

# Data warehouse (DW)

Dr. José Luis Zechinelli Martini

[jose Luis.zechinelli@udlap.mx](mailto:jose Luis.zechinelli@udlap.mx)

LIS – 3071

Este material está basado en los cursos del Dr. José Hernández Orallo, Universidad Politécnica de Valencia y de la Dra. Ofelia Cervantes Villagómez, UDLAP

## Plan

- Almacenes de datos:
  - Definición y objetivo
  - Aplicaciones
  - Arquitectura
- Diseño y construcción:
  - Modelo multidimensional
  - Pasos en el diseño de un almacén de datos
  - Esquemas de representación
- Tecnología OLAP:
  - Consultas y análisis
  - ROLAP, MOLAP, HOLAP
- Carga y mantenimiento

## Almacenes de datos (1)

- Un almacén de datos (*data warehouse*) es una colección de datos:
  - orientada a un dominio
  - integrada
  - no volátil
  - variante en el tiempo
- Para ayudar en la toma de decisiones [Immon 1992, 1996]

3

## Almacenes de datos (2)

- **Objetivo:** Análisis de datos para soportar la toma de decisiones:
  - Generalmente, la información que se necesita analizar 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 – *online transaccional processing* – OLTP)

4

## Almacenes de datos y OLAP

- La tecnología OLAP (*online analytical processing*):
  - Generalmente se asocia a los almacenes de datos
  - Aunque se pueden tener almacenes de datos sin OLAP y viceversa
- Los almacenes de datos y las técnicas OLAP:
  - Son las maneras más efectivas y tecnológicamente más avanzadas para integrar, transformar y combinar los datos
  - Facilitan al usuario o a otros sistemas el análisis de la información

5

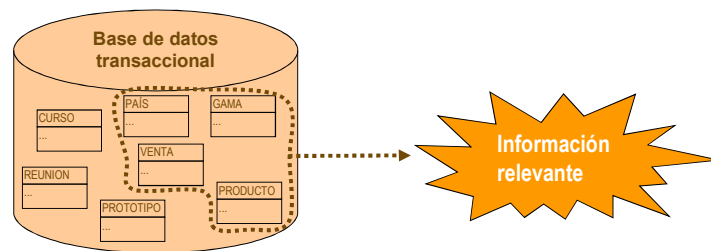
## Aplicaciones (1)

DOMINIO	APLICACIONES
Supermercados	Análisis del comportamiento de los consumidores Ventas cruzadas
Bancos	Investigación en contextos de fraude Crédito sujeto a identificación
Compañías aseguradoras	Modelos de tasación y de selección Análisis de causas de accidentes
Líneas aéreas y automóviles	Control de calidad Orden de previsión
Telecomunicaciones	Simulación de tasación

6

## Aplicaciones (2)

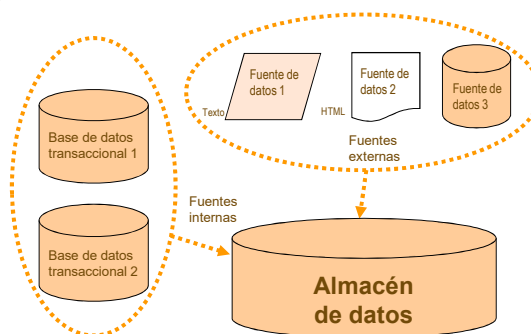
- Un almacén de datos se diseña:
  - Para consultar eficientemente información de las actividades básicas (ventas, compras, producción, ...) de una organización
  - No para soportar los procesos que se realizan en ella (gestión de pedidos, facturación, ...)



7

## Aplicaciones (3)

- Un almacén de datos integra datos recogidos de diferentes sistemas operacionales de la organización y/o fuentes externas



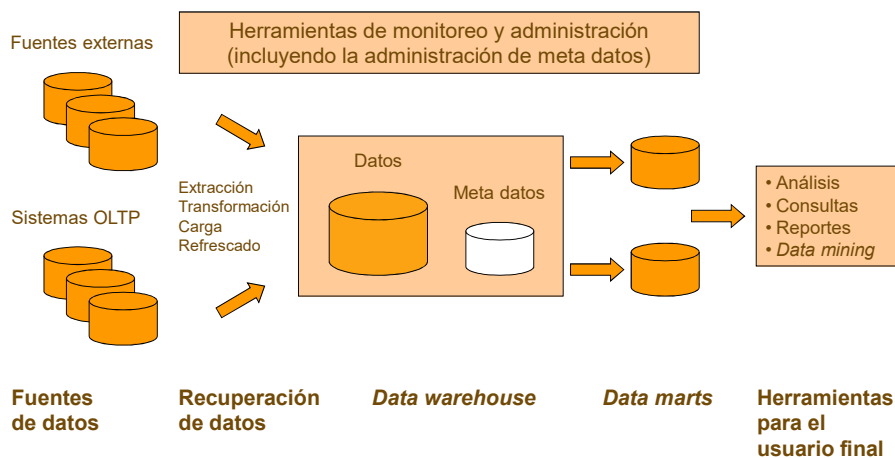
8

## Aplicaciones OLTP vs. OLAP

Sistemas operacionales	Almacenes de datos
<ul style="list-style-type: none"> <li>- Almacena datos actuales</li> <li>- Almacena datos de detalle</li> <li>- Bases de datos medianas (100Mb – 1Gb)</li> <li>- Los datos son dinámicos (actualizables)</li> <li>- Los procesos (transacciones) son repetitivos</li> <li>- El número de transacciones es elevado</li> <li>- Tiempo de respuesta pequeño (segundos)</li> <li>- Dedicado al procesamiento de transacciones</li> <li>- Orientado a los procesos de la organización</li> <li>- Soporta decisiones diarias</li> <li>- Sirve a muchos usuarios (administrativos)</li> </ul>	<ul style="list-style-type: none"> <li>- Almacena datos históricos</li> <li>- Almacena datos de detalle y datos agregados a distintos niveles</li> <li>- Bases de datos grandes (100Gb – 1Tb)</li> <li>- Los datos son estáticos</li> <li>- Los procesos (consultas) no son previsibles</li> <li>- El número de transacciones es bajo o medio</li> <li>- Tiempo de respuesta variable (segundos-horas)</li> <li>- Dedicado al análisis de datos</li> <li>- Orientado a la información relevante</li> <li>- Soporta decisiones estratégicas</li> <li>- Sirve a técnicos de dirección (gerentes)</li> </ul>

9

## Arquitectura general (1)



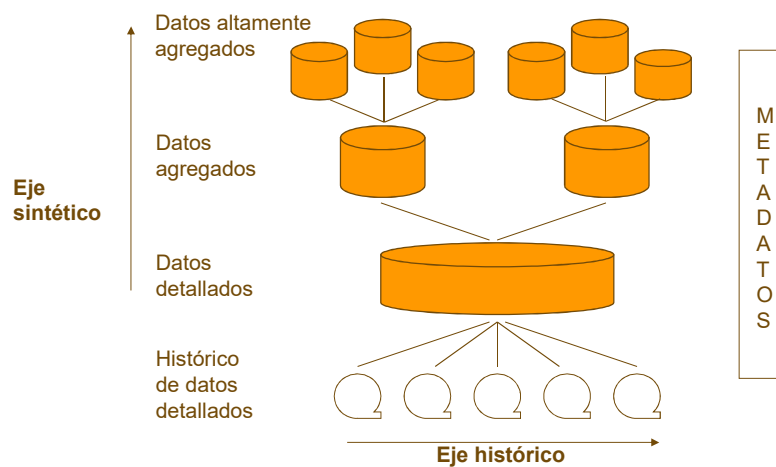
10

## Arquitectura general (2)

- **Sistema ETL** (*extract-transform-load*), realiza:
  - La extracción de las fuentes datos (transaccionales o externas)
  - El filtrado y transformación de los datos (limpieza, consolidación, ...)
  - La carga inicial del almacén (ordenación, agregaciones, ...)
  - El refresco del almacén (operación periódica), 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 y 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, ...

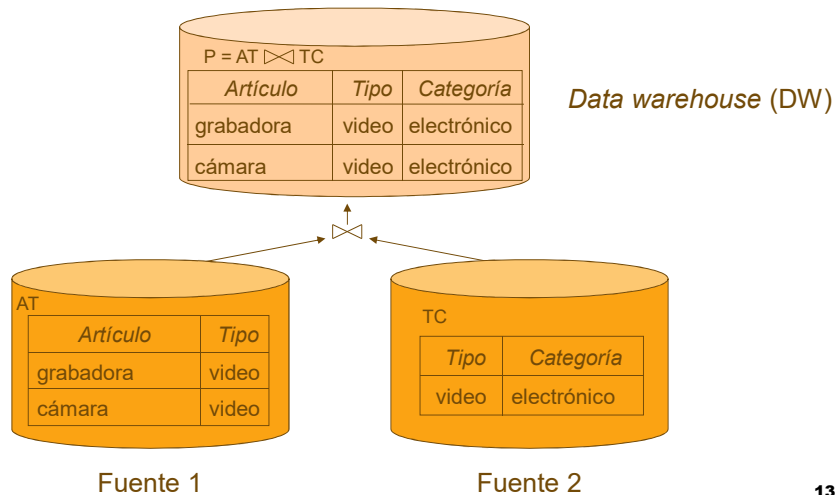
11

## Arquitectura interna



12

## Ejemplo de construcción



13

## Aspectos de construcción (1)

- Carga masiva de **datos**
- **Soluciones:**
  - Grupos de datos seleccionados
  - Ejecución paralela de los cálculos y de las entradas-salidas

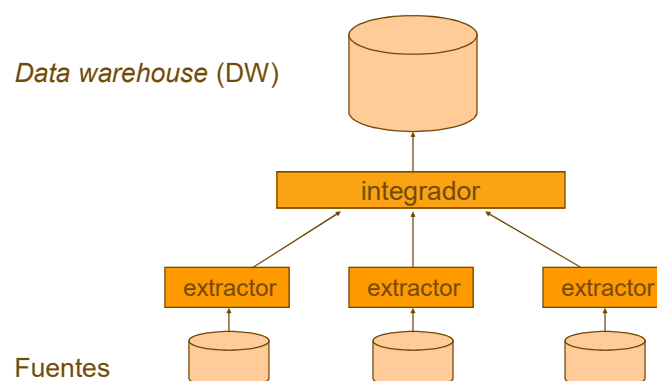
14

## Aspectos de construcción (2)

- Carga masiva de los **esquemas de dimensión** de los datos
- **Heterogeneidad** de fuentes:
  - Modelos de los datos
  - Esquema
  - Tipos de datos
  - Valores
- **Soluciones:**
  - Homogeneización de los datos
  - Técnicas de "limpieza de datos"

15

## Aspectos de construcción (3)

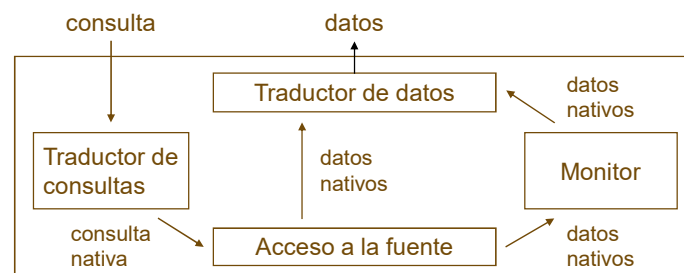


16



## Extractores

- Componentes (*wrappers*) de software que:
  - Traduce datos y consultas de un modelo de datos a otro
  - Detecta los cambios que se ejecutan en las fuentes



17

## Integradores

- Componentes de software que:
  - Descompone la «consulta» que define al almacén de datos como varias subconsultas dirigidas a las fuentes
  - Fusiona los datos que provienen de fuentes diferentes
  - Carga los datos fusionados en el almacén de datos

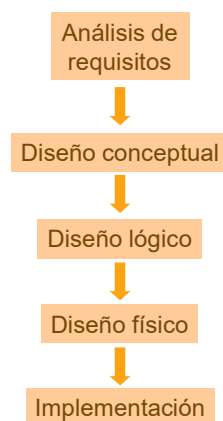
18

# Plan

- ✓ Almacenes de datos:
  - Definición y objetivo
  - Aplicaciones
  - Arquitectura
- Diseño y construcción:
  - Modelo multidimensional
  - Pasos en el diseño de un almacén de datos
  - Esquemas de representación
- Tecnología OLAP:
  - Consultas y análisis
  - ROLAP, MOLAP, HOLAP
- Carga y mantenimiento

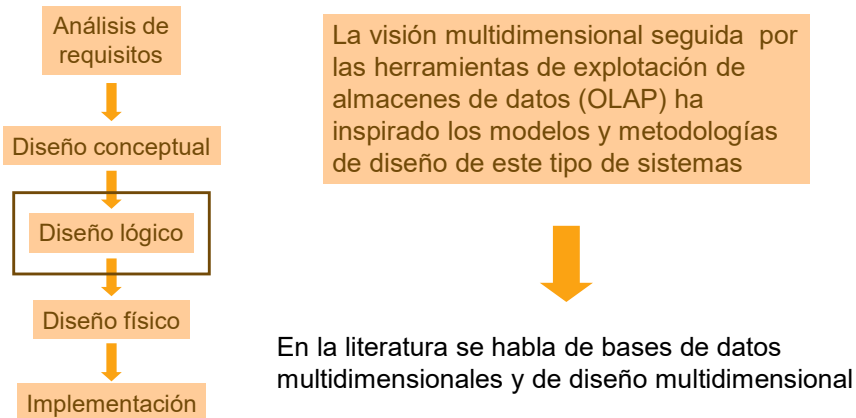
19

## Diseño de un SI (1)



20

## Diseño de un SI (2)



21

## Modelo multidimensional (1)

- En un esquema multidimensional se representa una **actividad** que es objeto de análisis (hecho) y las entidades 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 **entidad** se representa por un conjunto de atributos (atributos de dimensión)

22

## Modelo multidimensional (2)

- 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, ...)

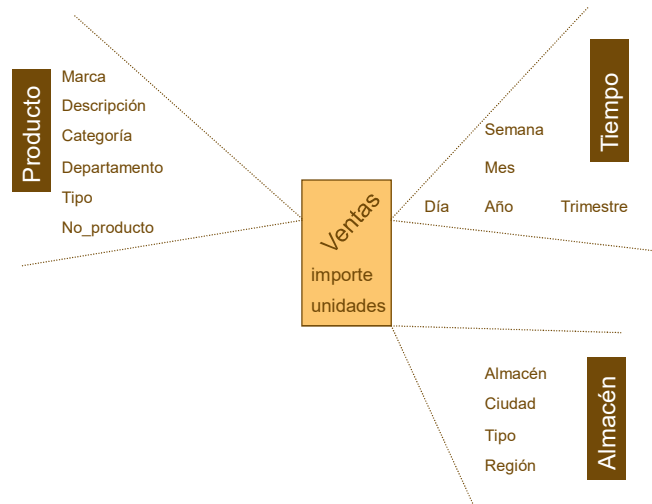
23

## Ejemplo (1)

- **Organización:** Cadena de supermercados
  - **Actividad objeto de análisis:** Ventas de productos
  - **Información registrada sobre una venta:** Del producto "Red Bull 250 ml" se han vendido en el almacén "Almacén no.1" el día 17.7.2011, 25 unidades por un importe de \$292.00 pesos
  - **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**

24

## Ejemplo (2)



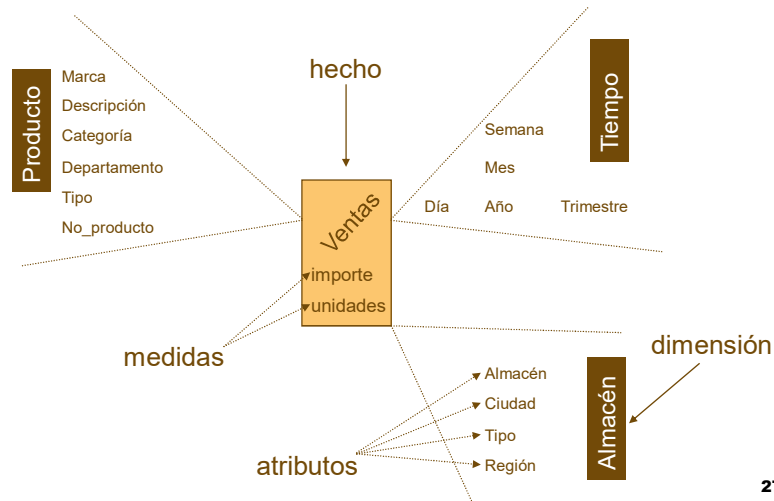
25

## Ejemplo (3)



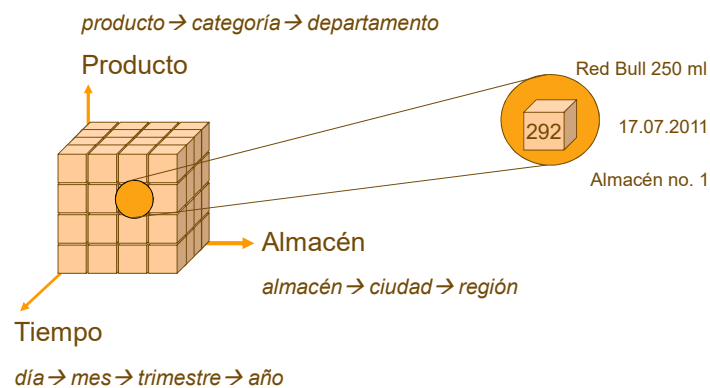
26

## Ejemplo (4)



27

## Ejemplo: Esquema multidimensional

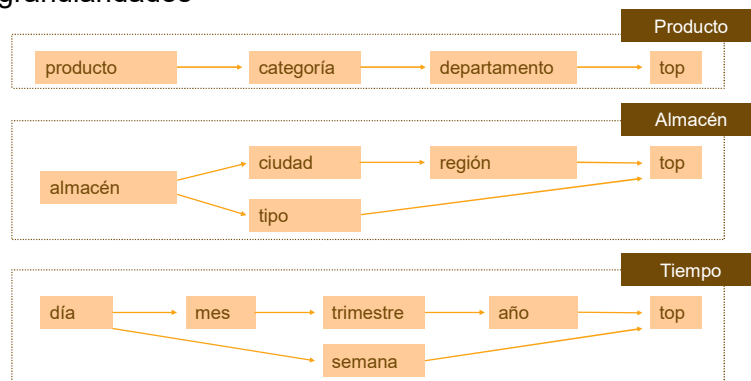


Un nivel de agregación para un conjunto de dimensiones se denomina cubo

28

## Ejemplo: Esquemas de dimensión

- Cada entidad es representada por un esquema de dimensión, i.e., una **jerarquía** que permite observar los datos bajo varias granularidades



29

## Plan

- ✓ Almacenes de datos:
  - Definición y objetivo
  - Aplicaciones
  - Arquitectura
- Diseño y construcción:
  - ✓ Modelo multidimensional
  - Pasos en el diseño de un almacén de datos
  - Esquemas de representación
- Tecnología OLAP:
  - Consultas y análisis
  - ROLAP, MOLAP, HOLAP
- Carga y mantenimiento

30

## Pasos en el diseño de un DW

- **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

31

## Paso 1: Elegir un proceso (1)

- **Proceso:** Actividad de la organización soportada por una aplicación OLTP de la cual se puede **extraer información** con el propósito de construir el almacén de datos:
  - Pedidos (de clientes)
  - Compras (a proveedores)
  - Facturación
  - Envíos
  - Ventas
  - Inventario
  - ...

32



## Paso 1: Elegir un proceso (2)

- **Ejemplo:** Cadena de supermercados
  - Cadena de supermercados con 300 almacenes en la que se comercializan unos 30,000 productos distintos
- **Actividad:** Ventas
  - La actividad a modelar son las ventas de productos en los almacenes de la cadena

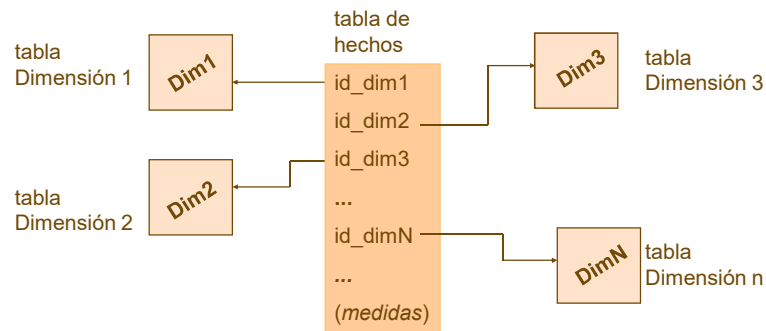
33

## Paso 2: Decidir el gránulo (1)

- **Gránulo:** Es el nivel de detalle con el que se necesita almacenar información sobre la actividad a modelar:
  - Define el **nivel atómico de datos** en el almacén de datos
  - Determina el **significado de la información** en la tabla de hechos
  - Es **determinado por las dimensiones** básicas del esquema:
    - transacción en el OLTP
    - información diaria
    - información semanal
    - información mensual ...

34

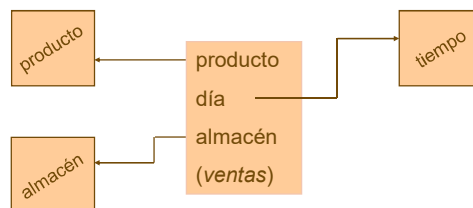
## Paso 2: Decidir el gránulo (2)



35

## Paso 2: Decidir el gránulo (3)

- **Ejemplo:** Cadena de supermercados
- **Gránulo:** Se desea almacenar información sobre las ventas diarias de cada producto en cada almacén de la cadena:
  - Determina las dimensiones básicas del esquema
  - Define el significado de las *tuplas* de la tabla de hechos



36

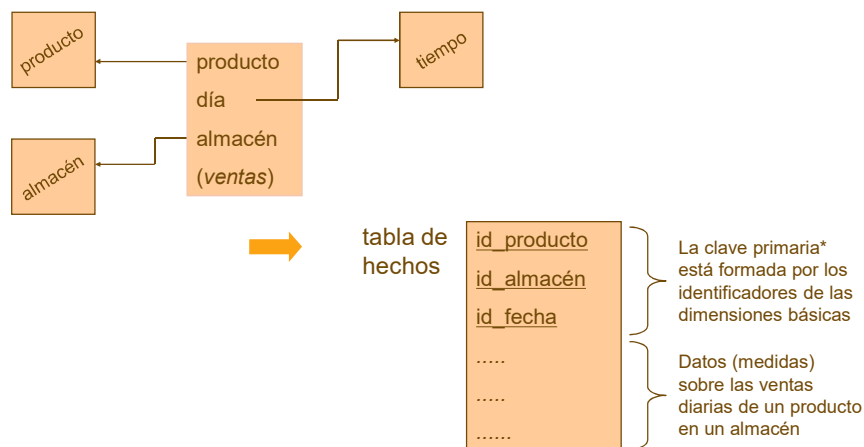
## Paso 2: Decidir el gránulo (4)

- **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, e.g., 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 fino** (gránulo) no porque se vaya a consultar el almacén a ese nivel, sino porque ello permite estudiar y clasificar (analizar) la información desde muchos puntos de vista

37

## Paso 2: Decidir el gránulo (5)



\* pueden existir excepciones a esta regla general

38

## Paso 3: Identificar dimensiones (1)

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

39

## Paso 3: Identificar dimensiones (2)

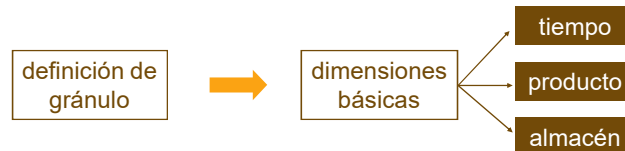
tabla  
Dimensión 1



40

## Paso 3: Identificar dimensiones (3)

### ■ Ejemplo: Cadena de supermercados



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

41

## Paso 3: Dimensión Tiempo (1)

### ■ La dimensión Tiempo está presente en todo DW:

- Normalmente un DW 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:

- **No. de día, no. de semana, no. de año:** Valores absolutos del calendario gregoriano que permiten hacer ciertos cálculos aritméticos
- **Día de la semana {lunes, martes, miércoles, ...}:** Permite hacer análisis sobre los días concretos de la semana, e.g., ventas en sábado, ventas en lunes, ...

42

## Paso 3: Dimensión Tiempo (2)

- Atributos frecuentes (continuación):
  - **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 o 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, otoño, invierno }
  - **Evento especial**: permite marcar días de eventos especiales (final de futbol, elecciones, ...)
- Jerarquías: día → mes → trimestre → año; día → semana

43

## Paso 3: 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?)
  - **Atributos frecuentes**: Identificador (código estándar), descripción, marca, categoría, departamento, peso, unidades de peso, tipo de envase, producto dietético, fórmula, unidades por envase, ...
  - **La dimensión Producto debe contener el mayor número posible de atributos descriptivos que permitan un análisis flexible; un número frecuente es de 50 atributos**
- Jerarquías: producto → categoría → departamento

44

## Paso 3: Dimensión Almacén

- La dimensión Almacén representa la información geográfica básica:
  - Esta dimensión suele ser creada explícitamente recopilando información externa que sólo tiene sentido en el DW y que no la tiene en un OLTP, e.g., número de habitantes de la ciudad de la tienda, caracterización del tipo de población de la ciudad, ...
  - **Atributos frecuentes:** Identificador (código interno), nombre, dirección, ciudad, región, 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:
  - almacén → ciudad → región (jerarquía geográfica)
  - almacén → tipo (jerarquía de ventas)

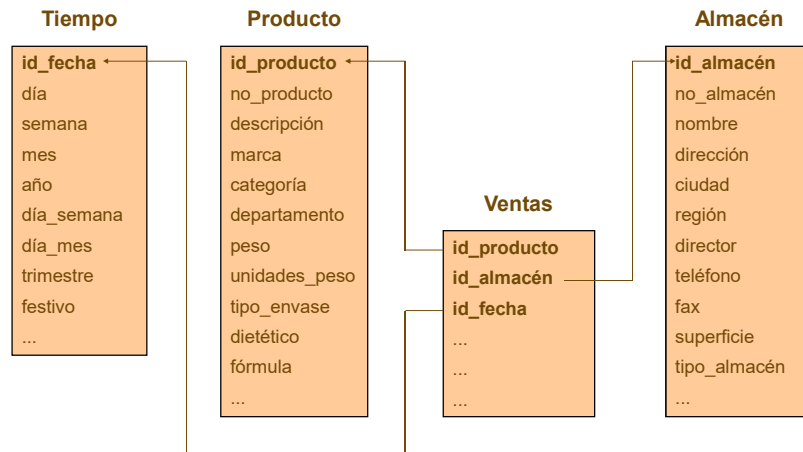
45

## Paso 3: Esquema relacional (1)

Tiempo	Producto	Almacén
id_fecha	id_producto	id_almacén
día	no_producto	no_almacén
semana	descripción	nombre
mes	marca	dirección
año	categoría	ciudad
día_semana	departamento	región
día_mes	peso	director
trimestre	unidades_peso	teléfono
festivo	tipo_envase	fax
...	dietético	superficie
	fórmula	tipo_almacén
	...	...

46

## Paso 3: Esquema relacional (2)



47

## Paso 4: Decidir que almacenar (1)

- **Medidas o hechos:** Información (sobre la actividad) que se desea almacenar en cada *tupla* de la tabla de hechos y que será el objeto del análisis:
  - Importe
  - Unidades
  - Número de clientes
  - ...

**Nota:** Algunos datos de la aplicación OLTP coincidirán con los valores de los atributos de las dimensiones, en el almacén de datos representan medidas, e.g., el importe de venta de un producto

48

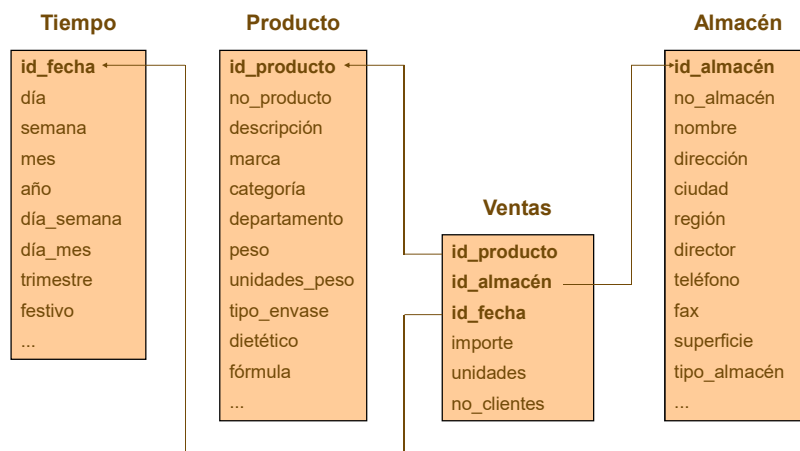


## Paso 4: Decidir que almacenar (2)

- **Ejemplo:** Cadena de supermercados
- **Gránulo:** Se desea almacenar información sobre las ventas diarias de cada producto en cada almacén de la cadena
- **Medidas** requeridas para representar las ventas:
  - El importe total de las ventas del producto en el día
  - El número total de unidades vendidas del producto en el día
  - El número total de clientes distintos que han comprado el producto en el día

49

## Paso 4: Esquema relacional



50

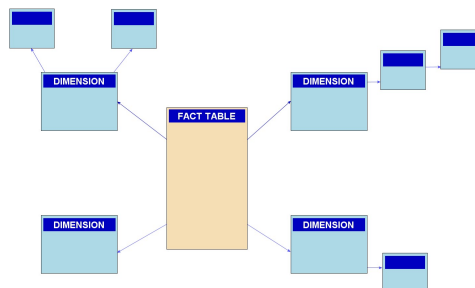
# Plan

- ✓ Almacenes de datos:
  - Definición y objetivo
  - Aplicaciones
  - Arquitectura
- Diseño y construcción:
  - ✓ Modelo multidimensional
  - ✓ Pasos en el diseño de un almacén de datos
  - Esquemas de representación
- Tecnología OLAP:
  - Consultas y análisis
  - ROLAP, MOLAP, HOLAP
- Carga y mantenimiento

51

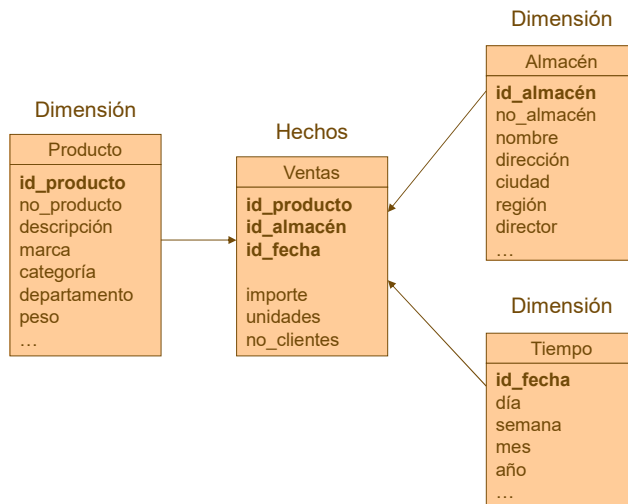
# Esquemas de representación

- **Estrella:** si las dimensiones son representadas usando una sola tabla
- **Copo de nieve:** si las tablas que representan las dimensiones están normalizadas



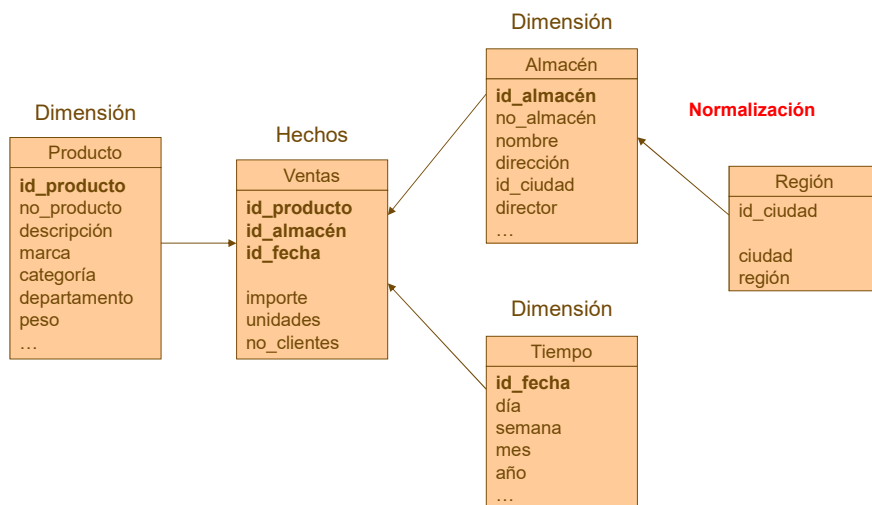
52

## Esquema en estrella



53

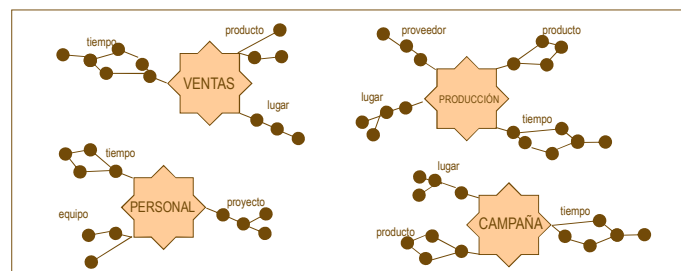
## Copo de nieve (snow flake)



54

## Constelación

- Contiene múltiples tablas de hechos:
  - Las tablas de dimensiones pueden estar compartidas entre más de una tabla de hechos
  - Cuando el número de las tablas vinculadas aumenta, la arquitectura puede llegar a ser muy compleja y difícil de mantener



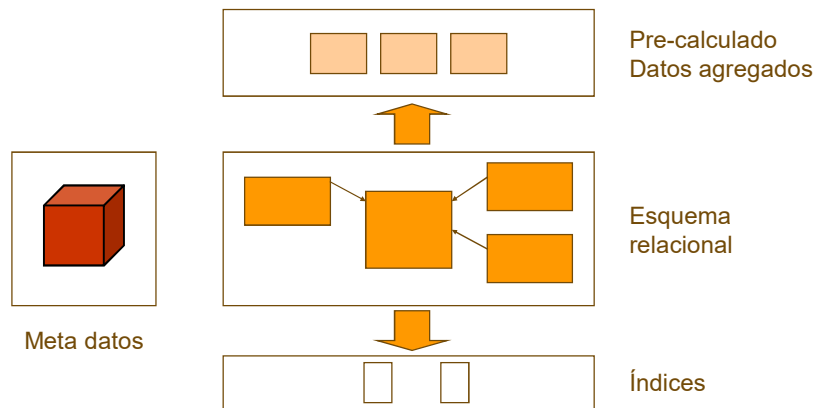
55

## Data marts

- **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
- El almacén de datos puede estar formado por **varios data marts** y, opcionalmente, por tablas adicionales

56

## *Data warehouse* relacional



57

## Plan

- ✓ Almacenes de datos:
  - Definición y objetivo
  - Aplicaciones
  - Arquitectura
- ✓ Diseño y construcción:
  - Modelo multidimensional
  - Pasos en el diseño de un almacén de datos
  - Esquemas de representación
- Tecnología OLAP:
  - Consultas y análisis
  - ROLAP, MOLAP, HOLAP
- Carga y mantenimiento

58

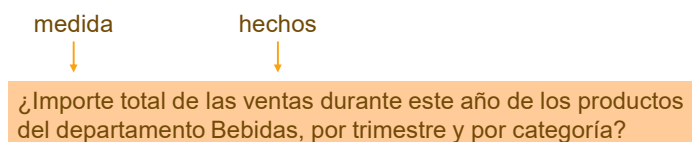
## Consulta y análisis

- Las herramientas 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, e.g., mediante una sentencia SELECT-FROM-WHERE

59

## Consultas OLAP (1)

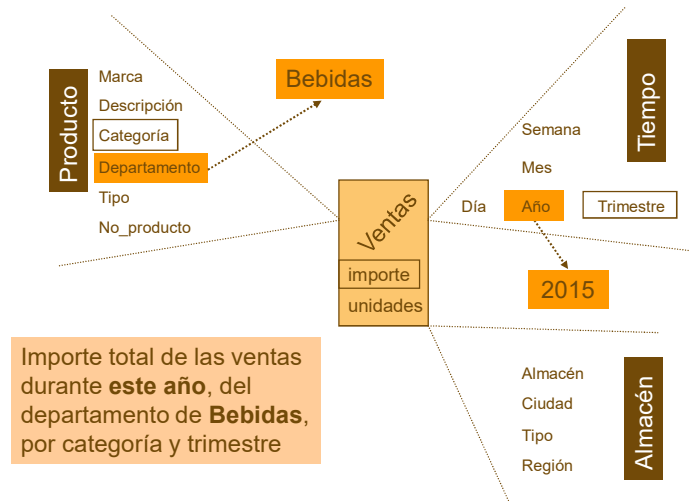
- Una consulta consiste generalmente en obtener **medidas** sobre la tabla de **hechos** parametrizadas por atributos de las dimensiones y restringidas por condiciones impuestas sobre las dimensiones:



- **Restricciones:** Ventas durante este año, productos del departamento Bebidas
- **Parámetros de la consulta:** Por trimestre y por categoría de producto

60

## Consultas OLAP (2)



61

## Consultas OLAP (3)

Categoría	Trimestre	Importe
Refrescos	T1	2000
Refrescos	T2	1000
Refrescos	T3	3000
Refrescos	T4	2000
Jugos	T1	1000
Jugos	T2	1500
Jugos	T3	8000
Jugos	T4	2400

- Presentación tabular (relacional) de los datos seleccionados
- Se asumen dos categorías en el departamento de bebidas: Refrescos y Jugos

62

## Consultas OLAP (4)

- Presentación matricial (multidimensional) de los datos seleccionados:

Categoría / Trimestre	T1	T2	T3	T4
Refrescos	2000	1000	3000	2000
Jugos	1000	15000	8000	2400

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

63

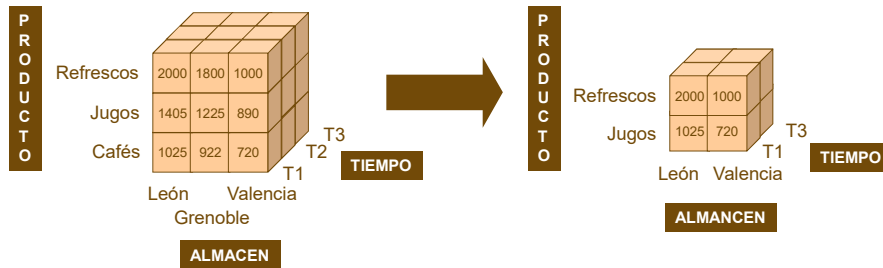
## Operadores OLAP

- Lo interesante no es poder realizar consultas que, en cierto modo, se pueden hacer con selecciones, proyecciones, uniones y agrupamientos tradicionales
- Lo realmente interesante de las herramientas OLAP son sus **operadores de refinamiento** o manipulación de consultas:
  - *Slice'n-Dice*
  - *Roll*
  - *Drill*
  - *Pivot*

64



## OLAP: *Slice'n-Dice* (1)

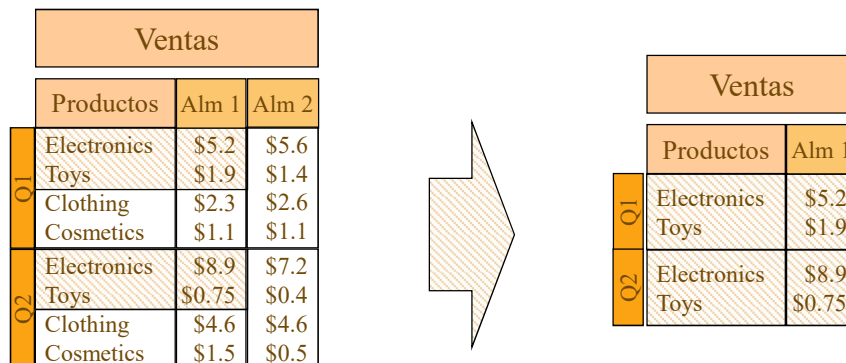


*Slice'n-Dice* →

Producto != "Cafés"  
Almacén != "Grenoble"  
Tiempo != "T2"

65

## OLAP: *Slice'n-Dice* (2)



66

## OLAP: *Roll, Drill* (1)

- 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

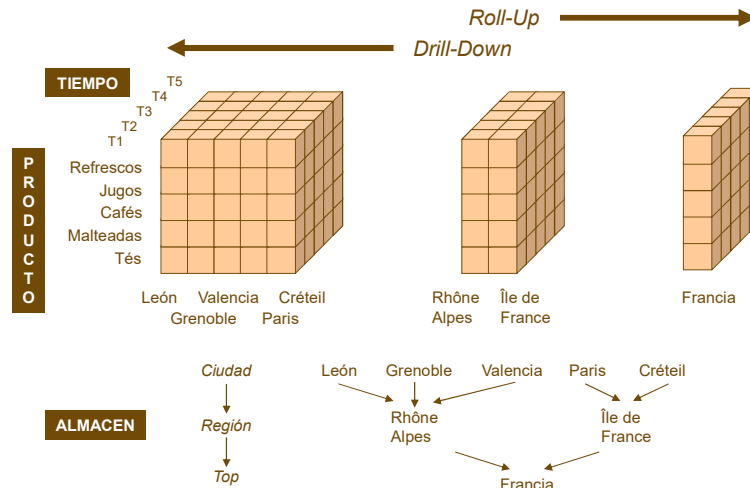
67

## OLAP: *Roll, Drill* (2)

- Las operaciones de agregación y disgregación se pueden hacer:
  - Sobre atributos de una dimensión sobre los que se ha definido una jerarquía: ***Roll-Up, Drill-Down***:
    - departamento ⇔ categoría ⇔ producto (Producto)
    - año ⇔ trimestre ⇔ mes ⇔ día (Tiempo)
  - Sobre dimensiones independientes: ***Roll-Across, Drill-Across***:
    - Producto ⇔ Almacén ⇔ Tiempo

68

## OLAP: *Roll-Up, Drill-Down*



69

## OLAP: *Drill-Down*

Categoría	Trimestre	Importe		Categoría	Trimestre	Mes	Importe
Refrescos	T1	2000	➔	Refrescos	T1	Enero	1000
Refrescos	T2	1000		Refrescos	T1	Febrero	500
Refrescos	T3	3000		Refrescos	T1	Marzo	500
Refrescos	T4	2000					
Jugos	T1	1000					
Jugos	T2	1500					
Jugos	T3	8000					
Jugos	T4	2400					

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

70

## OLAP: *Drill-Across* (1)

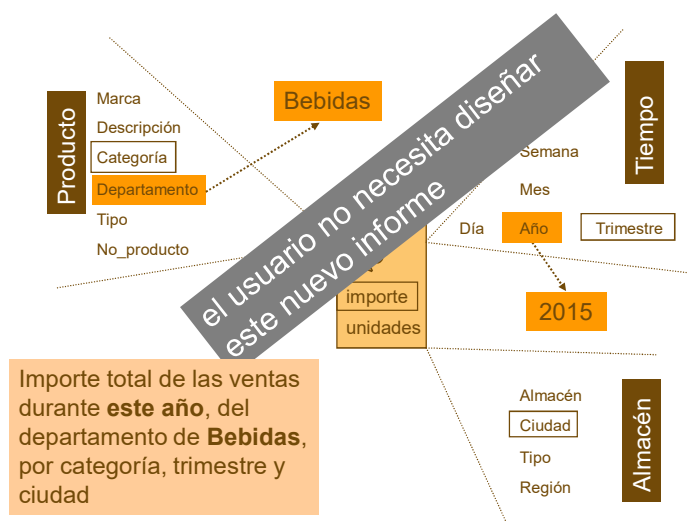
- Si se desea introducir la dimensión Almacén en el análisis anterior e incluir un nuevo criterio de agrupación sobre la ciudad del almacén:

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

- **Restricciones:** Ventas durante este año, productos del departamento Bebidas
- **Parámetros de la consulta:** Por trimestre, por categoría de producto y por ciudad del almacén

71

## OLAP: *Drill-Across* (2)



72

## OLAP: *Drill-Across* (3)

Categoría	Trimestre	Importe		Categoría	Trimestre	Ciudad	Importe
Refrescos	T1	2000	➔	Refrescos	T1	Valencia	1000
Refrescos	T2	1000	➔	Refrescos	T2	León	1000
Refrescos	T3	3000		Refrescos	T2	Valencia	300
Refrescos	T4	2000		Refrescos	T2	León	700
Jugos	T1	1000					
Jugos	T2	1500					
Jugos	T3	8000					
Jugos	T4	2400					

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

73

## OLAP: *Roll-Across* (1)

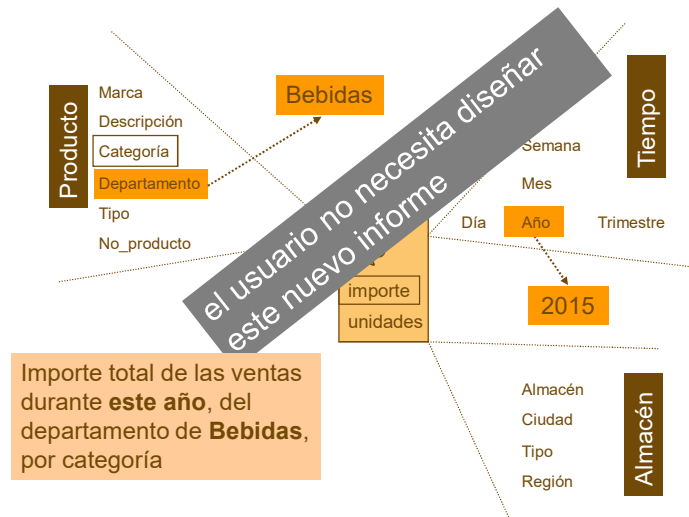
- Si se desea eliminar el criterio de agrupación sobre la dimensión Tiempo en la consulta original:

¿Importe total de las ventas durante este año de los productos del departamento Bebidas por categoría?

- **Restricciones:** Ventas durante este año, productos del departamento Bebidas
- **Parámetros de la consulta:** Por categoría de producto

74

## OLAP: *Roll-Across* (2)



75

## OLAP: *Roll-Across* (3)

Categoría	Trimestre	Importe
Refrescos	T1	2000
Refrescos	T2	1000
Refrescos	T3	3000
Refrescos	T4	2000
Jugos	T1	1000
Jugos	T2	1500
Jugos	T3	8000
Jugos	T4	2400


  

Categoría	Importe
Refrescos	8000
Refrescos	12900

76

## OLAP: *Pivot*

Ventas			
	Productos	Alm 1	Alm 2
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



Ventas			
	Productos	Q1	Q2
Alm 1	Electronics	\$5.2	\$8.9
	Toys	\$1.9	\$0.75
	Clothing	\$2.3	\$4.6
	Cosmetics	\$1.1	\$1.5
Alm 2	Electronics	\$5.6	\$7.2
	Toys	\$1.4	\$0.4
	Clothing	\$2.6	\$4.6
	Cosmetics	\$1.1	\$0.5

77

## 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: *Slice'n-Dice*, *Roll-Up*, *Drill-Down*, *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

78

# Plan

- ✓ Almacenes de datos:
  - Definición y objetivo
  - Aplicaciones
  - Arquitectura
- ✓ Diseño y construcción:
  - Modelo multidimensional
  - Pasos en el diseño de un almacén de datos
  - Esquemas de representación
- Tecnología OLAP:
  - ✓ Consultas y análisis
  - ROLAP, MOLAP, HOLAP
- Carga y mantenimiento

79

# ROLAP, MOLAP, HOLAP

- 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

80



## 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 gestión de almacenes de datos:
  - Índices de mapa de bits
  - Índices de JOIN
  - Técnicas de particionamiento de los datos
  - Optimizadores de consultas
  - Extensiones del SQL (*group-by cube*, operadores OLAP)

81

## Sistemas ROLAP (2)

**select**                      Producto, Almacén, Tiempo,  
                                 SUM(importe) as Ventas  
  
**from**                        Ventas  
  
**group by cube**          Producto, Almacén, Tiempo

Producto	Almacén	Tiempo	importe
videocámara	Grenoble	15/10/00	1500
videocámara	Lyon	15/10/00	2000
videocasetera	Grenoble	15/10/00	8000
videocasetera	Lyon	15/10/00	12000

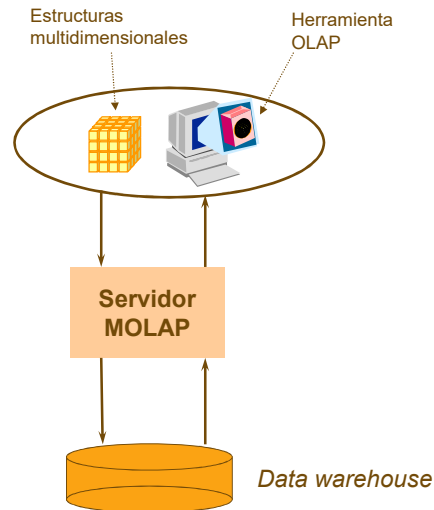
Group  
by cube

Producto	Almacén	Tiempo	Ventas
videocámara	Grenoble	15/10/00	1500
videocámara	Lyon	15/10/00	2000
videocámara	Grenoble	ALL	1500
videocámara	Lyon	ALL	2000
videocámara	ALL	15/10/00	3500
videocámara	ALL	ALL	3500
videocasetera	Grenoble	15/10/00	8000
videocasetera	Lyon	15/10/00	12000
videocasetera	Grenoble	ALL	8000
videocasetera	Lyon	ALL	12000
videocasetera	ALL	15/10/00	20000
videocasetera	ALL	ALL	20000
ALL	Grenoble	15/10/00	9500
ALL	Lyon	15/10/00	14000
ALL	Grenoble	ALL	9500
ALL	Lyon	ALL	14000
ALL	ALL	15/10/00	23500
ALL	ALL	ALL	23500

82

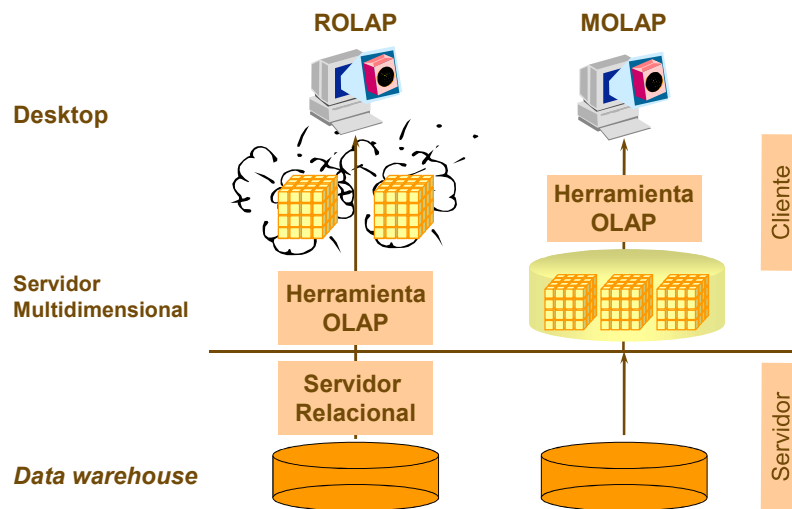
## Sistemas MOLAP

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



83

## Sistemas ROLAP y MOLAP



84

## Ventajas e inconvenientes

### ■ Sistemas 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

### ■ Sistemas MOLAP:

- Generalmente más eficientes que los ROLAP
- El coste de los cambios en la visión de los datos
- La construcción de las estructuras multidimensionales

85

## Plan

### ✓ Almacenes de datos:

- Definición y objetivo
- Aplicaciones
- Arquitectura

### ✓ Diseño y construcción:

- Modelo multidimensional
- Pasos en el diseño de un almacén de datos
- Esquemas de representación

### ✓ Tecnología OLAP:

- Consultas y análisis
- ROLAP, MOLAP, HOLAP

### ■ Carga y mantenimiento

86

## Carga y mantenimiento (1)

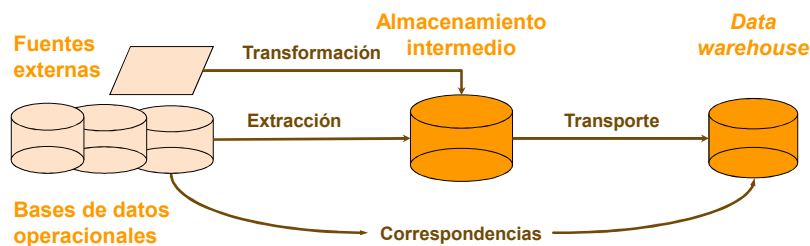
- El sistema encargado del mantenimiento del DW es el sistema ETT\* (*extraction, transformation, transportation*):
  - La construcción del sistema ETT es responsabilidad del equipo de desarrollo del almacén de datos
  - El sistema ETT es construido específicamente para cada almacén de datos, aproximadamente 50% del esfuerzo
  - En la construcción del ETT se pueden utilizar herramientas del mercado o programas diseñados específicamente
- Funciones del sistema ETT:
  - Carga inicial (*initial load*)
  - Mantenimiento o refresco periódico: Inmediato, diario, semanal, mensual, ... (*refreshment*)

\* Mejor conocido como ETL (*extract-transform-load*)

87

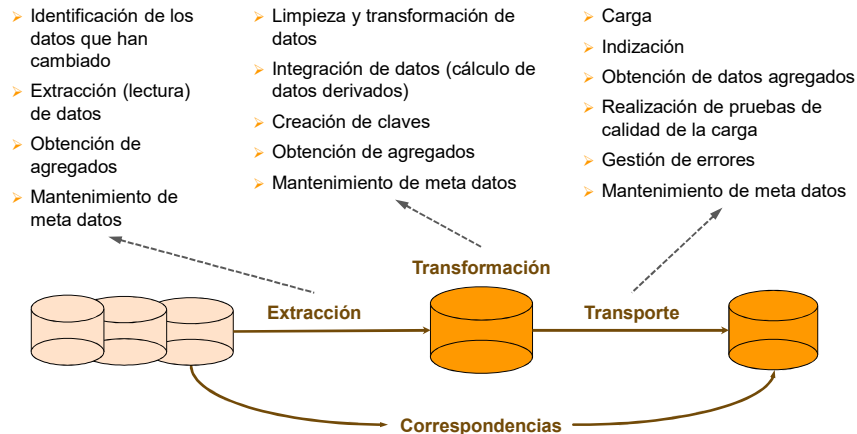
## Carga y mantenimiento (2)

- El **almacenamiento intermedio** permite:
  - Realizar transformaciones sin paralizar las bases de datos operacionales y el almacén de datos
  - Almacenar metadatos (correspondencias)
  - Facilitar la integración de fuentes externas



88

## Carga y mantenimiento (3)



89

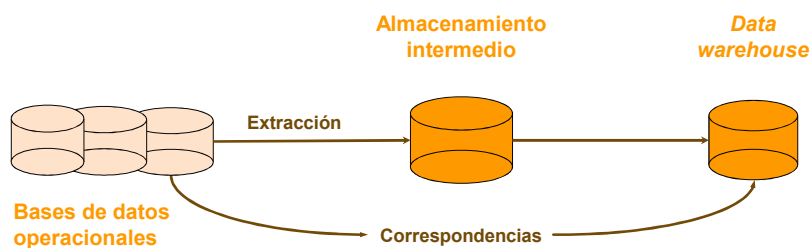
## Carga y mantenimiento (4)

- La **calidad de los datos** es la clave del éxito de un almacén de datos
- Estrategia** para tener datos de calidad:
  - Actuación de las correspondencias: 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*)

90

## Extracción (1)

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



91

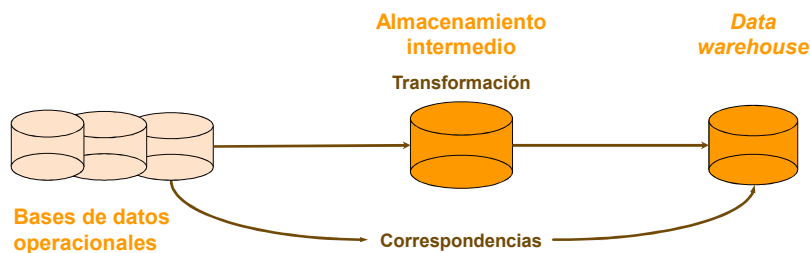
## Extracción (2)

- Lectura de los datos del sistema operacional:
  - Durante la carga inicial
  - Mantenimiento del DW
- Ejecución de la extracción:
  - Si los datos operacionales están mantenidos en un **SGBDR**, la extracción de datos se puede reducir a **consultas en SQL** o rutinas programadas
  - 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

92

## Transformación (1)

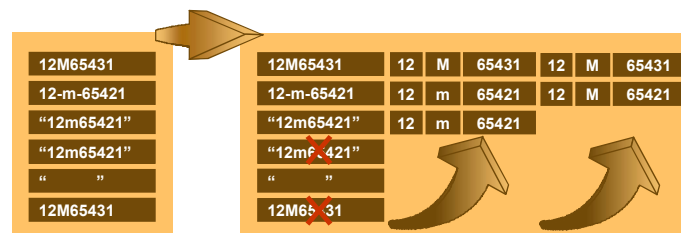
- **Transformar los datos extraídos** de las fuentes operacionales: Limpieza, estandarización (*cleansing*)
- **Calcular los datos derivados**: Aplicar las reglas de integración



93

## Transformación (2)

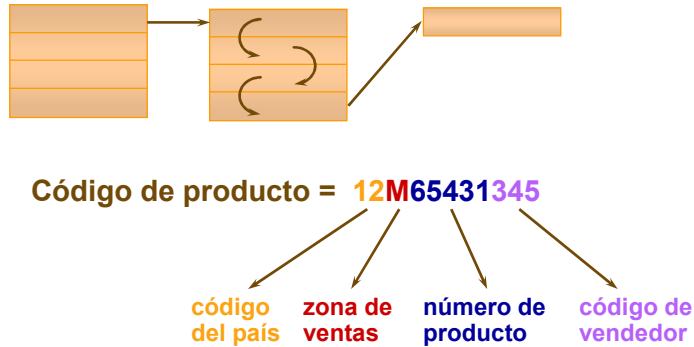
- En los datos operacionales existen **anomalías**: Datos desarrollos de forma independiente, fuentes heterogéneas, ...
- Eliminar anomalías:
  - **Limpieza de datos**: Corregir y completar datos, eliminar duplicados, eliminar datos no relevantes, ...
  - **Estandarización**: Codificación, formatos, unidades de medida, ...



94

## Transformación (3)

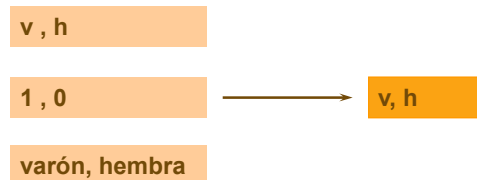
- **Claves con estructura:** Descomponer en valores atómicos



95

## Transformación (4)

- **Unificar codificaciones:** Existencia de codificaciones múltiples



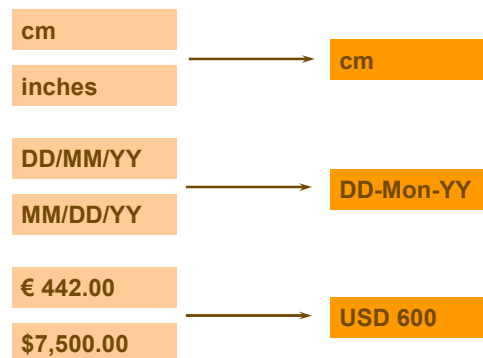
- Deben detectarse los valores erróneos

96



## Transformación (5)

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



97

## Transformación (6)

- **Valores duplicados:** Deben ser eliminados

- Subconsultas SQL
- Restricciones en el SGBDR



98

## Transformación (7)

- **Integridad referencial:** Debe reconstruirse

Departamento	Emp	Nombre	Departamento
10	1099	Smith	10
20	1289	Jones	20
30	1234	Doe	50
40	6786	Harris	60

99

## Transformación (8)

- **Creación de claves:** Eliminar información irrelevante

#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

Eliminar 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

100

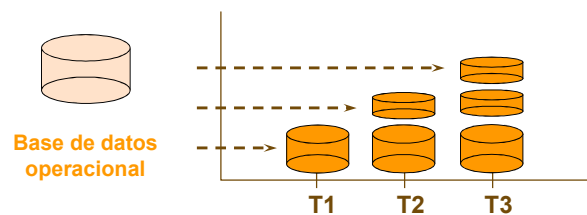
## Transporte – Carga (1)

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

101

## Transporte – Carga (2)

- Creación y mantenimiento:
  - **T1**: Crear el DW (base de datos)
  - **T2, T3**: En intervalos de tiempo fijos añadir cambios al DW
  - 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

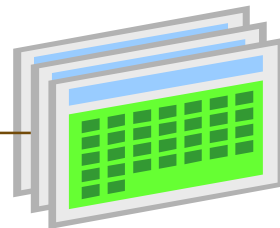


102

## Refresco (1)

- Función repetitiva que consiste en integrar los cambios producidos en las fuentes:
  - Los datos son relativos a un periodo de tiempo y deben ser incrementados periódicamente
  - Los datos son almacenados como fotos (*snapshots*)

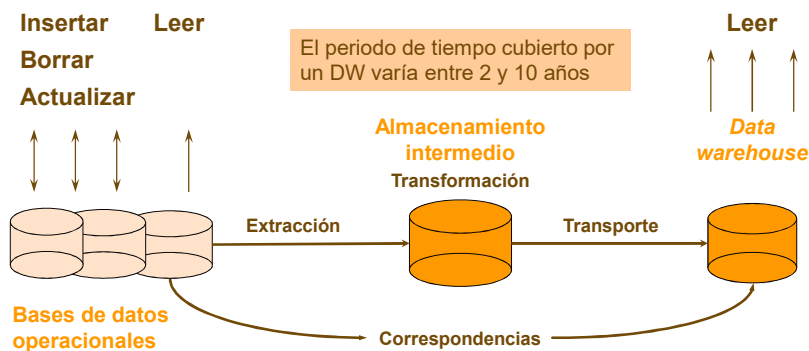
Tiempo	Datos
01/2015	Datos de enero
02/2015	Datos de febrero
03/2015	Datos de marzo



103

## Refresco (2)

- Los datos almacenados no son actualizados, sólo son incrementados, *i.e.*, no son volátiles



104

## Refresco (3)

- En el mantenimiento del DW antes de realizar la extracción es preciso identificar los 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

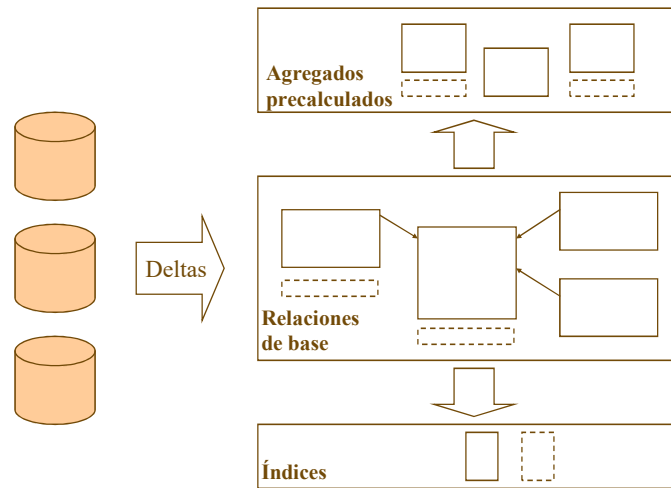
105

## Refresco (4)

- En general:
  - Evitar recalcular todo → volumen importante de datos
  - Integrar datos no existentes en el *data warehouse*
  - Actualizar los datos agregados de manera eficaz
- Para ello:
  - Identificar eficazmente las deltas → sólo se actualizan los datos que fueron modificados

106

# Refresco incremental



107