FUNDAMENTOS DE BASES DE DATOS

TEMA 6

Almacenes de Datos

T6 – Almacenes de Datos

Contenido

T6 – Almacenes de Datos

6.1. Concepto y arquitectura

*" Las organizaciones tienen un insaciable apetito de datos, pero frecuentemente les faltan las necesarias **enzimas** para digerirlos"*

" Neil Raden"

Presidente de Archer Decision Sciences

T6 – Almacenes de Datos

6.1. Concepto y arquitectura

CADENA DE VALOR

T6 – Almacenes de Datos

6.1. Concepto y arquitectura

INFORMACION = DATOS + SIGNIFICADO

El volumen de datos es irrelevante si a éstos no se les añade valor, es decir:

significado

T6 – Almacenes de Datos

6.1. Concepto y arquitectura

***Ejemplo*:**

2 8 9 1 1

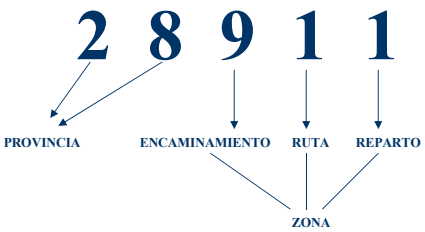
(es una simple sucesión de números sin ningún significado)

¡¡¡ NO ES INFORMACIÓN !!!

T6 – Almacenes de Datos

6.1. Concepto y arquitectura

Ahora vamos a añadir significado, convertirlo en **información**



De una sencilla sucesión de números, hemos creado **información**, mediante el significado que proporciona una estructura, la del Sistema de Codificación Postal

T6 – Almacenes de Datos

6.1. Concepto y arquitectura

OBJETIVO DE LOS S.I.

Proporcionar la **información necesaria** a la **persona adecuada** en el **momento oportuno**

T6 – Almacenes de Datos

6.1. Concepto y arquitectura

Significado orientado hacia las operaciones diarias

Proceso transaccional (OLTP) (*ON-LINE TRANSACTIONAL PROCESSING)*

Proceso analítico (OLAP) *(ON-LINE ANALYTICAL PROCESSING)*

Significado orientado hacia el negocio

T6 – Almacenes de Datos

6.1. Concepto y arquitectura

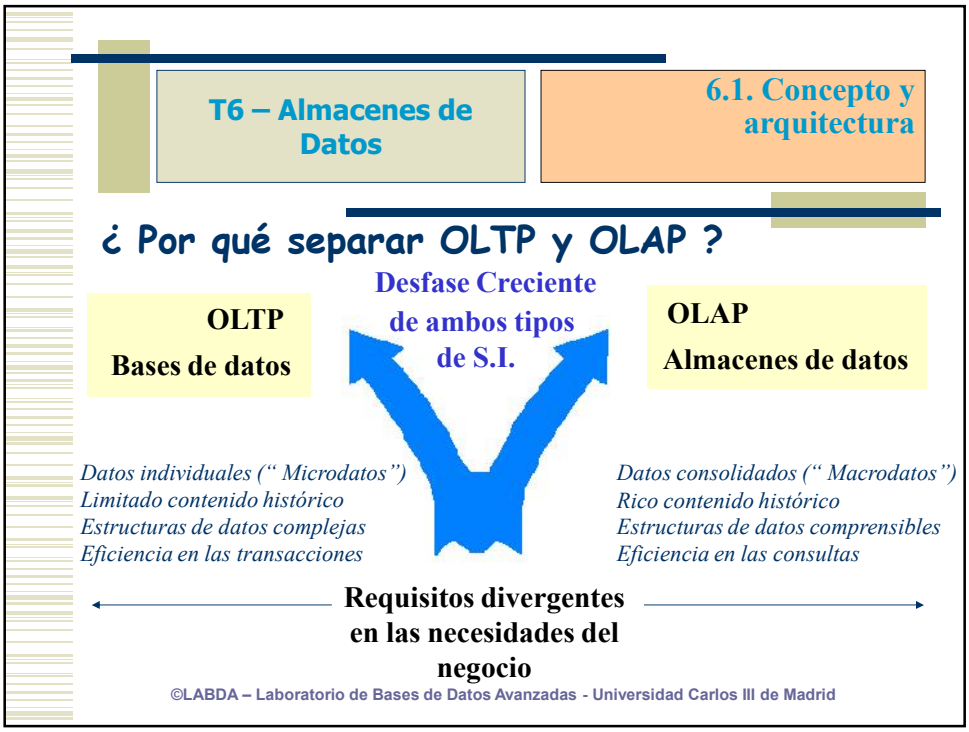
Para atender al:

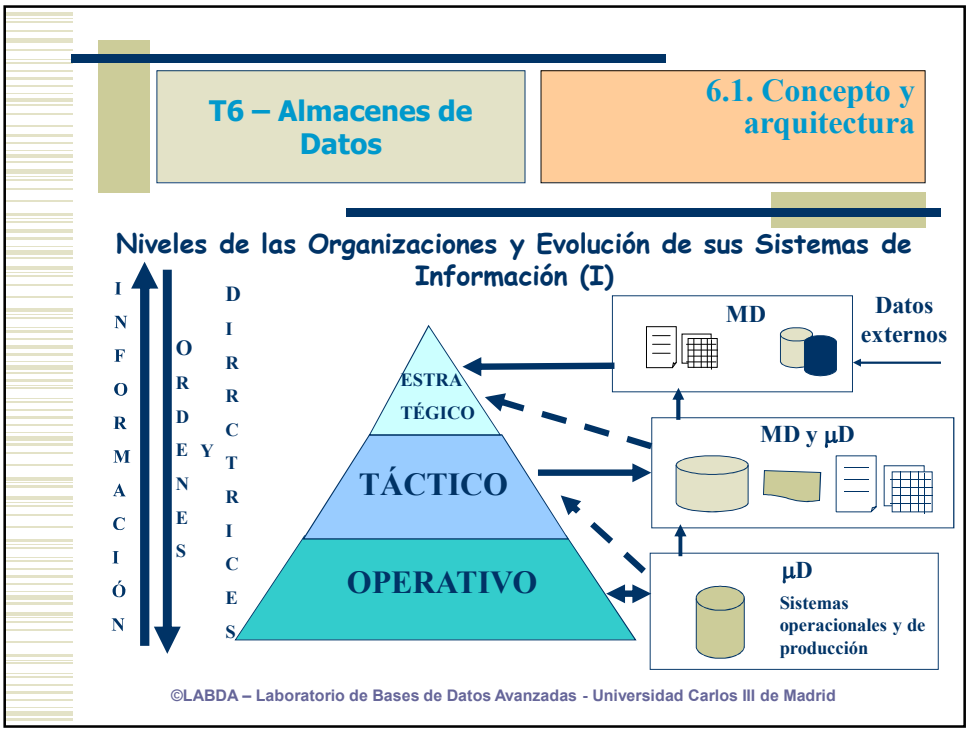
- Proceso transaccional

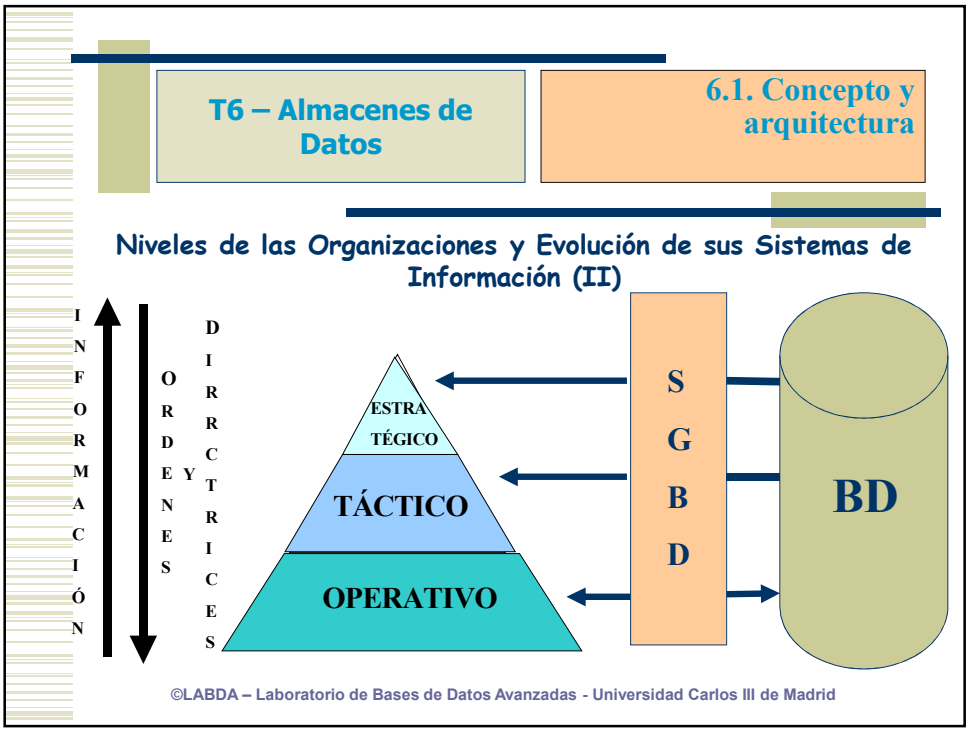
  → *Surgen las Bases de Datos*

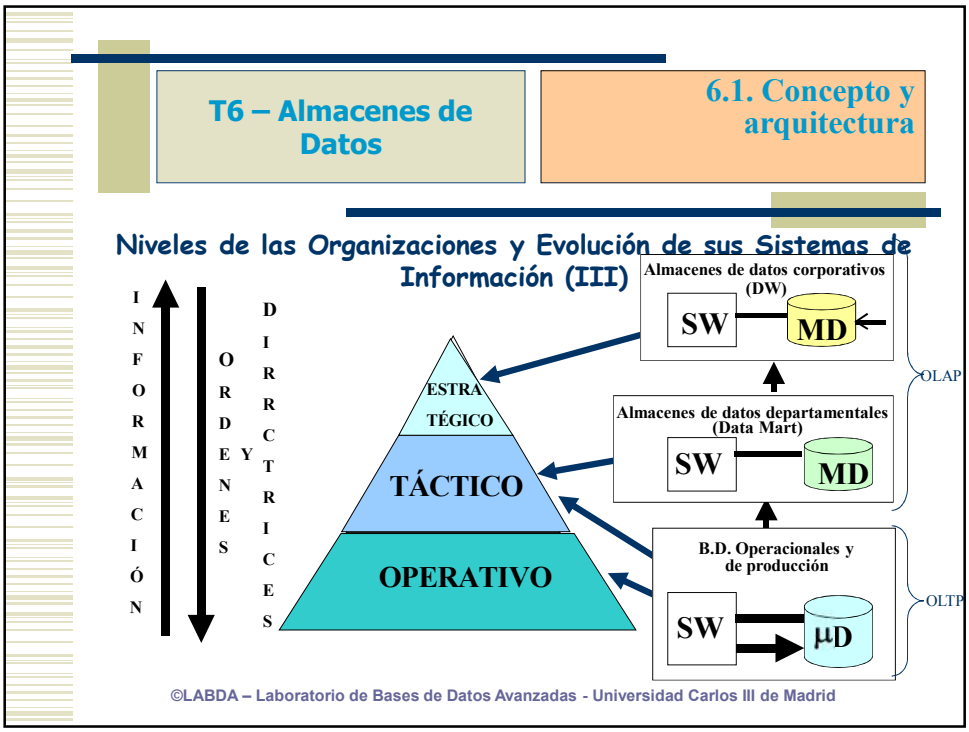
- Proceso analítico

  → *Surgen los Almacenes de Datos*


T6 – Almacenes de Datos
6.1. Concepto y arquitectura
¿ Por qué separar OLTP y OLAP ?
Desfase Creciente de ambos tipos de S.I.
OLTP
Bases de datos
OLAP
Almacenes de datos
Datos individuales (" Microdatos")
Limitado contenido histórico
Estructuras de datos complejas
Eficiencia en las transacciones
Datos consolidados (" Macrodatos")
Rico contenido histórico
Estructuras de datos comprensibles
Eficiencia en las consultas
Requisitos divergentes en las necesidades del negocio
©LABDA – Laboratorio de Bases de Datos Avanzadas - Universidad Carlos III de Madrid


T6 – Almacenes de Datos
6.1. Concepto y arquitectura
Niveles de las Organizaciones y Evolución de sus Sistemas de Información (I)
INFORMACIÓN
ORDENES Y DIRECTRICES
ESTRATÉGICO
TÁCTICO
OPERATIVO
MD
Datos externos
MD y µD
µD
Sistemas operacionales y de producción
©LABDA – Laboratorio de Bases de Datos Avanzadas - Universidad Carlos III de Madrid


T6 – Almacenes de Datos
6.1. Concepto y arquitectura
Niveles de las Organizaciones y Evolución de sus Sistemas de Información (II)
INFORMACIÓN
ÓRDENES Y DIRECTRICES
ESTRATÉGICO
TÁCTICO
OPERATIVO
SGBD
BD
©LABDA – Laboratorio de Bases de Datos Avanzadas - Universidad Carlos III de Madrid


T6 – Almacenes de Datos
6.1. Concepto y arquitectura
Niveles de las Organizaciones y Evolución de sus Sistemas de Información (III)
INFORMACIÓN
ÓRDENES Y DIRECTRICES
ESTRATÉGICO
TÁCTICO
OPERATIVO
Almacenes de datos corporativos (DW)
SW
MD
Almacenes de datos departamentales (Data Mart)
SW
MD
B.D. Operacionales y de producción
SW
μD
OLAP
OLTP
©LABDA – Laboratorio de Bases de Datos Avanzadas - Universidad Carlos III de Madrid

T6 – Almacenes de Datos

6.1. Concepto y arquitectura

Razones para mantener separados los datos OLAP de los datos OLTP

- Actualizaciones masivas y poco frecuentes
  - → *Por eficiencia, los datos del AD se pueden desnormalizar*
- Consultas muy largas
- Elevado consumo de recursos
  - → *Pueden interferir en el proceso transaccional*
- Fuentes de datos muy variadas
  - → *No solo transaccionales*

T6 – Almacenes de Datos

6.1. Concepto y arquitectura

OLTP vs. OLAP

	OLTP	OLAP
Usuario típico	Administrativo	Estratega
Usuario del sistema	Ejecución del negocio	Análisis del negocio
Interacción de usuario	Predeterminado	Ad-hoc
Unidad de trabajo	Transacción	Consulta
Característica de trabajo	Lectura / Escritura	Principalmente lectura
Registros accedidos	Cientos	Millones
Número de usuarios	Miles	Cientos
Focos	Entrada de datos	Salida de información

T6 – Almacenes de Datos

6.1. Concepto y arquitectura

Tipos de almacenes de datos

- **Corporativos** (*Data Warehouses*)
- **Departamentales** (*Data Marts*)

T6 – Almacenes de Datos

6.1. Concepto y arquitectura

"El almacenamiento de datos (*data warehousing*) y el procesamiento analítico en línea (*on - line analytical processing)* son elementos esenciales en el soporte de decisiones, que se están convirtiendo de forma creciente en un foco de la industria de las bases de datos"

Chandhuri, S.
Dayal, U.

"An overview of Data Warehousing and OLAP Technology"
ACM Sigmod Record, 26 (1) March, 1997

T6 – Almacenes de Datos

6.1. Concepto y arquitectura

TERMINOLOGIA

SISTEMAS DE GESTION DEL ALMACEN DE DATOS es un software de gestión para los almacenes de datos

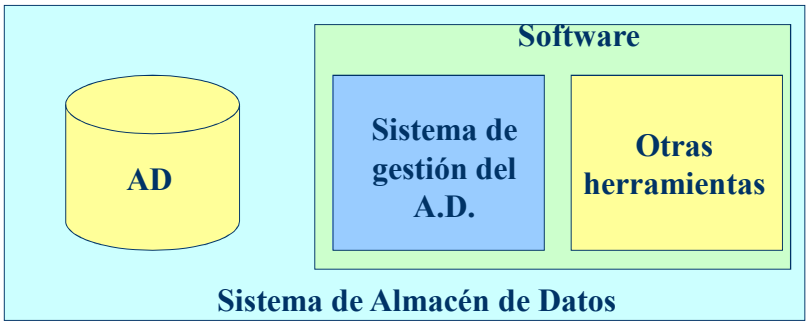
ALMACENES DE DATOS son **un tipo de bases de datos** que contienen datos seleccionados y organizados para satisfacer las necesidades del proceso OLAP

SISTEMAS DE ALMACENES DE DATOS son herramientas que permiten proporcionar soporte a la tecnología OLAP

T6 – Almacenes de Datos

6.1. Concepto y arquitectura

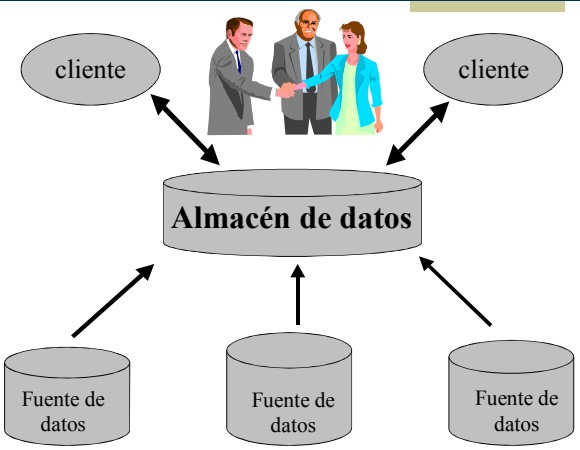
Componentes de un sistema de almacén de datos



T6 – Almacenes de Datos

6.1. Concepto y arquitectura

- **Múltiples clientes**
  analistas de negocios
  ejecutivos
  empresarios
  ....
- **Repositorio de datos**
  (Almacenamiento)
- **Múltiples fuentes**
  transaccionales
  históricas
  externas
  ....



T6 – Almacenes de Datos

6.1. Concepto y arquitectura

- Consumidores de Información
  - Expertos en negocio
  - Con capacidad de decisión
  - Cambian el enfoque de búsqueda dinámicamente; *"Dáme lo que te pido, que luego podré decirte lo que realmente quiero"*
  - Ajenos al OLTP

T6 – Almacenes de Datos

6.1. Concepto y arquitectura

¿Qué requiere el usuario OLAP?

- Conceptos familiares para el usuario final
  - Dimensiones, medidas y Jerarquías
- Acceso inmediato a los datos
- Información consistente
- Navegación y consulta sencillas
- Capacidades de generación de informes
- Datos precalculados
- Soporte de grandes volúmenes de datos

T6 – Almacenes de Datos

6.1. Concepto y arquitectura

¿Qué requiere el usuario OLAP?

- Flexibilidad de manejo y presentación
- Potentes capacidades de Análisis
  - Agregaciones, Comparaciones, Ratios, Rankings, Correlaciones, ...
  - Análisis de situaciones y escenarios "what-if", Contraste de Hipótesis, ...
  - Descubrimiento de patrones y tendencias, Previsiones, Series Temporales ...

T6 – Almacenes de Datos

6.1. Concepto y arquitectura

OLAP

Se trata de un término inventado para describir una aproximación dimensional interactiva al soporte de toma de decisiones (análisis desde la perspectiva de sus componentes o dimensiones, contemplando también los distintos niveles o jerarquías que éstas poseen).

Proceso, Dimensiones, Interacción, Análisis, Toma de decisiones

Este tipo de Análisis, se soporta mediante una visión dimensional:

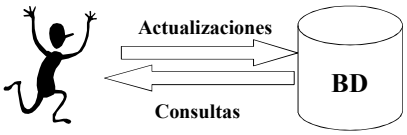
Modelo Multidimensional o Cubo

T6 – Almacenes de Datos

6.1. Concepto y arquitectura

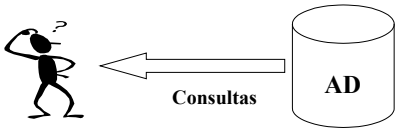
Sistemas OLTP :

Volátiles



Sistemas OLAP :

Poco volátiles



T6 – Almacenes de Datos

6.1. Concepto y arquitectura

¿ Qué es un almacén de datos (I) ?

- **Colección de diversos datos almacenados en soporte secundario**
  - Una solución para los problemas de integración de datos
  - Simple repositorio de información
- **Orientada a ciertos aspectos del negocio**
  - Organizada por materias, no por aplicaciones
  - Usada para análisis, explotación de datos...

T6 – Almacenes de Datos

6.1. Concepto y arquitectura

¿ Qué es un almacén de datos ? (II)

- **Optimizado de modo diferente al de las bases de datos orientadas a las operaciones**
- **Interfaz de usuario enfocada al ejecutivo**

T6 – Almacenes de Datos

6.1. Concepto y arquitectura

¿ Qué es un almacén de datos? (III)

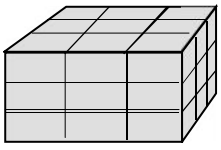
- **Gran volumen de datos (Gb, Tb)**
- **No volátil**
  - Histórico
  - Los atributos de tiempo son importantes
- **Actualizaciónes poco frecuentes, (**se añade información en procesos, en general, *" por lotes"* **)**

T6 – Almacenes de Datos

6.2. Modelo Multidimensional. Diseño

Modelo de Datos Multimensional (MDM)

Modelo de Datos (Estática y Dinámica) basado en estructuras multidimensionales



T6 – Almacenes de Datos

6.2. Modelo Multidimensional. Diseño

Bases de datos Multidimensionales - *BDM*

Base de Datos diseñada para *los sistemas de soporte de decisiones* en la cual los datos tienen una estructura matricial (multidimensional) para su almacenamiento. Este tipo de organización admite consultas más complejas.

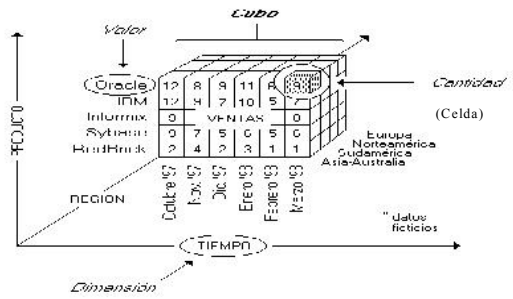
"Piense en una estructura de datos multimensional como de un cubo de Rubik, con datos que los usuarios pueden torcer y voltear de diferentes maneras, para trabajar con escenarios ¿qué sucedería si? y ¿qué ha sucedido?"

Lee Thé, (1995)
Editor de Datamation

T6 – Almacenes de Datos

6.2. Modelo Multidimensional. Diseño

Ejemplo: Estructura multidimensional para el almacenamiento de datos multidimensionales



T6 – Almacenes de Datos

6.2. Modelo Multidimensional. Diseño

MDM: Metáfora del Cubo

Dimensiones

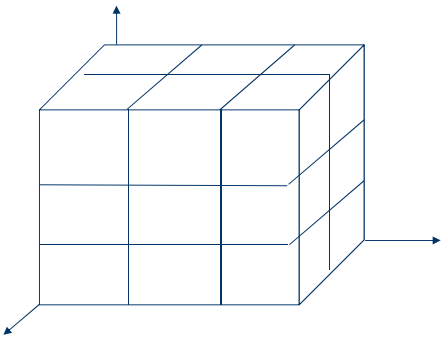
Atributos de dimensión

Atributos de hecho

Funciones resumen

Cubo

Celda



T6 – Almacenes de Datos

6.2. Modelo Multidimensional. Diseño

Estática: Elementos de un MDM (I)

- **Esquema de hecho** (esquema de cubo): es el objeto a analizar
  Ejemplos: empleados, ventas, stocks...
- **Atributos de hecho o de síntesis, medidas** (*measures*): atributos de tipo cuantitativo cuyos valores (cantidades) se obtienen generalmente por aplicación de una función estadística que resume un conjunto de valores en un único valor

  Ejemplos: nº de empleados, cantidad vendida, precio medio,...
- **Funciones resumen:** funciones de tipo estadístico que se aplican a los atributos de hecho

  Ejemplos: frecuencia, suma, media, máximo, etc

T6 – Almacenes de Datos

6.2. Modelo Multidimensional. Diseño

Estática: Elementos de un MDM (II)

- **Dimensiones**: cada uno de los ejes en un espacio multidimensional

  Ejemplos: tiempo, espacio, productos, intervalos del nº de empleados, departamentos

- **Atributos de Dimensión o de Clasificación**: atributos de tipo cualitativo (sus valores son modalidades) que suministran el contexto en el que se obtienen las medidas en un esquema de hecho

  Ejemplos: días, semanas, ciudades, provincias...

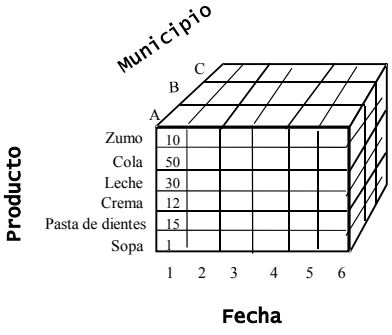
- **Jerarquías**: varios atributos de dimensión unidos mediante una relación de tipo jerárquico

  Ejemplos: día -> semana -> mes -> año

T6 – Almacenes de Datos

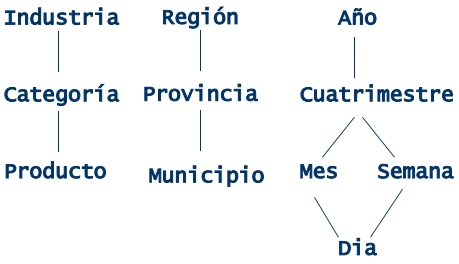
6.2. Modelo Multidimensional. Diseño

Ejemplo:



Dimensiones: Producto, Ciudad, Fecha

Jerarquías



T6 – Almacenes de Datos

6.2. Modelo Multidimensional. Diseño

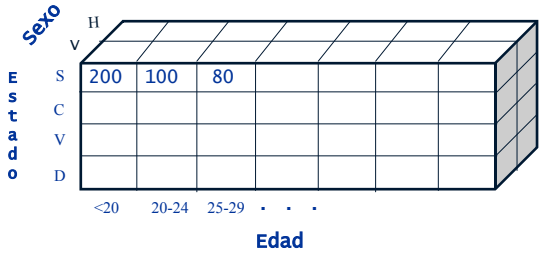
Ejemplo: Empleados por edad, sexo y estado civil

SEXO	EC \ EDAD	<20	20 - 24	25 - 29	. . . . .	TOTAL
V	S	200	100	80	. . . . .	
	C	70	150	210	. . . . .	
	V	2	10	25	. . . . .	
	D	4	30	55	. . . . .	
M	S	210	105	60	. . . . .	
	C	80	100	230	. . . . .	
	V	2	5	35	. . . . .	
	D	4	33	40	. . . . .	
TOTAL		572	533	735	. . . . .	

T6 – Almacenes de Datos

6.2. Modelo Multidimensional. Diseño

Ejemplo: Empleados por edad, sexo y estado civil



T6 – Almacenes de Datos

6.2. Modelo Multidimensional. Diseño

Metodología de Kimball (1996)

1. Entrevistas con usuarios finales y administradores de las bases de datos:
   - fijar las necesidades de consulta teniendo en cuenta los datos disponibles procedentes de los sistemas heredados
   - información sobre los datos disponibles para poblar el almacén
2. Identificación de las tablas de hechos
3. La granularidad de cada tabla de hechos
4. Las dimensiones de cada tabla de hechos
5. Los hechos, incluyendo los precalculados
6. Los atributos de las dimensiones con una dexripción completa y una terminología propia
7. Cómo hacer el seguimiento de las dimensiones variantes en el tiempo
8. Las agregaciones, modos de consulta y otras decisiones de almacenamiento físico
9. La duración histórica del almacen de datos
10. La urgencia con la que el dato es extraído y cargado en el almacén.

T6 – Almacenes de Datos

6.2. Modelo Multidimensional. Diseño

Paso 1: Selección del proceso
Tema objetivo de un mercado de datos concreto. Proceso de negocio.

Paso 2: Selección de la granularidad
Decidir qué es lo que va a representar cada registro de la tabla de hechos. Esta granularidad también nos dará las dimensiones y la granularidad de las mismas.

Paso 3: Identificación y construcción de las dimensiones
Las dimensiones establecen el contexto para realizar preguntas acerca de los hechos contenidos en la tabla. Un buen conjunto de dimensiones hace que los datos sean comprensibles y fáciles de utilizar.

Paso 4: Selección de los hechos
Vendrá determinado por la granularidad y las dimensiones elegidas

T6 – Almacenes de Datos

6.2. Modelo Multidimensional. Diseño

ROLAP: Tipos de Diseño

- Esquema en estrella
- Esquema en copo de nieve
- Constelación de estrellas

T6 – Almacenes de Datos

6.2. Modelo Multidimensional. Diseño

ROLAP: Tipos de Diseño

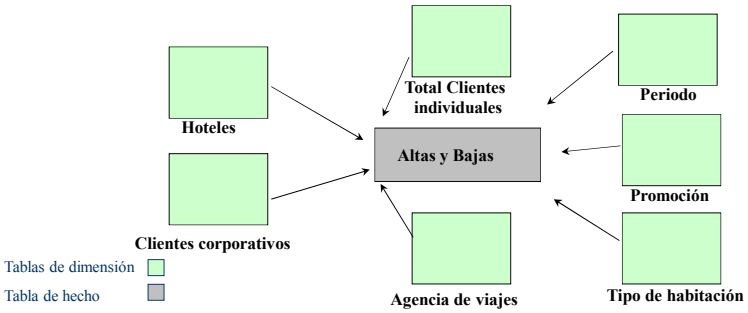
- **Esquema en estrella:** Esquema relacional adaptado a la representación de datos multidimensionales. Se basa en una serie de tablas que representan dimensiones unidas mediante claves ajenas, a una principal que actúa como nexo y almacena datos agregados y precalculados
  `(Tablas no normalizadas)`
- **Esquema en copo de nieve:** Variante del esquema de estrella que presenta las tablas de dimensión estructuradas a más de un nivel
  `(Tablas normalizadas)`

T6 – Almacenes de Datos

6.2. Modelo Multidimensional. Diseño

Esquema en estrella

Esquema relacional adaptado a la representación de datos multidimensionales. Se basa en una serie de tablas que representan dimensiones unidas mediante claves ajenas, a una principal que actúa como nexo (Tablas no normalizadas)



T6 – Almacenes de Datos

6.2. Modelo Multidimensional. Diseño

Tabla de Hecho

Contenido

- **Clave:** Concatenación de las claves de todas las tablas de dimensión asociadas al EH (claves ajenas que referencian a las claves de las correspondientes dimensiones)
- **Atributos de Hecho:** Atributos, en general de tipo numérico, que contienen:
  - Datos resumen, muchas veces de tipo aditivo **SUMA** (Euros, Cantidades, ...)
  - Indicación de inexistencia de valor (nulo)
  - Resultado de aplicar un método

Características

- Filas con pocas columnas (pocos atributos)
- Nº filas: desde millones a más de miles de millones (tantas como celdas tenga el cubo)
- Acceso, en general, vía dimensiones

T6 – Almacenes de Datos

6.2. Modelo Multidimensional. Diseño

Tablas de dimensión

- Definen las dimensiones de negocio en *términos familiares* para los usuarios
- Filas con numerosas columnas de texto, altamente descriptivas
- Normalmente menos de un millón de filas
- Combinadas con las tablas de hecho mediante claves ajenas
- Altamente indexadas
- A veces se desnormaliza (estructura "**copo de nieve**")

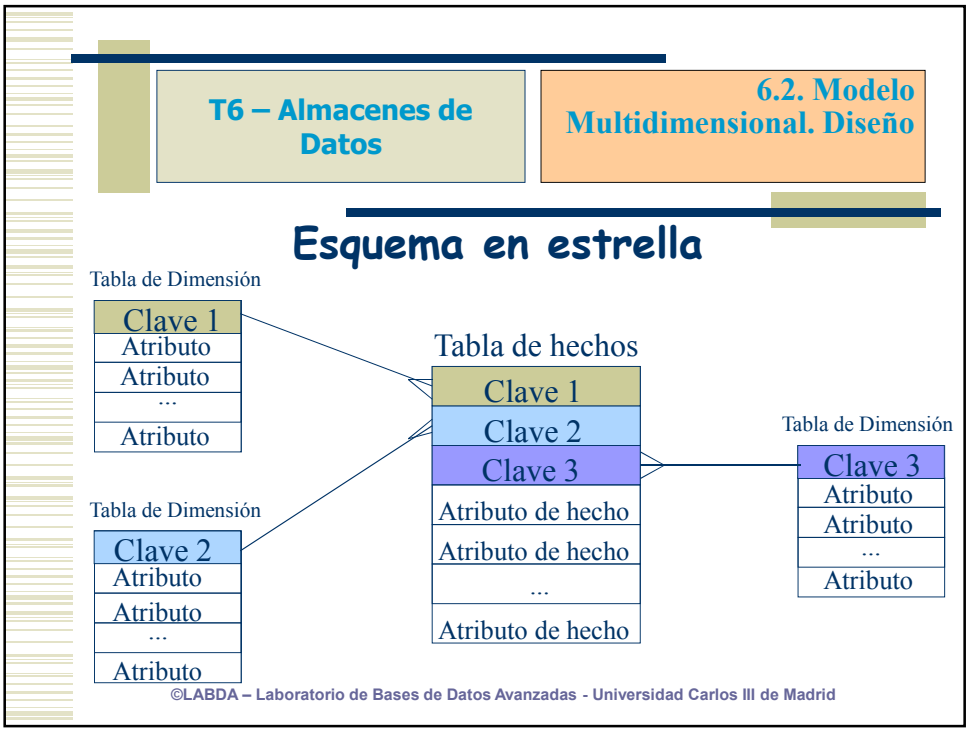
T6 – Almacenes de Datos

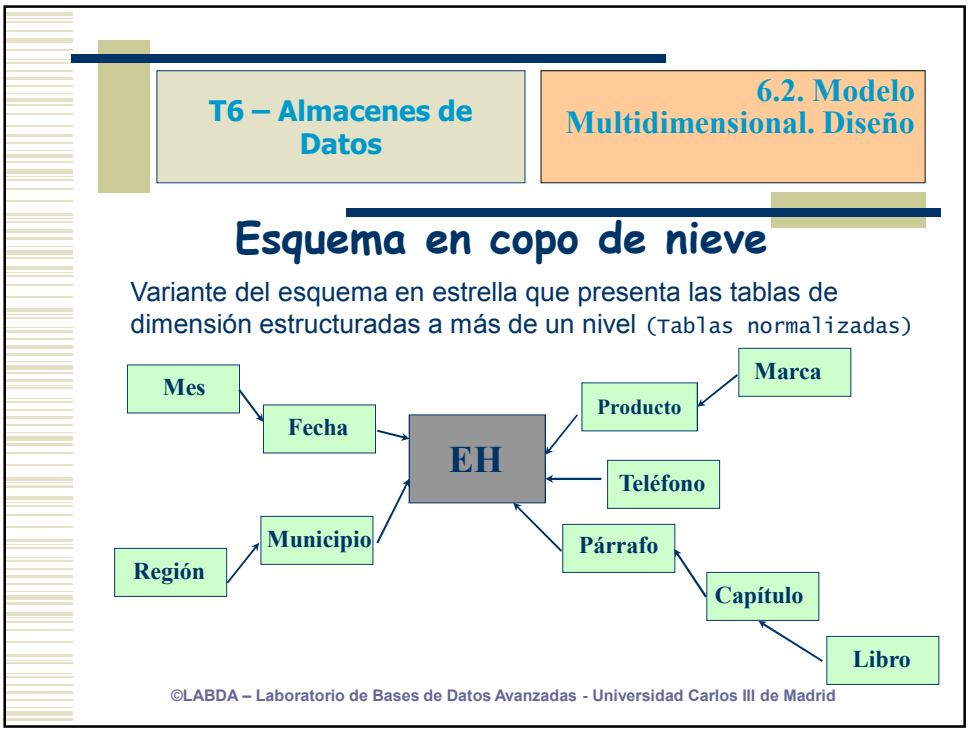
6.2. Modelo Multidimensional. Diseño

Dimensiones típicas

- Periodos de tiempo
- Región geográfica
- Productos
- Promociones
- Clientes
- Ventas representativas (comprador)
- Acceso (frecuente, permanente)
- Tipo de habitación (doble, individual, suite, ...)
- Medicinas (gratuita, hospitalaria, particular, ...)
- Vendedor (distribuidor, almacén)

. . . . . .


T6 – Almacenes de Datos
6.2. Modelo Multidimensional. Diseño
Esquema en estrella
Tabla de Dimensión
Clave 1
Atributo
Atributo
...
Atributo
Tabla de Dimensión
Clave 2
Atributo
Atributo
...
Atributo
Tabla de hechos
Clave 1
Clave 2
Clave 3
Atributo de hecho
Atributo de hecho
...
Atributo de hecho
Tabla de Dimensión
Clave 3
Atributo
Atributo
...
Atributo
©LABDA – Laboratorio de Bases de Datos Avanzadas - Universidad Carlos III de Madrid


T6 – Almacenes de Datos
6.2. Modelo Multidimensional. Diseño
Esquema en copo de nieve
Variante del esquema en estrella que presenta las tablas de dimensión estructuradas a más de un nivel (Tablas normalizadas)
Mes
Fecha
Región
Municipio
EH
Producto
Marca
Teléfono
Párrafo
Capítulo
Libro
©LABDA – Laboratorio de Bases de Datos Avanzadas - Universidad Carlos III de Madrid

T6 – Almacenes de Datos

6.2. Modelo Multidimensional. Diseño

Constelación de estrellas

Varios esquemas en estrella y / o en copo de nieve que comparten dimensiones