elmasri & Shamkant B. Navathe:
“Fundamentos de Sistemas de Bases de Datos”.
Addison-Wesley, 2002 [3ª edición]. ISBN 84-782-9051-6.
- Henry F. Korth, Abraham Silberschatz & S. Sudarshan:
“Fundamentos de Bases de Datos”.
Mc-Graw Hill, 2002 [4ª edición]. ISBN 84-481-3654-3.

OLAP vs. OLTP

OLTP [On-Line Transaction Processing]

Aplicaciones típicas de gestión

- ✓ Tareas repetitivas.
- ✓ Tareas muy bien estructuradas.
- ✓ Transacciones cortas (actualizaciones generalmente).

Prioridad: Gestión de transacciones

Las transacciones se realizan sobre grandes *bases de datos* a las cuales se puede acceder eficientemente mediante índices, ya que cada operación afecta sólo a unos pocos registros.

Es de vital importancia garantizar la “acidez” de las transacciones (atomicidad, consistencia, aislamiento y durabilidad).

OLAP [On-Line Analytical Processing]

Sistemas de ayuda a la decisión (DSS):

- ✓ Consultas muy complejas (muchos datos y funciones de agregación).
- ✓ Actualizaciones poco frecuentes.

Prioridad: Procesamiento de consultas

Los *data warehouses* (DW) almacenan datos resumidos de tipo histórico. La optimización de las consultas y el tiempo de respuesta son primordiales.

Data warehousing

Problema

Las organizaciones manejan enormes cantidades de datos...

... en distintos formatos.

... que residen en distintas bases de datos.

... organizados utilizando distintos tipos de gestores de bases de datos

Consecuencia

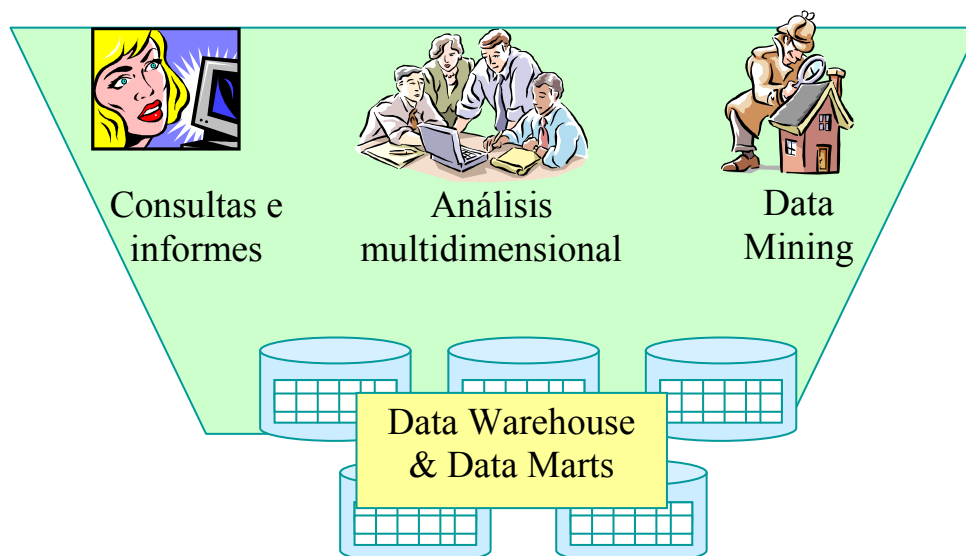
Resulta difícil acceder y utilizar todos los datos en aplicaciones de análisis (las cuales requieren extraer, preparar e integrar los datos)

Data Warehousing

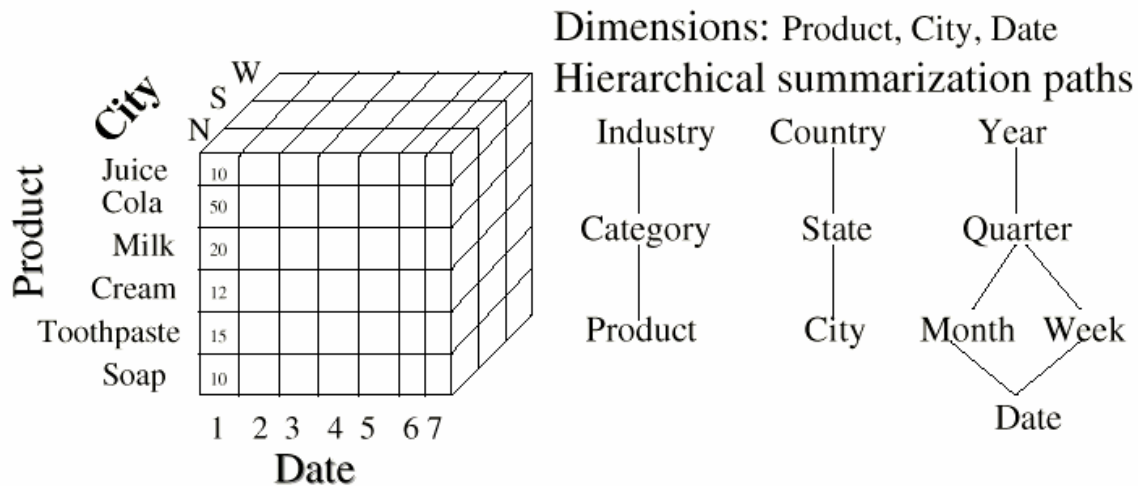
Diseño de procesos e implementación de herramientas que proporcionen información completa, oportuna, correcta y entendible en la toma de decisiones.

Técnicas utilizadas en data warehousing

- ✓ Acceso a fuentes de datos heterogéneas:
 - Limpieza de datos
 - Filtrado de datos
 - Transformación de datos
- ✓ Almacenamiento de los datos: Estructura de datos multidimensional:



El modelo multidimensional



Los datos en un DW se modelan en *data cubes* (“cubos de datos” sería su traducción literal), estructuras multidimensionales (hipercubos, en concreto) cuyas operaciones más comunes son:

- ✓ *Roll up* (incremento en el nivel de agregación de los datos).
- ✓ *Drill down* (incremento en el nivel de detalle, opuesto a roll up).
- ✓ *Slice* (reducción de la dimensionalidad de los datos mediante selección).
- ✓ *Dice* (reducción de la dimensionalidad de los datos mediante proyección).
- ✓ *Pivotaje o rotación* (reorientación de la visión multidimensional de los datos).

Modelado multidimensional

Modelos de datos como conjuntos de medidas descritas por dimensiones.

- Adecuado para resumir y organizar datos (p.ej. hojas de cálculo).
- Enfocado para trabajar sobre datos de tipo numérico.
- Más simple: Más fácil de visualizar y entender que el modelado E/R.

Dimensiones

Perspectivas o entidades respecto a las cuales una organización quiere mantener sus datos organizados (p.ej. tiempo, localización, clientes, proveedores...)

Miembros

Nombres o identificadores que marcan una posición dentro de la dimensión.

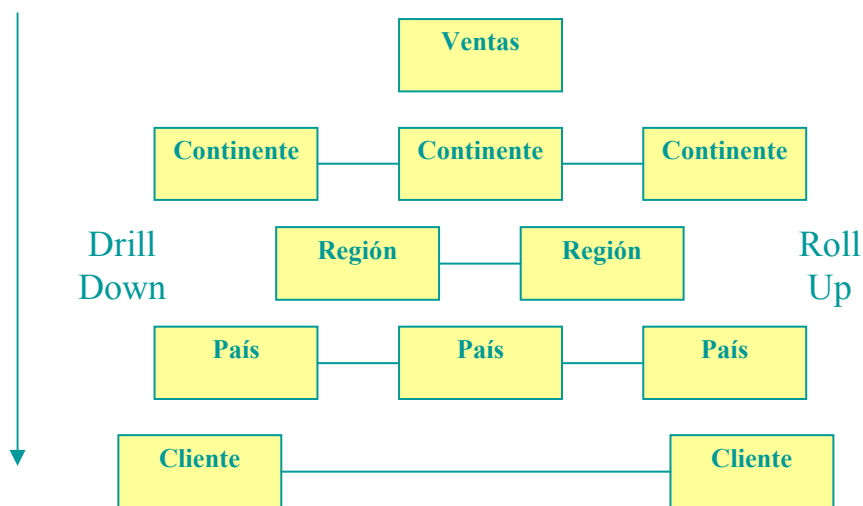
Ejemplos:

Meses, trimestres y años son miembros de la dimensión tiempo.

Ciudades, regiones y países son miembros de la dimensión localización.

Jerarquías

Los miembros de las dimensiones se suelen organizar en forma de jerarquías.



Organización jerárquica de las dimensiones

Hechos

Colecciones de datos relacionados compuestas por medidas y un contexto.

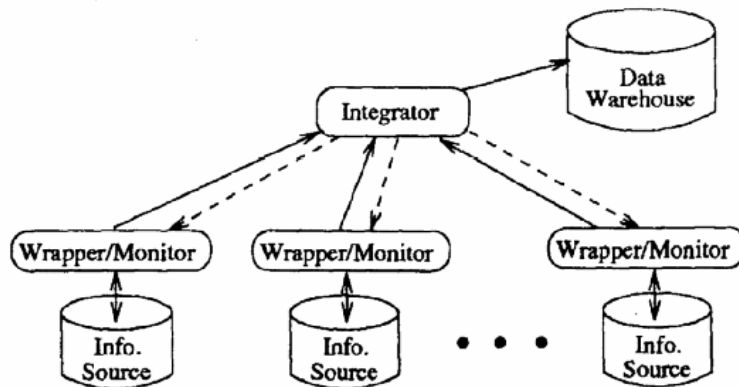
- Las dimensiones determinan el contexto de los hechos.
- Cada hecho particular está asociado a un miembro de cada dimensión.

Medidas

Atributos numéricos asociados a los hechos (lo que realmente se mide).

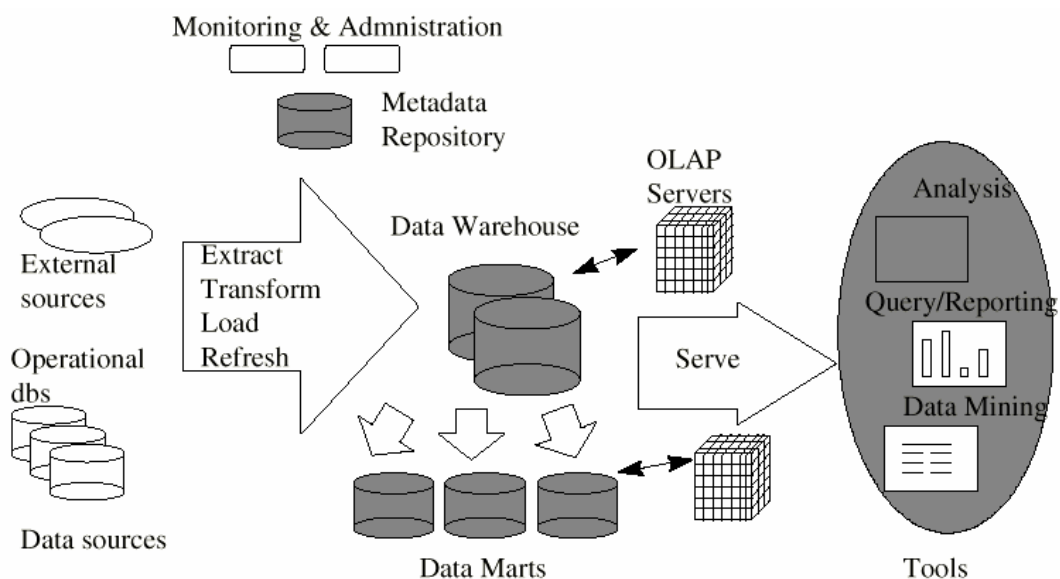
Ejemplos: Volumen de las ventas, coste asociado a un producto, número de transacciones efectuadas, porcentaje de beneficios...

Arquitectura



- Las fuentes de información pueden incluir bases de datos relacionales, bases de conocimiento, documentos en distintos formatos (p.ej. XML)...
- Los wrappers (encapsuladores) se encargan de extraer los datos de las distintas fuentes y transmitirlos al data warehouse.
- Los monitores están en contacto directo con las fuentes de datos para detectar los cambios que se puedan producir en ellas.
- El integrador es el responsable de filtrar, resumir y unificar la información proveniente de las distintas fuentes.

Funcionamiento detallado de un DW



Implementación

El data warehouse se mantiene aparte de las bases de datos operacionales y se puede implementar utilizando servidores OLAP multidimensionales (MOLAP) o servidores OLAP relacionales (ROLAP).

MOLAP (Multidimensional OLAP)

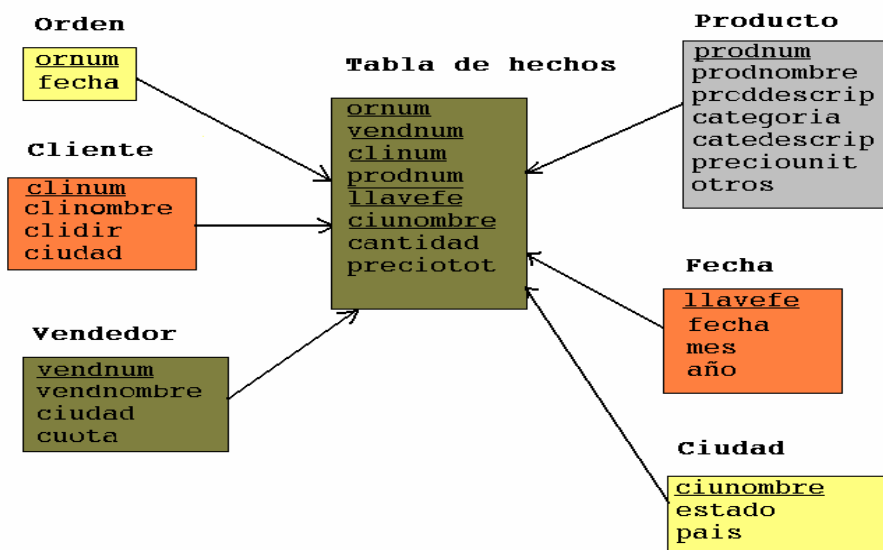
Los datos se almacenan en estructuras de datos multidimensionales (matrices multidimensionales sobre las que se realizan directamente las operaciones OLAP).

ROLAP (Relational OLAP)

El DW se implementa como una base de datos relacional (las operaciones multidimensionales OLAP se traducen en operaciones relacionales estándar).

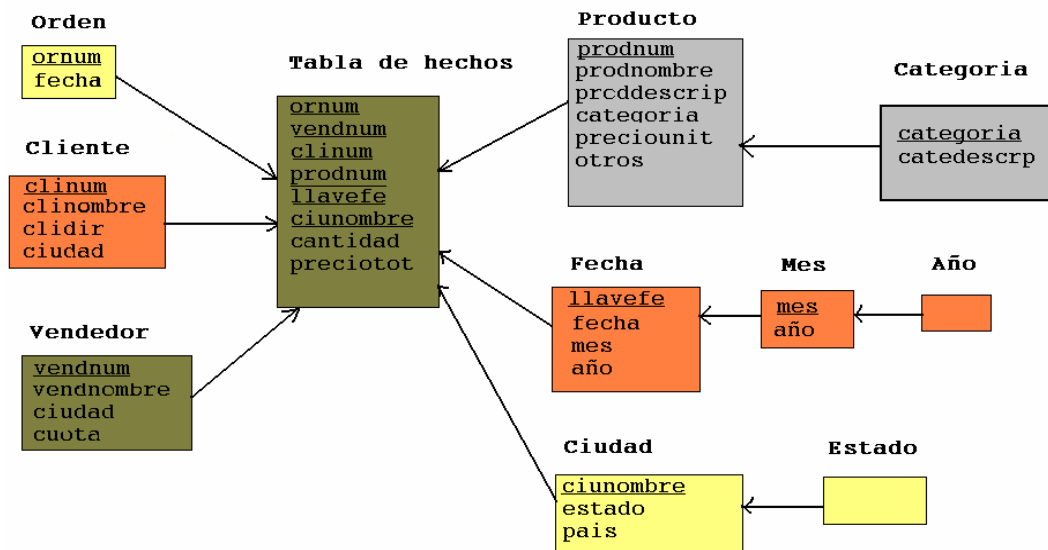
Esquema en estrella (star)

Una tabla de hechos y una tabla adicional por cada dimensión



Esquema en copo de nieve (snowflake)

Refleja la organización jerárquica de las dimensiones



Constelaciones de hechos

Los esquemas en estrella y bola de nieve pueden generalizarse con la inclusión de distintas tablas de hechos que comparten todas o algunas de las dimensiones.