PARTE II:
ALMACENES DE DATOS

** Transparencias basadas parcialmente en el "tutorial DW" de Matilde Celma*

José Hernández Orallo

jorallo@dsic.upv.es

Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia

Temario

1. Introducción
   1.1. Finalidades y Evolución de los Sistemas de Información.
   1.2. Herramientas para la Toma de Decisiones: diferencias e interrelación.
   1.3. Almacenes de Datos, OLAP y Minería de Datos: definición e interrelación.
2. Almacenes de Datos
   2.1. Introducción a los Almacenes de Datos.
   2.2. Arquitectura de un Sistema de Almacén de Datos.
   2.3. Explotación de un Almacén de Datos: Herramientas OLAP.
   2.4. Sistemas ROLAP y MOLAP.
   2.5. Carga y Mantenimiento de un Almacén de Datos.
   2.6. Diseño de un almacén de Datos.
   2.7. Líneas de Investigación Abiertas.
3. Minería de Datos
   3.1. Introducción a la Minería de Datos (DM)
   3.2. El proceso de KDD
   3.3. Técnicas de Minería de Datos
   3.4. Web Mining
   3.5. Líneas de Investigación Abiertas

Objetivos Parte II

- Conocer las ventajas y casos donde es aconsejable recopilar información interna y externa en un Almacén de Datos.
- Conocer el modelo multidimensional de los almacenes de datos y los operadores de refinamiento asociados: *drill, roll, slice & dice, pivot*.
- Conocer la arquitectura y diferentes implementaciones (ROLAP, MOLAP) de Almacenes de Datos.
- Reconocer pautas para el diseño y mantenimiento de ADs.

Introducción a los Almacenes de Datos

OBJETIVO:
Análisis de Datos para el Soporte en la Toma de Decisiones.

- Generalmente, **la información** que se quiere investigar sobre un cierto dominio de la organización **se encuentra en bases de datos y otras fuentes muy diversas**, tanto internas como externas.
- Muchas de estas fuentes son las que se utilizan para el trabajo diario (**bases de datos operacionales**).

Introducción a los Almacenes de Datos

- Sobre estas mismas bases de datos de trabajo ya se puede extraer conocimiento (visión tradicional).
- Uso de la base de datos transaccional para varios cometidos:
  - Se mantiene el trabajo transaccional diario de los sistemas de información originales (conocido como **OLTP, *On-Line Transactional Processing***).
  - Se hace análisis de los datos en tiempo real sobre la misma base de datos (conocido como **OLAP, *On-Line Analytical Processing***).

Introducción a los Almacenes de Datos

- Uso de la base de datos transaccional para varios cometidos:
- PROBLEMAS:
  - perturba el trabajo transaccional diario de los sistemas de información originales ("*killer queries*"). Se debe hacer por la noche o en fines de semana.
  - la base de datos está diseñada para el trabajo transaccional, no para el análisis de los datos. Generalmente no puede ser en tiempo real (era AP pero no OLAP).

Introducción a los Almacenes de Datos

- Se desea operar eficientemente con esos datos...
  - los costes de almacenamiento masivo y conectividad se han reducido drásticamente en los últimos años,
- parece razonable recoger los datos (información histórica) en **un sistema separado y específico**.

NACE EL DATA-WAREHOUSING

- *Data warehouses* (Almacenes o Bodegas de Datos)

Introducción a los Almacenes de Datos

Almacenes de Datos (AD) (data warehouse)

motivación

disponer de Sistemas de Información de apoyo a la toma de decisiones*

disponer de *bases de datos* que permitan *extraer conocimiento* de la información histórica almacenada en la organización

objetivos

análisis de la organización | previsiones de evolución | diseño de estrategias

* DSS: Decision Support Systems

Introducción a los Almacenes de Datos

Almacenes de datos

Base de Datos diseñada con un objetivo de explotación distinto que el de las bases de datos de los sistemas operacionales.

Sistema Operacional (OLTP)	BD orientada al proceso
Sistema de Almacén de Datos (DW)	BD orientada al análisis

Introducción a los Almacenes de Datos

Almacenes de Datos

definición

colección de datos diseñada para dar apoyo a los procesos de toma de decisiones

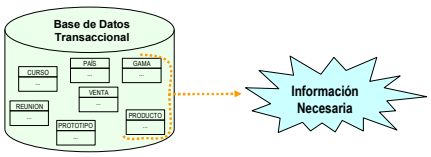
características

- orientada hacia la información* relevante de la organización
- integrada
- variable en el tiempo
- no volátil

** subject oriented, not process oriented*

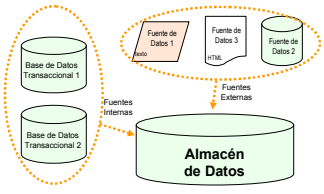
Introducción a los Almacenes de Datos

AD: Orientado hacia la información relevante de la organización → se diseña para consultar eficientemente información relativa a las actividades (ventas, compras, producción, ...) básicas de la organización, no para soportar los procesos que se realizan en ella (gestión de pedidos, facturación, etc).



Introducción a los Almacenes de Datos

AD: Integrado → integra datos recogidos de diferentes sistemas operacionales de la organización (y/o fuentes externas).



Introducción a los Almacenes de Datos

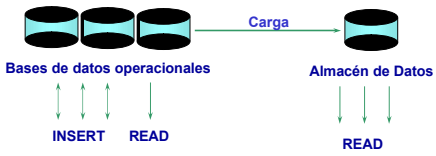
AD: Variable en el tiempo ➡ los datos son relativos a un periodo de tiempo y deben ser incrementados periódicamente.

Los datos son almacenados como fotos (snapshots) correspondientes a periodos de tiempo.

13

Introducción a los Almacenes de Datos

AD: No volátil ➡ los datos almacenados no son actualizados, sólo son incrementados.



El periodo de tiempo cubierto por un AD varía entre 2 y 10 años.

14

Introducción a los Almacenes de Datos

Almacenes de Datos

ventajas para las organizaciones

- rentabilidad de las inversiones realizadas para su creación
- aumento de la competitividad en el mercado
- aumento de la productividad de los técnicos de dirección

15

Introducción a los Almacenes de Datos

Almacenes de Datos

problemas

- infravaloración del esfuerzo necesario para su diseño y creación
- infravaloración de los recursos necesarios para la captura, carga y almacenamiento de los datos
- incremento continuo de los requisitos de los usuarios
- privacidad de los datos

16

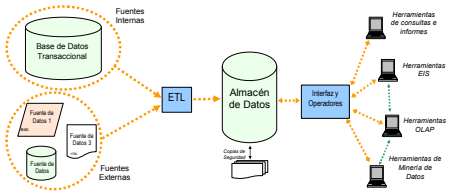
Introducción a los Almacenes de Datos

Sistema Operacional (OLTP)	Almacén de datos (DW)
- almacena datos actuales	- almacena datos históricos
- almacena datos de detalle	- almacena datos de detalle y datos agregados a distintos niveles
-bases de datos medianas (100Mb-1Gb)	- bases de datos grandes (100Gb-1Tb)
- los datos son dinámicos (actualizables)	- los datos son estáticos
- los procesos (transacciones) son repetitivos	- los procesos no son previsibles
- el número de transacciones es elevado	- el número de transacciones es bajo o medio
- tiempo de respuesta pequeño (segundos)	- tiempo de respuesta variable (segundos-horas)
- dedicado al procesamiento de transacciones	- dedicado al análisis de datos
- orientado a los procesos de la organización	- orientado a la información relevante
- soporta decisiones diarias	- soporta decisiones estratégicas
- sirve a muchos usuarios (administrativos)	- sirve a técnicos de dirección

17

Arquitectura de un Almacén de Datos

- La Arquitectura de un AD viene determinada por su situación central como fuente de información para las herramientas de análisis.



18

Arquitectura de un Almacén de Datos

- Componentes:
  - Sistema ETL (*Extraction, Transformation, Load*): realiza las funciones de *extracción* de las fuentes de datos (transaccionales o externas), *transformación* (limpieza, consolidación, ...) y la *carga* del AD, realizando:
    - extracción de los datos.
    - filtrado de los datos: limpieza, consolidación, etc.
    - carga inicial del almacén: ordenación, agregaciones, etc.
    - refresco del almacén: operación periódica que propaga los cambios de las fuentes externas al almacén de datos
  - Repositorio Propio de Datos: información relevante, metadatos.
  - Interfaces y Gestores de Consulta: permiten acceder a los datos ys sobre ellos se conectan herramientas más sofisticadas (OLAP, EIS, minería de datos).
  - Sistemas de Integridad y Seguridad: se encargan de un mantenimiento global, copias de seguridad, ...

19

Arquitectura de un Almacén de Datos

- Organización (Externa) de Los Datos…

Las herramientas de explotación de los almacenes de datos han adoptado un **modelo multidimensional de datos**.

Se ofrece al usuario una visión multidimensional de los datos que son objeto de análisis.

20

Arquitectura de un Almacén de Datos

EJEMPLO

Organización: Cadena de supermercados.

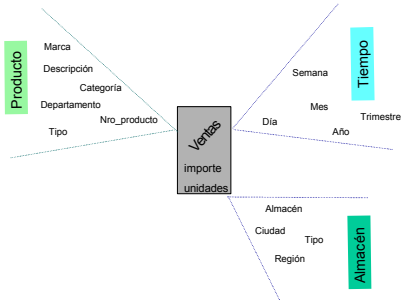
Actividad objeto de análisis: ventas de productos.

Información registrada sobre una venta: "del **producto** "Tauritón 33cl" se han vendido en el **almacén** "Almacén nro.1" el **día** 17/7/2003, 5 **unidades** por un **importe** de 103,19 euros."

Para hacer el análisis no interesa la venta individual (ticket) realizada a un cliente sino las ventas diarias de productos en los distintos almacenes de la cadena.

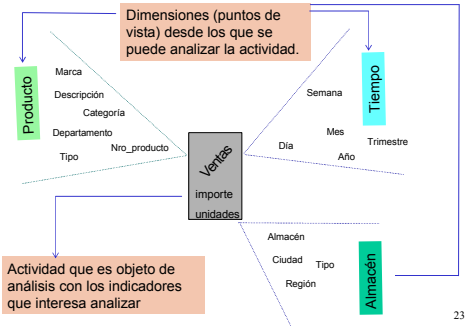
21

Arquitectura de un Almacén de Datos



22

Arquitectura de un Almacén de Datos



23

Arquitectura de un Almacén de Datos

Modelo multidimensional:

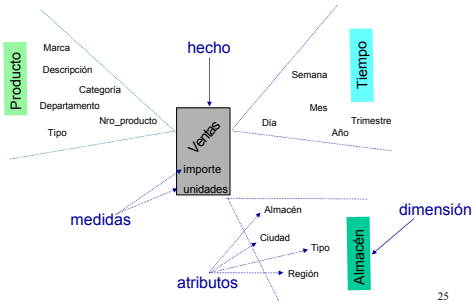
✓ en un esquema multidimensional se representa una actividad que es objeto de análisis (hecho) y las dimensiones que caracterizan la actividad (dimensiones).

✓la información relevante sobre el hecho (actividad) se representa por un conjunto de indicadores (medidas o atributos de hecho).

✓la información descriptiva de cada dimensión se representa por un conjunto de atributos (atributos de dimensión).

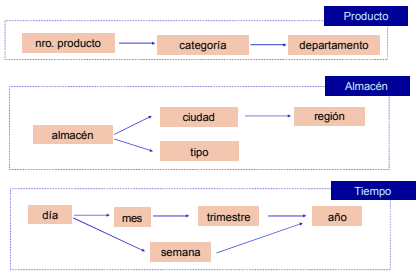
24

Arquitectura de un Almacén de Datos



Arquitectura de un Almacén de Datos

Entre los atributos de una dimensión se definen jerarquías



Arquitectura de un Almacén de Datos

Este esquema multidimensional recibe varios nombres:

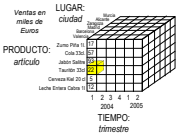
- estrella: si la jerarquía de dimensiones es lineal

- estrella jerárquica o copo de nieve: si la jerarquía no es lineal.

Arquitectura de un Almacén de Datos

- Se pueden obtener hechos a diferentes niveles de agregación:
  - obtención de medidas sobre los hechos parametrizadas por atributos de las dimensiones y restringidas por condiciones impuestas sobre las dimensiones

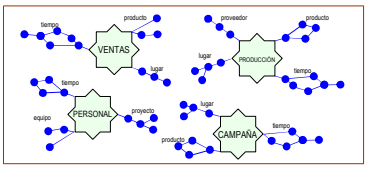
***HECHO**: "El primer trimestre de 2004 la empresa vendió en Valencia por un importe de 22.000 euros del producto tauritón 33 cl."*



- Un nivel de agregación para un conjunto de dimensiones se denomina cubo.

Arquitectura de un Almacén de Datos

- ¿Se puede recopilar toda la información necesaria en un único esquema estrella o copo de nieve?
  - NO : necesidad de varios esquemas.
- Cada uno de estos esquemas se denomina datamart.



Almacén formado por 4 datamarts.

Arquitectura de un Almacén de Datos

- El almacén de datos puede estar formado por varios datamarts y, opcionalmente, por tablas adicionales.

Data mart ⇒ subconjunto de un almacén de datos, generalmente en forma de estrella o copo de nieve.

- ✓ se definen para satisfacer las necesidades de un departamento o sección de la organización.
- ✓ contiene menos información de detalle y más información agregada.

Herramientas OLAP

- ✓ Las herramientas de OLAP presentan al usuario una visión multidimensional de los datos (esquema multidimensional) para cada actividad que es objeto de análisis.
- ✓ El usuario formula consultas a la herramienta OLAP seleccionando atributos de este esquema multidimensional sin conocer la estructura interna (esquema físico) del almacén de datos.
- ✓ La herramienta OLAP genera la correspondiente consulta y la envía al gestor de consultas del sistema (p.ej. mediante una sentencia SELECT).

31

Herramientas OLAP

una consulta a un almacén de datos consiste generalmente en la obtención de medidas sobre los hechos parametrizadas por atributos de las dimensiones y restringidas por condiciones impuestas sobre las dimensiones

medida — hecho

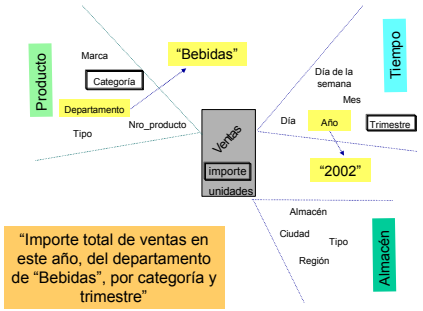
¿ "Importe total de las ventas durante este año de los productos del departamento *Bebidas*, por **trimestre** y por **categoría**" ?.

Restricciones: productos del departamento Bebidas, ventas durante este año

Parámetros de la consulta: por categoría de producto y por trimestre

32

Herramientas OLAP

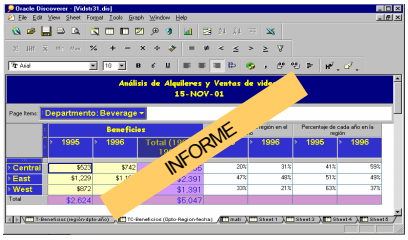


"Importe total de ventas en este año, del departamento de "Bebidas", por categoría y trimestre"

33

Herramientas OLAP

trimestre — categoría — importe



34

Herramientas OLAP

Categoría	Trimestre	Ventas
Refrescos	T1	2000000
Refrescos	T2	1000000
Refrescos	T3	3000000
Refrescos	T4	2000000
Zumos	T1	1000000
Zumos	T2	1500000
Zumos	T3	8000000
Zumos	T4	2400000

Presentación tabular (relacional) de los datos seleccionados

Se asumen dos categorías en el departamento de *Bebidas*: Refrescos y Zumos.

35

Herramientas OLAP

trimestre / categoría	T1	T2	T3	T4
Refrescos	2000000	1000000	3000000	2000000
Zumos	1000000	1500000	8000000	2400000

Presentación matricial (multidimensional) de los datos seleccionados

Los parámetros de la consulta ("por trimestre" y "por categoría") determinan los criterios de agrupación de los datos seleccionados (ventas de productos del departamento *Bebidas* durante este año). La agrupación se realiza sobre dos dimensiones (Producto, Tiempo).

36

Herramientas OLAP

- Lo interesante no es poder realizar consultas que, en cierto modo, se pueden hacer con selecciones, proyecciones, concatenaciones y agrupamientos tradicionales.
- Lo realmente interesante de las herramientas OLAP son sus **operadores de refinamiento o manipulación de consultas**.
  - DRILL
  - ROLL
  - SLICE & DICE
  - PIVOT

37

Herramientas OLAP

El carácter agregado de las consultas en el Análisis de Datos, aconseja la definición de nuevos operadores que faciliten la agregación (consolidación) y la disgregación (división) de los datos:

- ✓ agregación (roll): permite eliminar un criterio de agrupación en el análisis, agregando los grupos actuales.
- ✓ disgregación (drill): permite introducir un nuevo criterio de agrupación en el análisis, disgregando los grupos actuales.

38

Herramientas OLAP

Si se desea introducir la dimensión Almacén en el análisis anterior e incluir un nuevo criterio de agrupación sobre la ciudad del almacén:

¿ "Importe total de las ventas durante este año de los productos del departamento *Bebidas*, por **trimestre**, por **categorías** y por **ciudad del almacén**" ?.

Restricciones: productos del departamento Bebidas, ventas durante este año

Parámetros de la consulta: por categoría de producto, por trimestre y por ciudad del almacén.

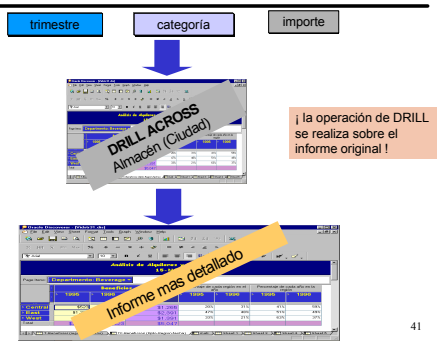
39

Herramientas OLAP

"Importe total de ventas en este año, del departamento de "Bebidas", por categoría, trimestre y ciudad"

40

Herramientas OLAP



¡ la operación de DRILL se realiza sobre el informe original !

41

Herramientas OLAP

Categoría	Trimestre	Ventas
Refrescos	T1	2000000
Refrescos	T2	1000000
Refrescos	T3	3000000
Refrescos	T4	2000000
Zumos	T1	1000000
Zumos	T2	1500000
Zumos	T3	8000000
Zumos	T4	2400000

drill-across

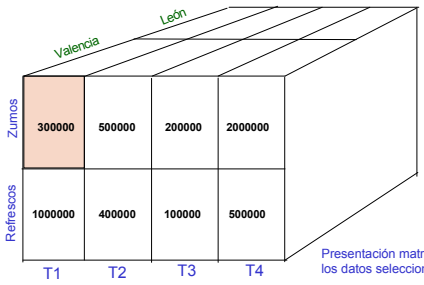
Categoría	Trimestre	Ciudad	Ventas
Refrescos	T1	Valencia	1000000
Refrescos	T1	León	1000000
Refrescos	T2	Valencia	400000
Refrescos	T2	León	700000

Cada grupo (categoría-trimestre) de la consulta original se disgrega en dos nuevos grupos (categoría-trimestre-ciudad) para las ciudades de León y Valencia.

* Se asumen dos ciudades: Valencia y León

42

Herramientas OLAP



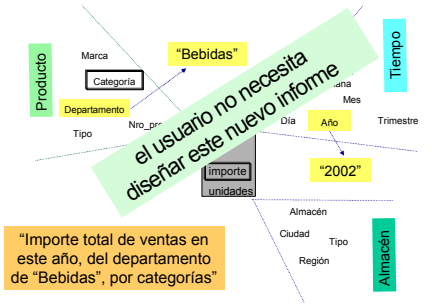
Presentación matricial de los datos seleccionados.

Herramientas OLAP

Si se desea eliminar el criterio de agrupación sobre la dimensión Tiempo en la consulta original:

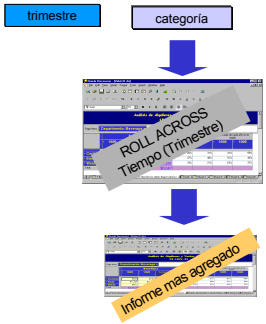
¿ "Importe total de las ventas durante este año de los productos del departamento *Bebidas*, por **categorías**" ?

Herramientas OLAP



"Importe total de ventas en este año, del departamento de "Bebidas", por categorías"

Herramientas OLAP



¡ la operación de ROLL se realiza sobre el informe original !

Herramientas OLAP

Categoría	Trimestre	Ventas
Refrescos	T1	2000000
Refrescos	T2	1000000
Refrescos	T3	3000000
Refrescos	T4	2000000
Zumos	T1	1000000
Zumos	T2	1500000
Zumos	T3	8000000
Zumos	T4	2400000

roll-across

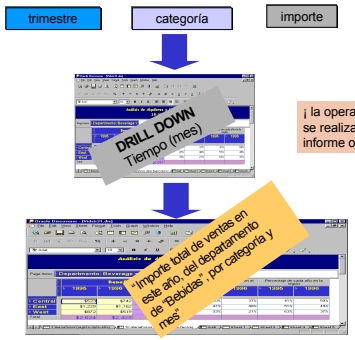
Categoría	Ventas
Refrescos	8000000
Zumos	12900000

Herramientas OLAP

Las operaciones de agregación (DRILL) y disgregación (ROLL) se pueden hacer sobre:

- ✓atributos de una dimensión sobre los que se ha definido una jerarquía: DRILL-DOWN, ROLL-UP
  - departamento – categoría - producto (Producto)
  - año - trimestre – mes - día (Tiempo)
- ✓sobre dimensiones independientes: DRILL-ACROSS, ROLL-ACROSS
  - Producto – Almacén -Tiempo

Herramientas OLAP



¡ la operación de DRILL se realiza sobre el informe original !

Herramientas OLAP

Categoría	Trimestre	Ventas
Refrescos	T1	2000000
Refrescos	T2	1000000
Refrescos	T3	3000000
Refrescos	T4	2000000
Zumos	T1	1000000
Zumos	T2	1500000
Zumos	T3	8000000
Zumos	T4	2400000

drill-down

Categoría	Trimestre	Mes	Ventas
Refrescos	T1	Enero	1000000
Refrescos	T1	Febrero	500000
Refrescos	T1	Marzo	500000

Cada grupo (categoría-trimestre) de la consulta original se disgrega en dos nuevos grupos (categoría-trimestre-mes).

Herramientas OLAP

Otras operaciones de OLAP:

- ✓ SLICE & DICE: seleccionar y proyectar datos en el informe.
- ✓ PIVOT: reorientación de las dimensiones en el informe.

Herramientas OLAP

Ventas

	Productos	Store1	Store2
Q1	Electronics	$5,2	$5,6
	Toys	$1,9	$1,4
	Clothing	$2,3	$2,6
	Cosmetics	$1,1	$1,1
Q2	Electronics	$8,9	$7,2
	Toys	$0,75	$0,4
	Clothing	$4,6	$4,6
	Cosmetics	$1,5	$0,5

PIVOT

Ventas

	Productos	Q1	Q2
Store 1	Electronics	$5,2	$8,9
	Toys	$1,9	$0,75
	Clothing	$2,3	$4,6
	Cosmetics	$1,1	$1,5
Store 2	Electronics	$5,6	$7,2
	Toys	$1,4	$0,4
	Clothing	$2,6	$4,6
	Cosmetics	$1,1	$0,5

Herramientas OLAP

Ventas

	Productos	Store1	Store2
Q1	Electronics	$5,2	$5,6
	Toys	$1,9	$1,4
	Clothing	$2,3	$2,6
	Cosmetics	$1,1	$1,1
Q2	Electronics	$8,9	$7,2
	Toys	$0,75	$0,4
	Clothing	$4,6	$4,6
	Cosmetics	$1,5	$0,5

SLICE & DICE

Ventas

	Productos	Store1
Q1	Electronics	$5,2
	Toys	$1,9
Q2	Electronics	$8,9
	Toys	$0,75

Herramientas OLAP

Las herramientas de OLAP se caracterizan* por:

- ✓ ofrecer una visión multidimensional de los datos (matricial).
- ✓ no imponer restricciones sobre el número de dimensiones.
- ✓ ofrecer simetría para las dimensiones.
- ✓ permitir definir de forma flexible (sin limitaciones) sobre las dimensiones: restricciones, agregaciones y jerarquías entre ellas.
- ✓ ofrecer operadores intuitivos de manipulación: *drill-down, roll-up, slice-and-dice, pivot.*
- ✓ ser transparentes al tipo de tecnología que soporta el almacén de datos (ROLAP o MOLAP).

*Subconjunto de las 12 reglas propuestas por E.F. Codd para A.D.

ROLAP y MOLAP

- El Almacén de Datos y las herramientas OLAP se pueden basar *físicamente* en varias organizaciones:

  Sistemas ROLAP
  - ✓ se implementan sobre tecnología relacional, pero disponen de algunas facilidades para mejorar el rendimiento (índices de mapas de bits, índices de JOIN).

  Sistemas MOLAP
  - ✓ disponen de estructuras de almacenamiento específicas (arrays) y técnicas de compactación de datos que favorecen el rendimiento del almacén.

  Sistemas HOLAP
  - ✓ sistemas híbridos entre ambos.

55

ROLAP y MOLAP

Sistemas ROLAP:

- ✓ El almacén de datos se construye sobre un SGBD Relacional.
- ✓ Los fabricantes de SGBD relacionales ofrecen extensiones y herramientas para poder utilizar el SGBDR como un Sistema Gestor de Almacenes de Datos.

56

ROLAP y MOLAP

Sistemas ROLAP:

Extensiones de los SGBD relacionales:

- ✓ índices de mapa de bits
- ✓ índices de JOIN
- ✓ técnicas de particionamiento de los datos
- ✓ optimizadores de consultas
- ✓ extensiones del SQL (operador CUBE, roll-up)

57

ROLAP y MOLAP

Sistemas MOLAP.

Sistema de propósito específico:

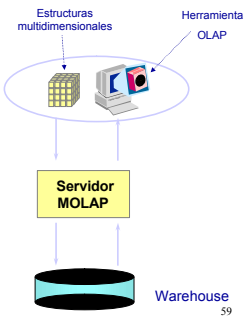
- ✓ estructuras de datos (arrays)
- ✓ técnicas de compactación.

El objetivo de los sistemas MOLAP es almacenar físicamente los datos en estructuras multidimensionales de forma que la representación externa y la representación interna coincidan.

58

ROLAP y MOLAP

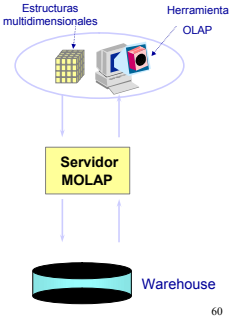
- El servidor MOLAP construye y almacena datos en estructuras multidimensionales.
- La herramienta de OLAP presenta estas estructuras multidimensionales.



59

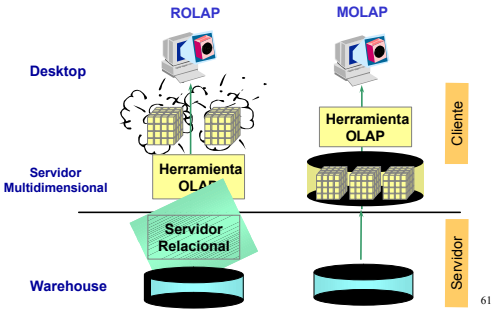
ROLAP y MOLAP

MOLAP:
- Datos
  - Arrays
  - Extraídos del almacén de datos
- almacenamiento y procesos eficientes
- la complejidad de la BD se oculta a los usuarios
- el análisis se hace sobre datos agregados y métricas o indicadores precalculados.



60

ROLAP y MOLAP



61

ROLAP y MOLAP

ROLAP/MOLAP: Ventajas e Inconvenientes:

ROLAP

- ✓ pueden aprovechar la tecnología relacional.
- ✓ pueden utilizarse sistemas relacionales genéricos (más baratos o incluso gratuitos).
- ✓ el diseño lógico corresponde al físico si se utiliza el diseño de Kimball.

MOLAP:

- ✓ generalmente más eficientes que los ROLAP.
- ✓ el coste de los cambios en la visión de los datos.
- ✓ la construcción de las estructuras multidimensionales.

62

Carga y Mantenimiento de un A.D.

El sistema encargado del mantenimiento del almacén de datos es el Sistema E.T.T* (Extracción - Transformación -Transporte)

- La construcción del Sistema E.T.T es responsabilidad del equipo de desarrollo del almacén de datos.
- El Sistema E.T.T es construido específicamente para cada almacén de datos. Aproximadamente 50% del esfuerzo.
- En la construcción del E.T.T se pueden utilizar herramientas del mercado o programas diseñados específicamente.

Funciones del Sistema E.T.T:

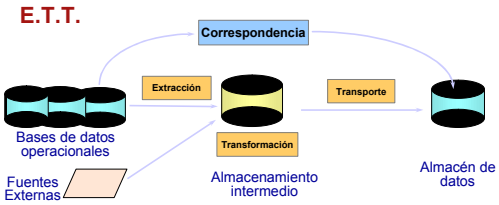
- Carga inicial. (initial load)
- Mantenimiento o *refresco* periódico: inmediato, diario, semanal, mensual,... (refreshment)

* Conocido también por "E.T.L: Extracción – Transformación – Load (carga)"

63

Carga y Mantenimiento de un A.D.

E.T.T.



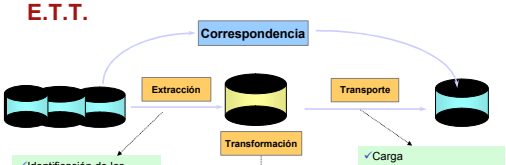
El Almacenamiento intermedio permite:

- Realizar transformaciones sin paralizar las bases de datos operacionales y el almacén de datos.
- Almacenar metadatos.
- Facilitar la integración de fuentes externas.

64

Carga y Mantenimiento de un A.D.

E.T.T.



- ✓Identificación de los datos que han cambiado
- ✓Extracción (lectura) de datos.
- ✓Obtención de agregados
- ✓Mantenimiento de metadata

- ✓Limpieza y transformación de datos
- ✓Integración de datos (cálculo de datos derivados)
- ✓Creación de claves
- ✓Obtención de agregados
- ✓Mantenimiento de metadata

- ✓Carga
- ✓ Indización
- ✓Obtención de datos agregados.
- ✓ Realización de pruebas de calidad de la carga.
- ✓Gestión de errores.
- ✓Mantenimiento de metadata

65

Carga y Mantenimiento de un A.D.

La "calidad de los datos" es la clave del éxito de un almacén de datos.

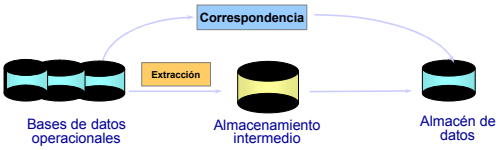
Definir una estrategia de calidad:

- actuación sobre los sistemas operacionales: modificar las reglas de integridad, los disparadores y las aplicaciones de los sistemas operacionales.
- documentación de las fuentes de datos.
- definición de un proceso de transformación.
- nombramiento de un responsable de calidad del sistema (*Data Quality Manager*).

66

Carga y Mantenimiento de un A.D.

Extracción.



– Programas diseñados para extraer los datos de las fuentes.
– Herramientas: *data migration tools, wrappers*, ...

Carga y Mantenimiento de un A.D.

Extracción: lectura de datos del sistema operacional.

a) durante la carga inicial .

b) mantenimiento del AD

Ejecución de la extracción:

a) si los datos operacionales están mantenidos en **un SGBDR**, **la extracción** de datos se puede reducir a consultas **en SQL** o rutinas programadas.

b) si los datos operacionales están en un **sistema propietario** (no se conoce el formato de los datos) **o** en **fuentes externas** textuales, hipertextuales u hojas de cálculo, **la extracción puede ser muy difícil** y puede tener que realizarse a partir de informes o volcados de datos proporcionados por los propietarios que deberán ser procesados posteriormente.

Carga y Mantenimiento de un A.D.

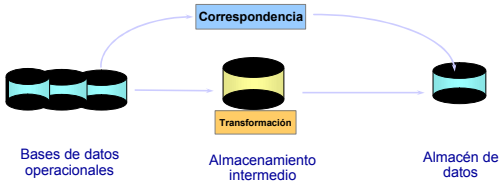
Extracción: en el mantenimiento/refresco del AD. Antes de realizar la extracción es preciso Identificar los Cambios.

Identificación de Cambios.

– Identificar los datos operacionales (relevantes) que han sufrido una modificación desde la fecha del último mantenimiento.
– Métodos
  - Carga total: cada vez se empieza de cero.
  - Comparación de instancias de la base de datos operacional.
  - Uso de marcas de tiempo (*time stamping*) en los registros del sistema operacional.
  - Uso de disparadores en el sistema operacional.
  - Uso del fichero de *log* (gestión de transacciones) del sistema operacional.
  - Uso de técnicas mixtas.

Carga y Mantenimiento de un A.D.

Transformación.



- Transformar los datos extraídos de las fuentes operacionales: limpieza, estandarización. (cleansing)
- Calcular los datos derivados: aplicar las leyes de derivación. (integration)

Carga y Mantenimiento de un A.D.

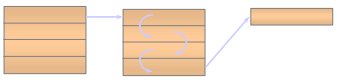
Transformación.

– En los datos operacionales existen anomalías: desarrollos independientes a lo largo del tiempo, fuentes heterogéneas, ..
– Eliminar anomalías:
  - Limpieza de datos: eliminar datos, corregir y completar datos, eliminar duplicados, ...
  - Estandarización: codificación, formatos, unidades de medida, ...

Carga y Mantenimiento de un A.D.

Transformación.

– Claves con estructura: descomponer en valores atómicos



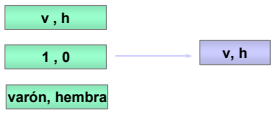
Código de producto = 12M65431345

código del país | zona de ventas | número de producto | código de vendedor

Carga y Mantenimiento de un A.D.

Transformación.

- Unificar codificaciones: existencia de codificaciones múltiples.



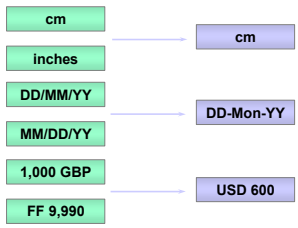
- Deben detectarse los valores erróneos.

73

Carga y Mantenimiento de un A.D.

Transformación.

- Unificar estándares:unidades de medida, unidades de tiempo,moneda,...



74

Carga y Mantenimiento de un A.D.

Transformación.

- Valores duplicados: deben ser eliminados.
  - SQL
  - restricciones en el SGBDR

75

Carga y Mantenimiento de un A.D.

Transformación.

- Integridad referencial: debe reconstruirse.

Departamento
10
20
30
40

Emp	Nombre	Departamento
1099	Smith	10
1289	Jones	20
1234	Doe	50
6786	Harris	60

76

Carga y Mantenimiento de un A.D.

Transformación. Creación de claves.

#1	Venta	1/2/98	12:00:01 Ham Pizza	$10.00
#2	Venta	1/2/98	12:00:02 Cheese Pizza	$15.00
#3	Venta	1/2/98	12:00:02 Anchovy Pizza	$12.00
#4	Devolución	1/2/98	12:00:03 Anchovy Pizza	- $12.00
#5	Venta	1/2/98	12:00:04 Sausage Pizza	$11.00

Claves sin significado

#dw1	Venta	1/2/98	12:00:01 Ham Pizza	$10.00
#dw2	Venta	1/2/98	12:00:02 Cheese Pizza	$15.00
#dw3	Venta	1/2/98	12:00:04 Sausage Pizza	$11.00

77

Carga y Mantenimiento de un A.D.

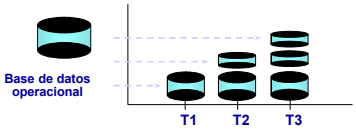
Transporte. (carga)

- La fase de Transporte consiste en mover los datos desde las fuentes operacionales o el almacenamiento intermedio hasta el almacén de datos y cargar los datos en las correspondientes estructuras de datos.
- La carga puede consumir mucho tiempo.
- En la carga inicial del AD se mueven grandes volúmenes de datos.
- En los mantenimientos periódicos del AD se mueven pequeños volúmenes de datos.
- La frecuencia del mantenimiento periódico está determinada por el gránulo del AD y los requisitos de los usuarios.

78

Carga y Mantenimiento de un A.D.

Transporte. Creación y mantenimiento de un AD.



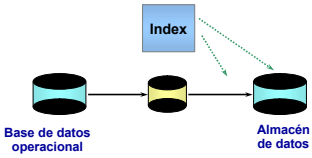
- Crear el AD (base de datos)
- En intervalos de tiempo fijos añadir cambios al AD. Se deben determinar las "ventanas de carga" más convenientes para no saturar la base de datos operacional.
- Ocasionalmente archivar o eliminar datos obsoletos que ya no interesan para el análisis.

79

Carga y Mantenimiento de un A.D.

Procesos posteriores a la carga: indización.

- Durante la carga:
  - carga con el índice habilitado
  - proceso tupla a tupla. (lento)
- Después de la carga:
  - carga con el índice deshabilitado
  - creación del índice (total o parcial). (rápido)

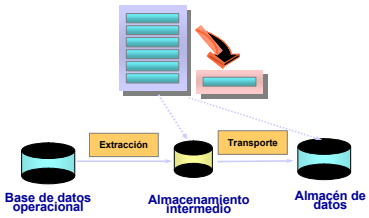


80

Carga y Mantenimiento de un A.D.

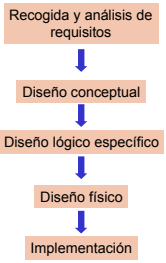
Procesos posteriores a la carga: obtención de agregados.

- Durante la extracción.
- Después de la carga (transporte).



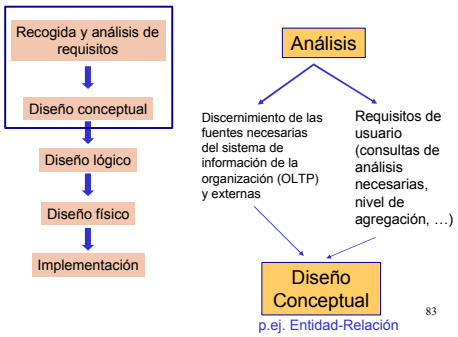
81

Diseño de un Almacén de Datos



82

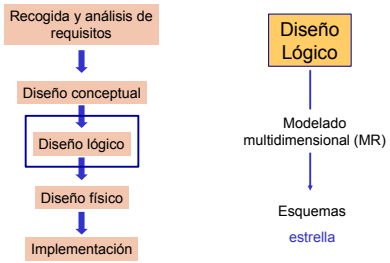
Diseño de un Almacén de Datos



p.ej. Entidad-Relación

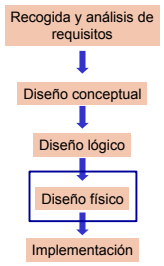
83

Diseño de un Almacén de Datos



84

Diseño de un Almacén de Datos



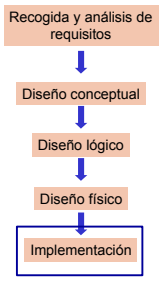
Diseño Físico

Definición del esquema ROLAP o MOLAP

Diseño del ETL

85

Diseño de un Almacén de Datos



Implementación

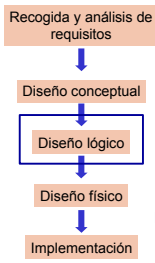
Carga del AD (ETL)

Preparación de las vistas de usuario (herramienta OLAP)

86

Diseño de un Almacén de Datos

- *Detallemos más ahora el Diseño Lógico...*



La visión multidimensional seguida por las herramientas de explotación de almacenes de datos (OLAP) ha inspirado los modelos y metodologías de diseño de este tipo de sistemas.

En la literatura se habla de "Bases de Datos Multidimensionales" y de "Diseño Multidimensional"

87

Diseño de un Almacén de Datos

Modelado multidimensional:

- ✓ en un esquema multidimensional se representa una actividad que es objeto de análisis (hecho) y las dimensiones que caracterizan la actividad (dimensiones).
- ✓ la información relevante sobre el hecho (actividad) se representa por un conjunto de indicadores (medidas o atributos de hecho).
- ✓ la información descriptiva de cada dimensión se representa por un conjunto de atributos (atributos de dimensión).

88

Diseño de un Almacén de Datos

Modelado multidimensional:

- ✓ el modelado multidimensional se puede aplicar utilizando distintos modelos de datos (conceptuales o lógicos).
- ✓ la representación gráfica del esquema multidimensional dependerá del modelo de datos utilizado (relacional, ER, UML, OO, ...).

89

Diseño de un Almacén de Datos

El desarrollo de la tecnología de almacenes de datos se ha caracterizado por:

- un temprano desarrollo industrial provocado por las demandas de los usuarios.
- el uso de metodologías de diseño centradas principalmente en los niveles lógico e interno. (la atención se ha centrado en mejorar la eficiencia en la ejecución de consultas)

Metodología de diseño basada en el modelo relacional: Modelo multidimensional de Kimball

No es muy congruente usar modelos para diseño conceptual….

90

Diseño de un Almacén de Datos

Pasos en el diseño del almacén de datos:

- Paso 1. Elegir un "proceso" de la organización para modelar.
- Paso 2. Decidir el gránulo (nivel de detalle) de representación del proceso.
- Paso 3. Identificar las dimensiones que caracterizan el proceso.
- Paso 4. Decidir la información a almacenar sobre el proceso.

91

Diseño de un Almacén de Datos

Paso 1. Elegir un *"proceso"* de la organización para modelar.

Proceso: actividad de la organización soportada por un OLTP del cual se puede extraer información con el propósito de construir el almacén de datos.

Pedidos (de clientes)

Compras (a suministradores)

Facturación

Envíos

Ventas

Inventario

…

92

Diseño de un Almacén de Datos

Ejemplo: Cadena de supermercados.

Cadena de supermercados con 300 almacenes en la que se expenden unos 30.000 productos distintos.

Actividad: *Ventas.*

La actividad a modelar son las ventas de productos en los almacenes de la cadena.

93

Diseño de un Almacén de Datos

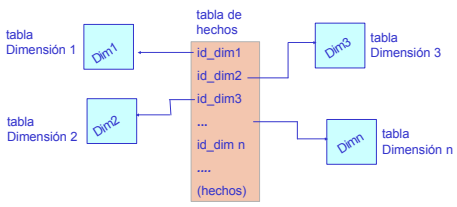
Paso 2. Decidir el gránulo (nivel de detalle) de representación.

Gránulo: es el nivel de detalle al que se desea almacenar información sobre la actividad a modelar.

- ✓ El gránulo define el nivel atómico de datos en el almacén de datos.
- ✓ El gránulo determina el significado de las tuplas de la tabla de hechos.
- ✓ El gránulo determina las dimensiones básicas del esquema
  - *transacción en el OLTP*
  - *información diaria*
  - *información semanal*
  - *información mensual. ….*

94

Diseño de un Almacén de Datos



95

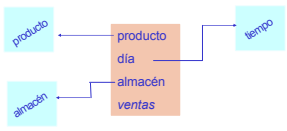
Diseño de un Almacén de Datos

Ejemplo: Cadena de supermercados.

Gránulo: "se desea almacenar información sobre las ventas diarias de cada producto en cada almacén de la cadena".

Gránulo:

- ✓define el significado de las tuplas de la tabla de hechos.
- ✓determina las dimensiones básicas del esquema.



96

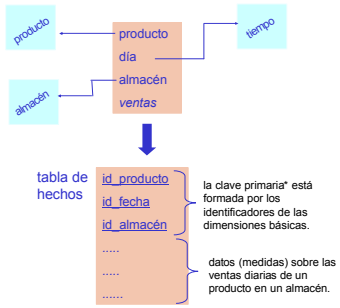
Diseño de un Almacén de Datos

- Gránulo inferior: no se almacena información a nivel de línea de ticket porque no se puede identificar siempre al cliente de la venta lo que permitiría hacer análisis del comportamiento (hábitos de compra) del cliente.
- Gránulo superior: no se almacena información a nivel semanal o mensual porque se perderían opciones de análisis interesantes: ventas en días previos a vacaciones, ventas en fin de semana, ventas en fin de mes, ...

En un almacén de datos se almacena información a un nivel de detalle (gránulo) fino no porque se vaya a interrogar el almacén a ese nivel sino porque ello permite clasificar y estudiar (analizar) la información desde muchos puntos de vista.

97

Diseño de un Almacén de Datos



* pueden existir excepciones a esta regla general

98

Diseño de un Almacén de Datos

Paso 3. Identificar las dimensiones que caracterizan el proceso.

- ✓ *Dimensiones*: dimensiones que caracterizan la actividad al nivel de detalle (gránulo) que se ha elegido.
  - *Tiempo* *(dimensión temporal: ¿cuándo se produce la actividad?)*
  - *Producto* *(dimensión ¿cuál es el objeto de la actividad?)*
  - *Almacén* *(dimensión geográfica: ¿dónde se produce la actividad?)*
  - *Cliente* *(dimensión ¿quién es el destinatario de la actividad?)*
- ✓ De cada *dimensión* se debe decidir los atributos (propiedades) relevantes para el análisis de la actividad.
- ✓ Entre los atributos de una dimensión existen jerarquías naturales que deben ser identificadas (día-mes-año)

99

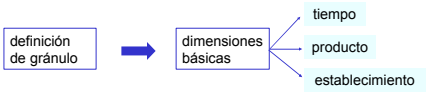
Diseño de un Almacén de Datos

tabla
Dimensión 1

100

Diseño de un Almacén de Datos

Ejemplo: Cadena de supermercados.



Nota: En las aplicaciones reales el número de dimensiones suele variar entre 3 y 15 dimensiones.

101

Diseño de un Almacén de Datos

Dimensión Tiempo:

- ✓ dimensión presente en todo AD porque el AD contiene información histórica sobre la organización.
- ✓ aunque el lenguaje SQL ofrece funciones de tipo DATE, una dimensión Tiempo permite representar otros atributos temporales no calculables en SQL.
- ✓ se puede calcular de antemano
- ✓ atributos frecuentes:
  - nro. de día, nro. de semana, nro. de año: valores absolutos del calendario juliano que permiten hacer ciertos cálculos aritméticos.
  - día de la semana (lunes, martes, miércoles,...): permite hacer análisis sobre días de la semana concretos (ej. ventas en sábado, ventas en lunes,..).

102

Diseño de un Almacén de Datos

Dimensión Tiempo:

- ✓ atributos frecuentes:
  - día del mes (1..31): permite hacer comparaciones sobre el mismo día en meses distintos (ventas el 1º de mes).
  - marca de fin de mes, marca de fin de semana : permite hacer comparaciones sobre el último día del mes o días de fin de semana en distintos meses.
  - trimestre del año (1..4): permite hacer análisis sobre un trimestre concreto en distintos años.
  - marca de día festivo: permite hacer análisis sobre los días contiguos a un día festivo.
  - estación (primavera, verano..)
  - evento especial: permite marcar días de eventos especiales (final de futbol, elecciones...)
- ✓ jerarquía natural:

  día - mes - trimestre -año

Diseño de un Almacén de Datos

Dimensión Producto:

- ✓ la dimensión Producto se define a partir del fichero maestro de productos del sistema OLTP.
- ✓ las actualizaciones del fichero maestro de productos deben reflejarse en la dimensión Producto (¿cómo?).
- ✓ la dimensión Producto debe contener el mayor número posible de atributos descriptivos que permitan un análisis flexible. Un número frecuente es de 50 atributos.
- ✓ atributos frecuentes: identificador (código estándar), descripción, tamaño del envase, marca, categoría, departamento, tipo de envase, producto dietético, peso, unidades de peso, unidades por envase, fórmula, ...
- ✓ jerarquías: producto-categoría-departamento

Diseño de un Almacén de Datos

Dimensión Establecimiento (*store*) :

- ✓ la dimensión Almacén representa la información geográfica básica.
- ✓ esta dimensión suele ser creada explícitamente recopilando información *externa* que sólo tiene sentido en el A.D y que no la tiene en un OLTP (número de habitantes de la ciudad del establecimiento, caracterización del tipo de población del distrito, ...)
- ✓ atributos frecuentes: identificador (código interno), nombre, dirección, distrito, región, ciudad, país, director, teléfono, fax, tipo de almacén, superficie, fecha de apertura, fecha de la última remodelación, superficie para congelados, superficie para productos frescos, datos de la población del distrito, zona de ventas, ...
- ✓ jerarquías:
  - establecimiento - distrito - ciudad - región - país (jerarquía geográfica)
  - establecimiento - zona_ventas - región_ventas (jerarquía de ventas)

Diseño de un Almacén de Datos

Tiempo

id_fecha
día
semana
mes
año
día_semana
día_mes
trimestre
festivo
....

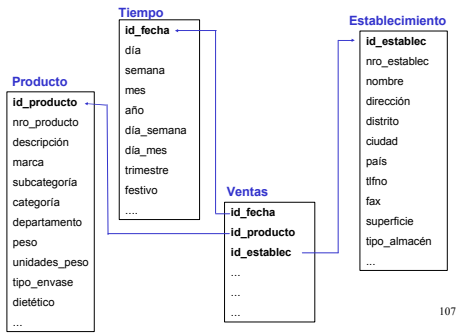
Establecimiento

id_establec
nro_establec
nombre
dirección
distrito
ciudad
país
tlfno
fax
superficie
tipo_almacén
...

Producto

id_producto
nro_producto
descripción
marca
subcategoría
categoría
departamento
peso
unidades_peso
tipo_envase
dietético
...

Diseño de un Almacén de Datos



Diseño de un Almacén de Datos

Paso 4. Decidir la información a almacenar sobre el proceso.

Hechos: información (sobre la actividad) que se desea almacenar en cada tupla de la tabla de hechos y que será el objeto del análisis.

Precio

Unidades

Importe

....

Nota: algunos datos que en el OLTP coincidirían con valores de atributos de dimensiones, en el almacén de datos pueden representar hechos. (Ejemplo: el precio de venta de un producto).

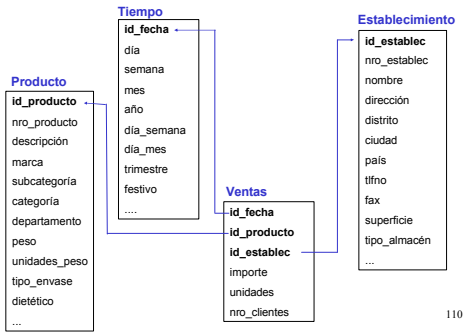
Diseño de un Almacén de Datos

Ejemplo: Cadena de supermercados.

Gránulo: "se desea almacenar información sobre las ventas diarias de cada producto en cada establecimiento de la cadena".

- importe total de las ventas del producto en el día
- número total de unidades vendidas del producto en el día
- número total de clientes distintos que han comprado el producto en el día.

Diseño de un Almacén de Datos



Diseño de un Almacén de Datos

Otras orientaciones de diseño:

- ✓ usar claves sin significado:
- ✓ evitar normalizar.
- ✓ incluir la dimensión Tiempo.
- ✓ dimensiones "que cambian".
- ✓ definición de agregados.

Diseño de un Almacén de Datos

Otras orientaciones de diseño:

✓uso de claves sin significado.

- en un almacén de datos debe evitarse el uso de las claves del sistema operacional.
- las claves de las dimensiones deben ser generadas artificialmente: claves de tipo entero (4 bytes) son suficiente para dimensiones de cualquier tamaño (2^{32} valores distintos).
- la dimensión TIEMPO debe tener también una clave artificial.

Inconvenientes del uso de las claves del sistema operacional:

- ✓ en el OLTP se puede decidir reutilizar valores de la clave no utilizados actualmente.
- ✓ en el OLTP se puede decidir cambiar la codificación de las claves.

Diseño de un Almacén de Datos

Otras Orientaciones de diseño:

✓evitar normalizar.

Si se define una tabla de dimensión para cada dimensión identificada en el análisis, es frecuente que entre el conjunto de atributos de la tabla aparezcan dependencias funcionales que hacen que la tabla no esté en 3ª F.N.

Evitar normalizar:

- ✓ el ahorro de espacio no es significativo
- ✓se multiplican los JOIN durante las consultas.

Diseño de un Almacén de Datos

Otras Orientaciones de diseño:

- ✓ siempre introducir la dimensión Tiempo.

En un almacén de Datos muchas consultas son restringidas y parametrizadas por criterios relativos a periodos de tiempo (último mes, este año, ...).

Diseño de un Almacén de Datos

Otras orientaciones de diseño:

✓dimensiones "que cambian".

Se considera relevante el caso en que, en el mundo real, para un valor de una dimensión, cambia el valor de un atributo que es significativo para el análisis sin cambiar el valor de su clave.

Ejemplo: En un A.D existe la dimensión CLIENTE. En la tabla correspondiente un registro representa la información sobre el cliente "María García" cuyo estado civil cambia el 15-01-1994 de *soltera* a *casada*. El estado civil del cliente es utilizado con frecuencia en el análisis de la información.

Existen tres estrategias para el tratamiento de los cambios en las dimensiones:

Tipo 1: Realizar la modificación.

Tipo 2: Crear un nuevo registro.

Tipo 3: Crear un nuevo atributo.

115

Diseño de un Almacén de Datos

Otras orientaciones de diseño:

✓definición de agregados.

¡En un almacén de datos es usual consultar información agregada!

El almacenamiento de datos agregados por distintos criterios de agregación en la tabla de hechos mejora la eficiencia del AD.

Estrategias de almacenamiento de datos agregados:

- ✓ Estrategia 1: definir nuevas tablas de hechos (resp. de dimensiones) para almacenar la información agregada (resp. la descripción de los niveles de agregación).
- ✓ Estrategia 2: insertar en la tabla de hechos (resp. dimensiones) tuplas que representan la información agregada (resp. los niveles de agregación).

116

Líneas de Investigación Abiertas

Resúmenes:

- ✓ Widom, J. *Research problems in data warehousing*
  Actas de la International Conference on Information and Knowledge Management (CIKM95), ACM Press. 1995
- ✓ Chaudhuri, S., Dayal, U. *An overview of data warehousing and OLAP technology*.
  SIGMOD Records. 26(1), pp. 65-74, 1997.
- ✓ Wu, Ch., Buchmann, P. *Research issues in data warehousing*
  Datebanksysteme in Büro, Technik und Wissenchaft (BTW), Informatik Aktuell, pp. 61-62. Springer, 1997

117

Líneas de Investigación Abiertas

Resúmenes:

- ✓ Samtani, S., Kumar, V., Kambayashi, Y.
  Recent advances and research problems in data warehousing.
  Actas de la International Conference on Conceptual Modeling (ER)
  LNCS 1507,Springer, 1998
- ✓ Gardner, S.R.
  Building the data warehouse.
  Communications of the ACM 41(9), pp. 52-60, 1998.
- ✓ Dinter, B., Sapia, C. Hölfing, G., Blaschka, M.
  OLAP market and research: initiating the cooperation.
  Journal of Computer Science and Information Management, 2(3), 1999

118

Líneas de Investigación Abiertas

Conferencias especializadas en DW:

- ✓ International Worshop on Data Warehousing and OLAP. (DOLAP)
- ✓ International Workshop on Data Warehouse and Data Mining. (DWDM)
- ✓ Interantional Workshop on Design and Management of Data Warehouses. (DMDW)
- ✓ International Conference on Data Warehousing and Knowledege Discovery. (DaWaK)

119

Líneas de Investigación Abiertas

Conferencias especializadas en BD:

- ✓ International Conference of Very Large Databases. (VLDB)
- ✓ International Conference on Data Engineering. (ICDE)
- ✓ Interantional Conference on Conceptual Modeling. (ER)
- ✓ International Conference on Extending Database Technology (EDBT).
- ✓ International Conference on Database Theory (ICDT).

120

Líneas de Investigación Abiertas

Direcciones de interés:

- http://www.cs.toronto.edu/~mendel/dwbib.html
- http://www.olapcouncil.org/research/
- http://www.ceur-ws.org/
- http://www.cis.drexel.edu/faculty/song/dolap.html
- http://www-db.stanford.edu/warehousing/

121

Líneas de Investigación Abiertas

- ✓ Diseño de Almacenes de Datos: modelos conceptuales, metodogías de diseño.
- ✓ Carga y ETL: recuperación de fallos durante la carga. Planificación de cargas y refrescos.
- ✓ Limpieza y Transformación
- ✓ Mantenimiento de Almacenes de Datos: mantenimiento de vistas materializadas.
- ✓ Implementación de Almacenes de Datos.
- ✓ Diseño Físico, optimizaciones para ROLAP, estructuras para MOLAP.
- ✓ Repartición de tareas OLAP entre el cliente y el servidor.

122