



# DATAWAREHOUSE

# DATAWAREHOUSE



- ▶ Surge de la necesidad por parte de la empresa de aprovechar los cada vez mas numerosos datos en línea para tomar mejores decisiones sobre sus actividades:
  - Artículos que deben tener en inventario.
  - Modo de dirigirse mejor a los cliente para aumentar sus ventas.

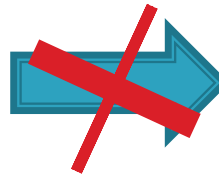
# DATAWAREHOUSE

- Una empresa automovilística puede darse cuenta que la mayor parte de los vehículos de pequeño tamaño los compran mujeres jóvenes cuyos ingresos anuales superan los 50.000\$



# DATAWAREHOUSE

- Esta empresa dirige su publicidad para atraer mas mujeres de estas características para que compren este tipo de vehículos.
- Así podría evitar desperdiciar dinero intentando atraer a otras categorías de consumidores para que compren esos vehículos



# DATAWAREHOUSE.

## Concepto

- ▶ Un Datawarehouse (DW) es una base de datos que almacena información para la toma de decisiones. Dicha información es construida a partir de bases de datos que registran las transacciones de los negocios de la organización



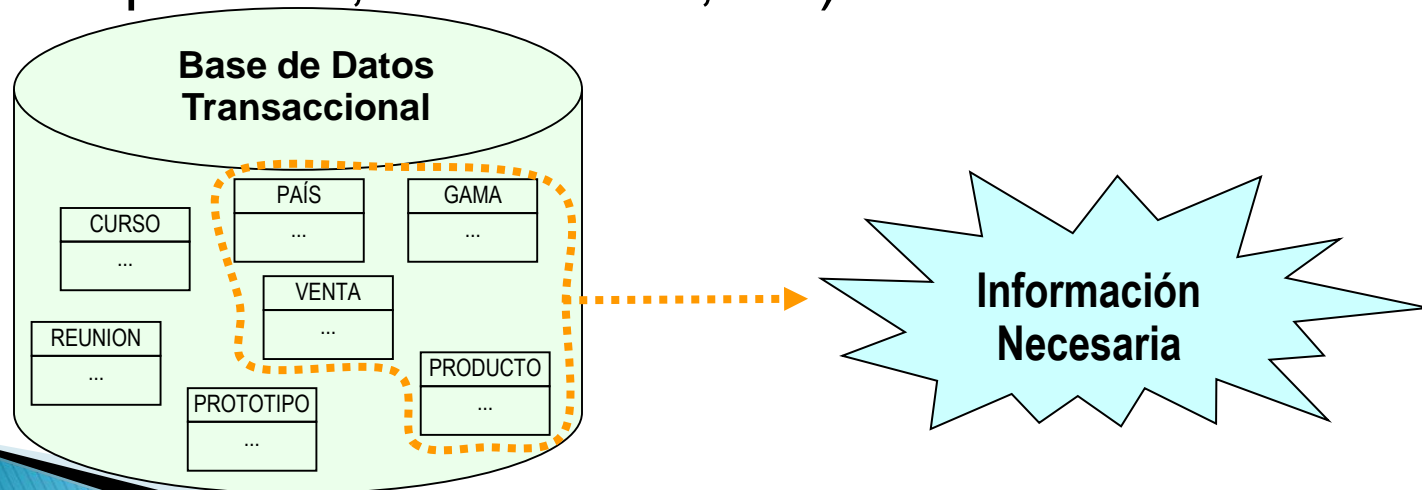
# DATAWAREHOUSE

## ▶ Características

- Recolección de datos que se encuentran orientados a sujetos o temas.
- Sus datos se integran en un único repositorio desde diversas fuentes de la empresa
- Sus datos almacenados son por lo general menos volátiles que en un sistema transaccional.
- Se usa para el soporte del proceso de toma de decisiones gerenciales.

# Características

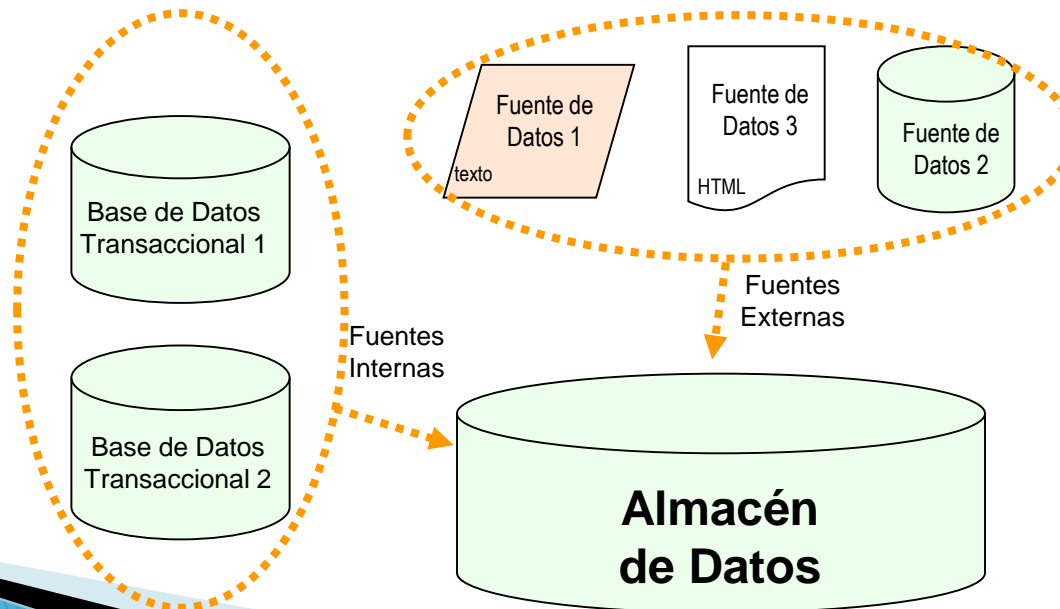
- ▶ Orientado hacia la información relevante de la organización
  - se diseña para consultar eficientemente información relativa a las actividades (ventas, compras, producción, ...) básicas de la organización, no para soportar los procesos que se realizan en ella (gestión de pedidos, facturación, etc).



# Características

## ► Integrado

- integra datos recogidos de diferentes sistemas operacionales de la organización (y/o fuentes externas).

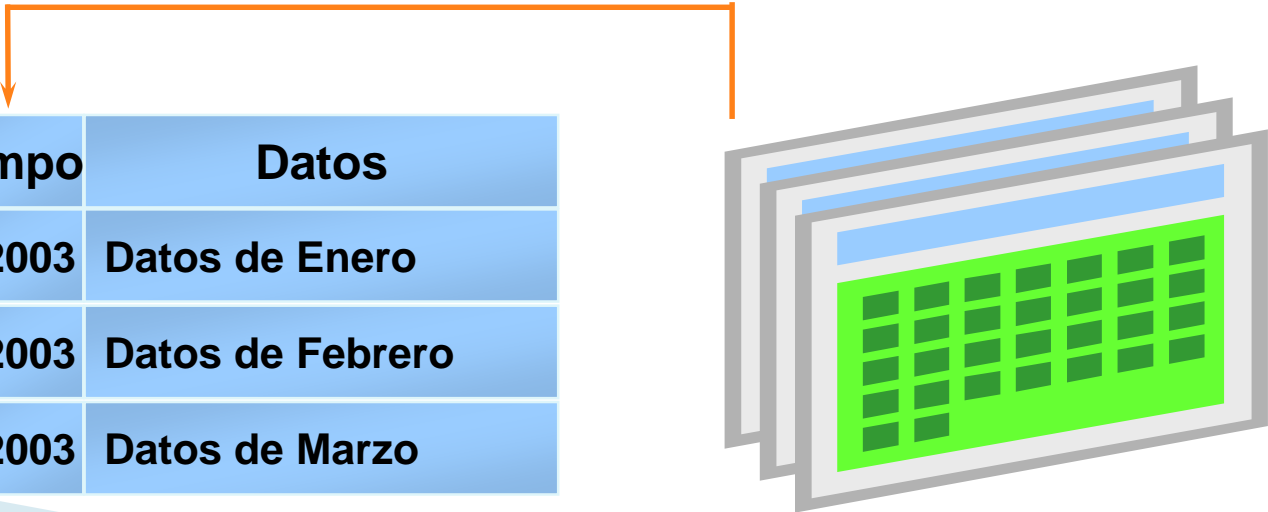




# Características

- ▶ 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.

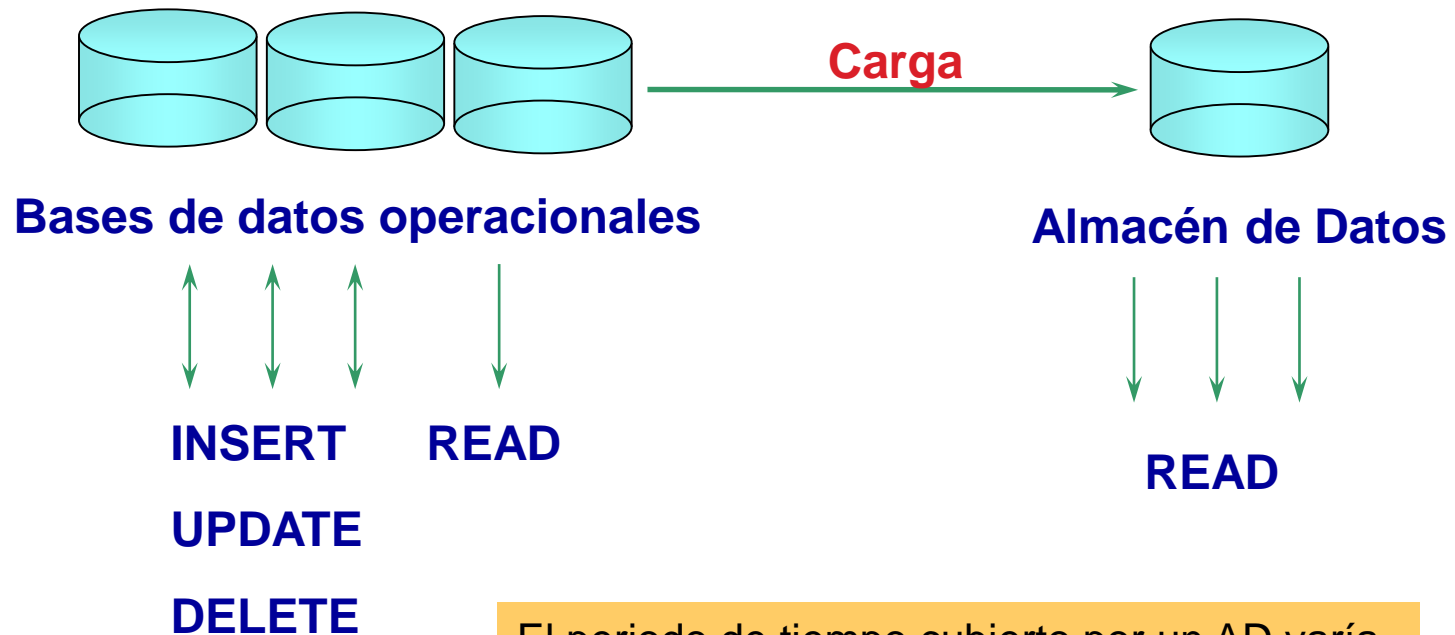


Tiempo	Datos
01/2003	Datos de Enero
02/2003	Datos de Febrero
03/2003	Datos de Marzo

# Características

## ► 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.

# Diferencias entre un datawarehouse y base de datos operacional

	Base de Datos Operacional	Datawarehouse
<b>Datos Del negocio</b>	Operacionales	Del negocio
<b>Uso de los datos</b>	Procesamiento repetitivo	Procesamiento analítico
<b>Orientación del diseño</b>	A la Aplicación (basada en Entidad Relación)	Al Tema o Sujeto (star schema, snowflake)
<b>Estructura de datos</b>	Muchas tablas altamente normalizadas	Pocas tablas con cierto grado de desnormalización
<b>Datos en el tiempo</b>	Actuales	Actuales + históricos
<b>Detalle de los datos</b>	Altamente detallados	Detallados + resúmenes
<b>Cambios en los datos</b>	Continuos	Datos más estables
<b>Cantidad de usuarios</b>	Más que en Datawarehouse	Menos que en la operacional.
<b>Tamaño de Base de Datos</b>	100 MB - GB	100 GB-TB
<b>Cantidad de registros accedidos en una operación</b>	Decenas	Millones

# Diferencias(2)

## Base de Datos Operativa

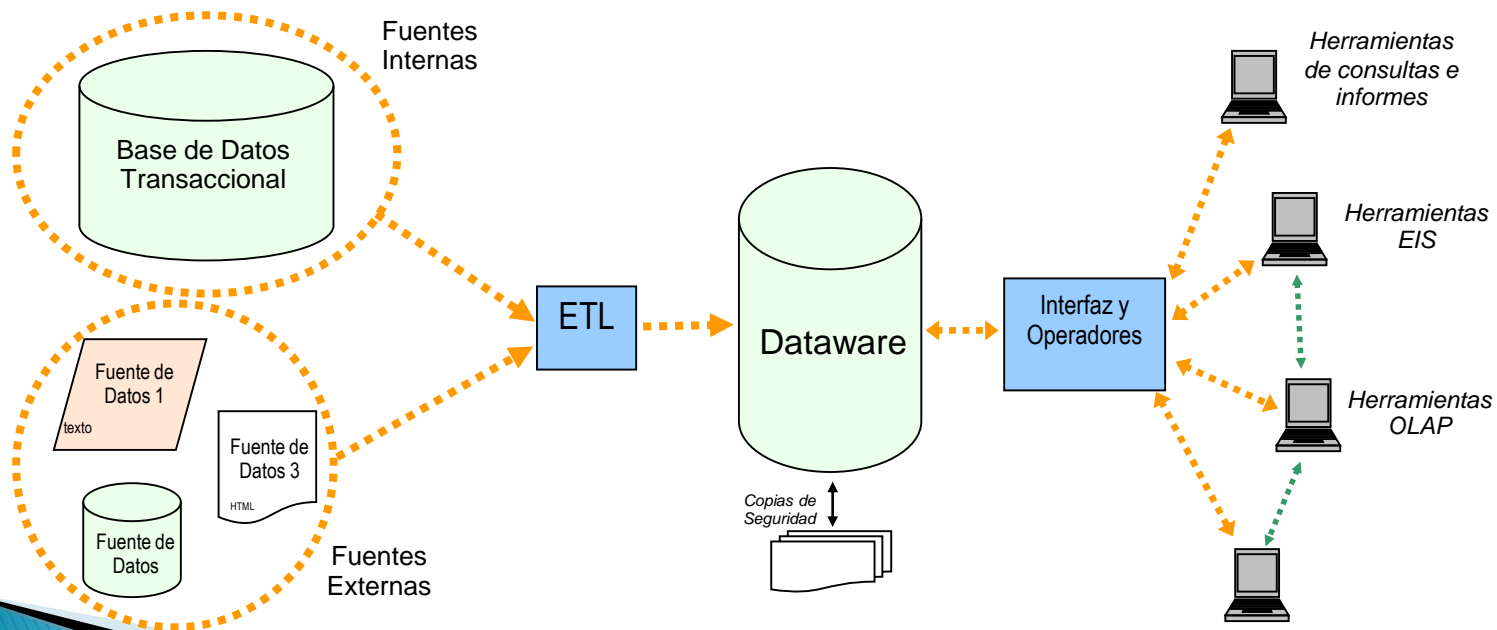
- Almacena la información de un sector del negocio.
- Se actualiza a medida que llegan datos que deban ser almacenados. Se opera mediante los cuatro mecanismos clásicos “añadir-eliminar-modificar-consulta”.
- Se orienta hacia la elaboración de informes periódicos.
- Suele manejar “pequeños” volúmenes de datos.
- Entorno optimizado para muchas transacciones (con gran cantidad de actualizaciones).
- Sirve de infraestructura al día a día de las funciones de explotación de una empresa.

## Un Datawarehouse:

- Almacena información integrada de los distintos sectores del negocio.
- Su actualización se realiza a intervalos regulares (típicamente una al día) dentro de un proceso controlado, y tras realizar un preprocesado de los datos que se van a almacenar.
- Su orientación es hacia la consulta del estado del negocio y obtención de información para ayuda en la toma de decisiones estratégicas.
- Se ofrece información bajo demanda (análisis mediante el uso de herramientas de generación de informes que consultan el datawarehouse).
- Refleja el modelo de negocio, frente al modelo de proceso.

# Arquitectura de un DW

- ▶ Viene determinada por su situación central como fuente de información para las herramientas de análisis.



# Arquitectura de un DW

## ► 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 DW, 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 y sobre ellos se conectan herramientas más sofisticadas (OLAP, minería de datos).
- Sistemas de Integridad y Seguridad: se encargan de un mantenimiento global, copias de seguridad,



# ETL

- ▶ La tarea más difícil y que más tiempo consume en la construcción de un DW es extraer, transformar y cargar los datos en el dataware.

# Procesamiento Analítico en Línea. OLAP(On-Line Analytical Processing )

- ▶ Se define como análisis rápido de información multidimensional compartida.
- ▶ Herramientas OLAP (para análisis de datos en DW): Frontales para el acceso a los datos del DW (o bases de datos multidimensionales también denominadas OLAP) basados en el modelo de datos multidimensional.

# 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).

# 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

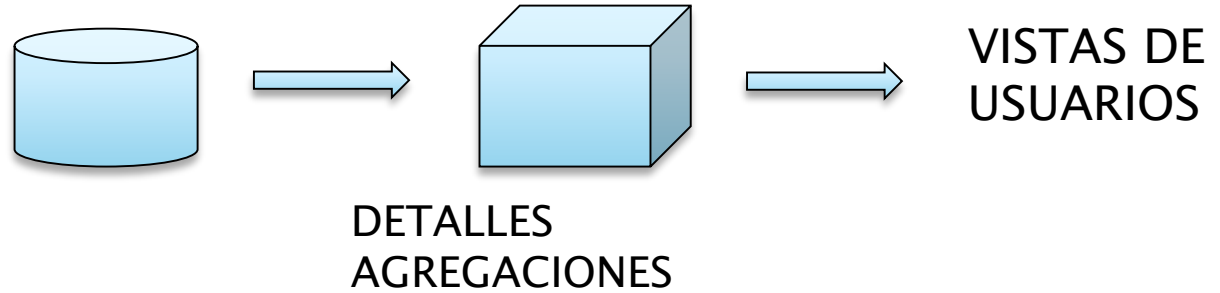
- ✓ disponen de estructuras de almacenamiento específicas (arrays) y técnicas de compactación de datos que favorecen el rendimiento del almacén.

## Sistemas HOLAP

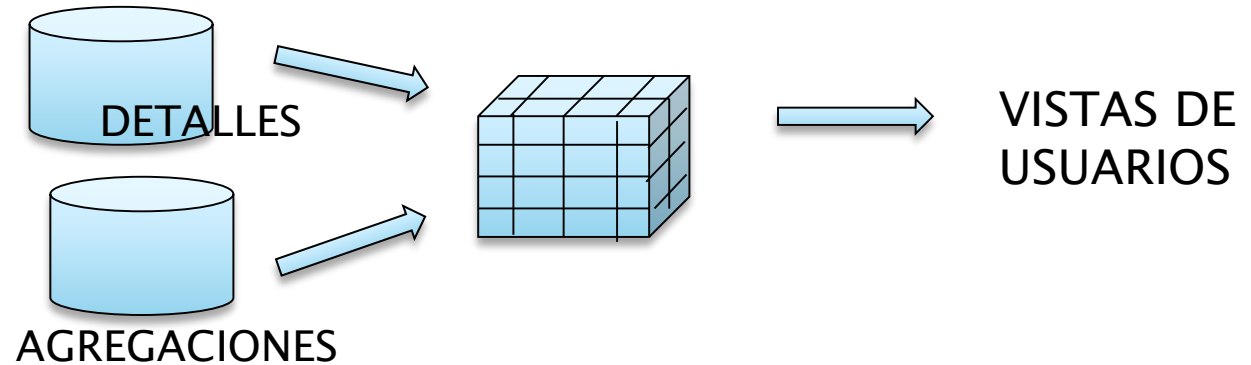
- ✓ sistemas híbridos entre ambos.

# ROLAP Y MOLAP

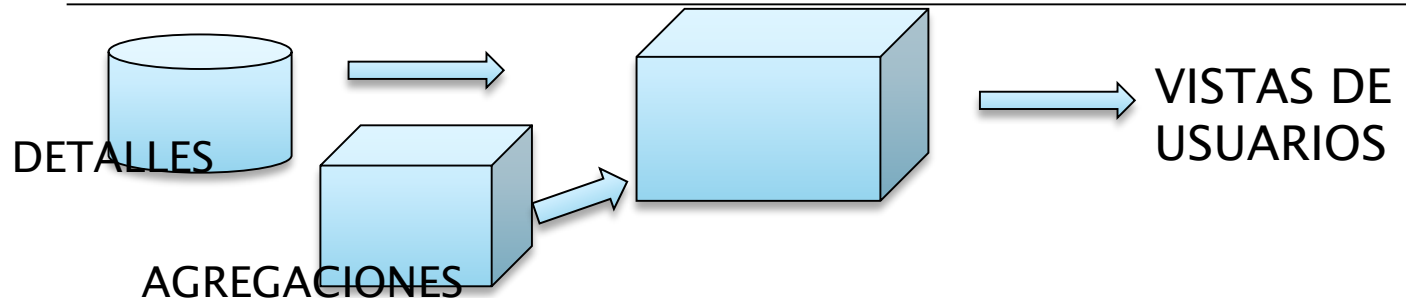
MOLAP



ROLAP



HOLAP



# Arquitectura de un DW

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

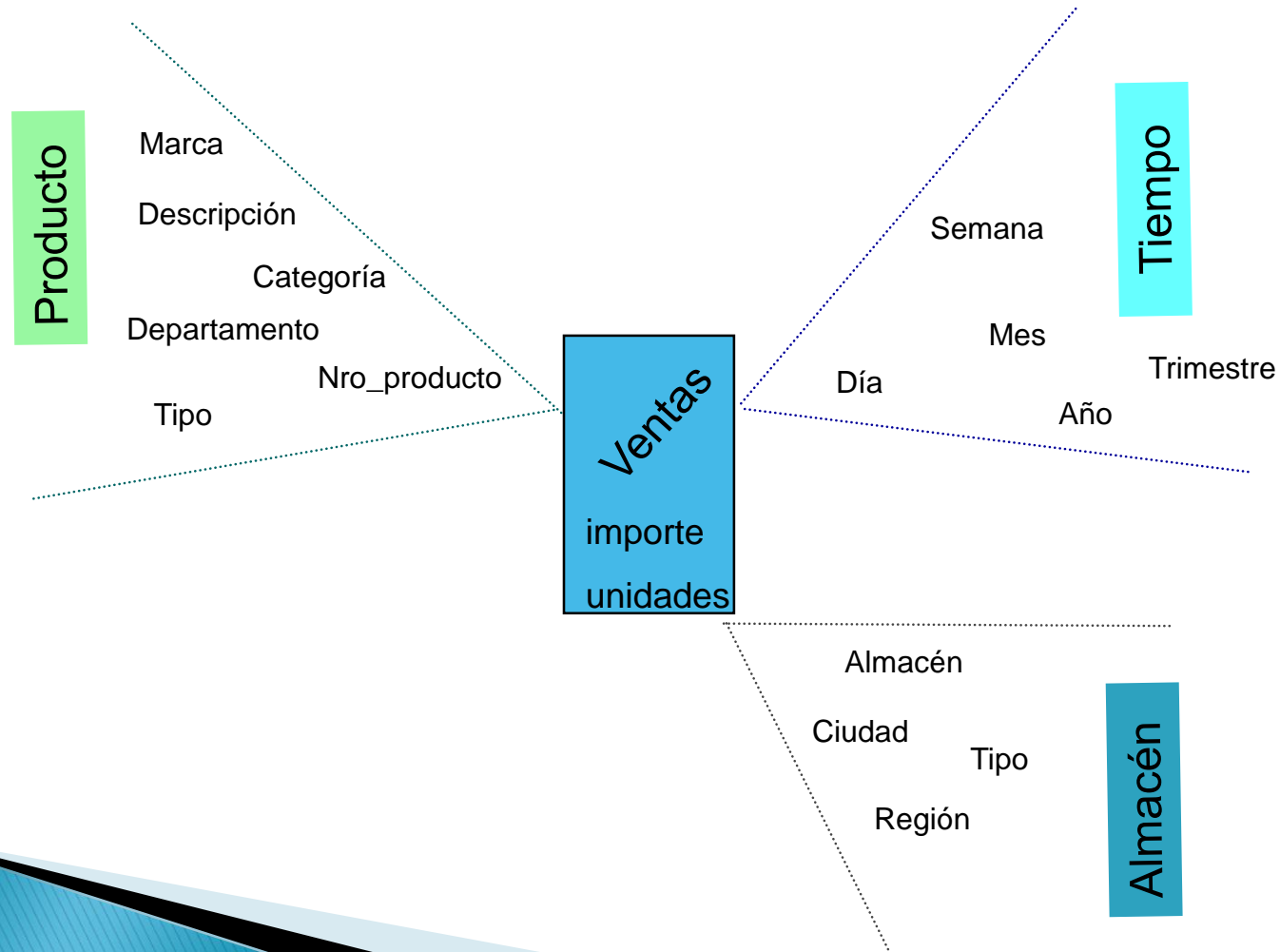
- EJEMPLO

- Organización: Cadena de supermercados.
- Actividad objeto de análisis: ventas de productos.
- Información registrada sobre una venta:
  - “del **producto** “Pharmaton 33cl” se han vendido en el **almacén** “Almacén nro.1” el **día** 17/7/2016, 2 **unidades** por un **importe** de 103.000 Gs.”

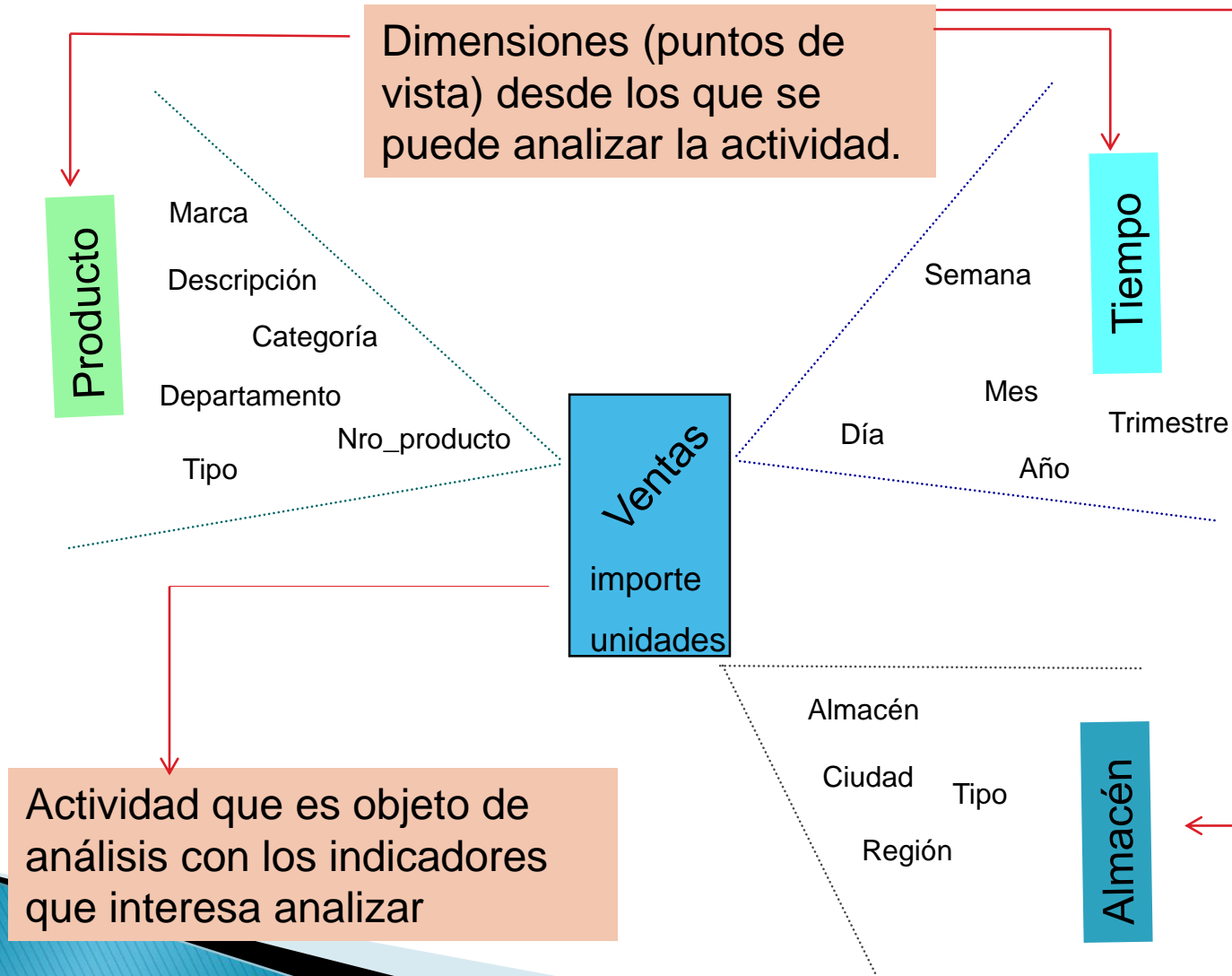
Para hacer el análisis no interesa la venta individual (ticket) realizada a un cliente sino las ventas diarias de productos en los distintos almacenes de la cadena.



# Arquitectura de un DW



# Arquitectura de un DW



# Arquitectura de un DW

## Modelo multidimensional:

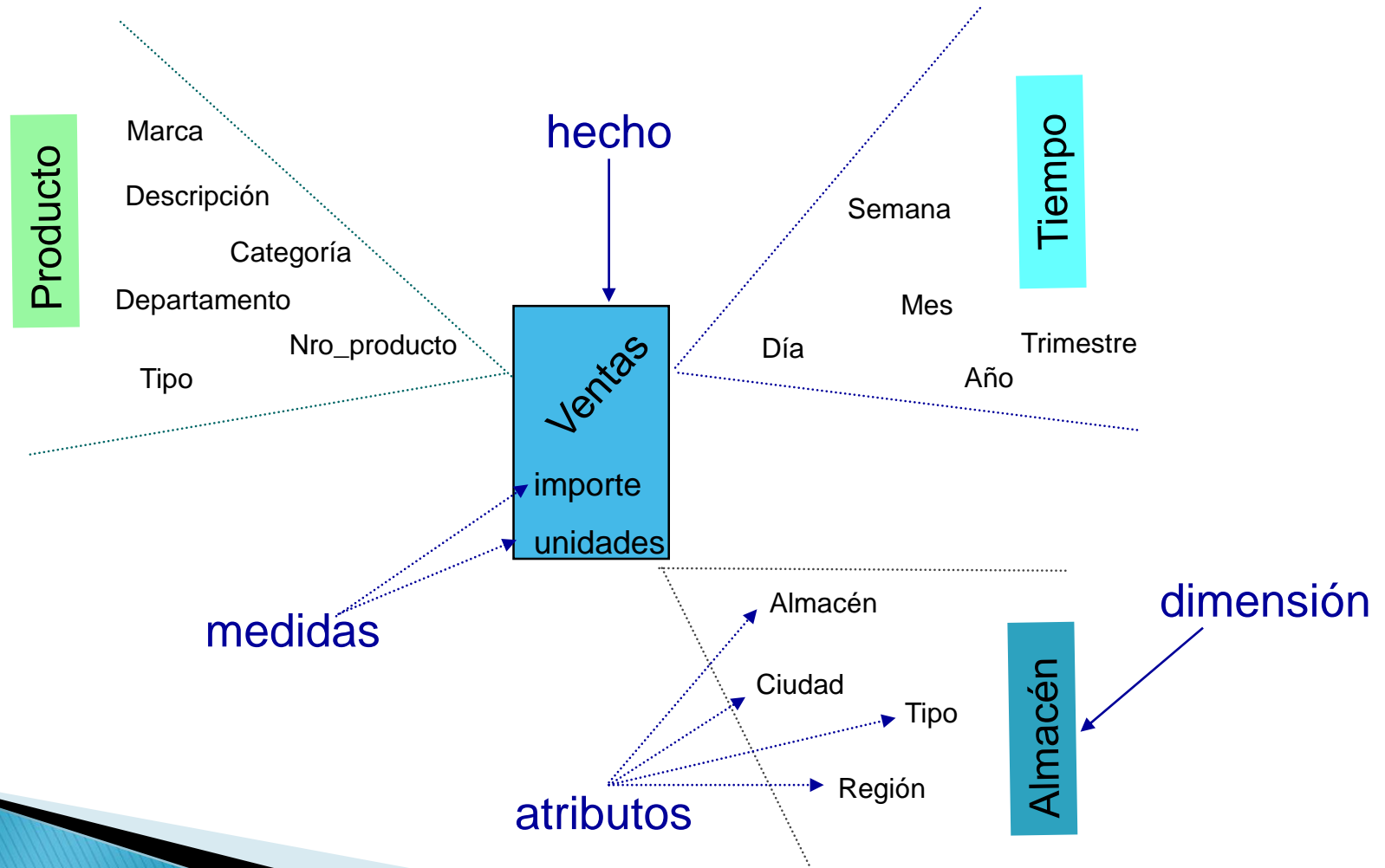
- ✓ en un esquema multidimensional se representa una actividad que es objeto de análisis (**hecho**) y las dimensiones que caracterizan la actividad (**dimensiones**).
- ✓ la información relevante sobre el **hecho** (actividad) se representa por un conjunto de indicadores (**medidas o atributos de hecho**).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

# Arquitectura de un DW

**Dimensiones:** Representan factores por lo que se analiza un determinado área del negocio. Son pequeñas y usualmente están desnormalizadas.

**Hechos:** Son el objeto de los análisis y están relacionados con las dimensiones. Son tablas muy grandes y suelen estar desnormalizadas. Se a menudo incluyen diferentes agregaciones como máximo, mínimo, media

# Arquitectura de un DW



# CUBOS OLAP

- ▶ Los cubos OLAP son representaciones específicas y segmentadas del Datawarehouse, en donde se realiza el cruce y conexión de los datos.
- ▶ Es una base de datos que posee diversas dimensiones, ampliando las posibilidades que hasta el momento ofrecían las conocidas hojas de cálculo.
- ▶ En otras palabras la forma de ver nuestro Datawarehouse es mediante los Cubos OLAP.



# CUBOS 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

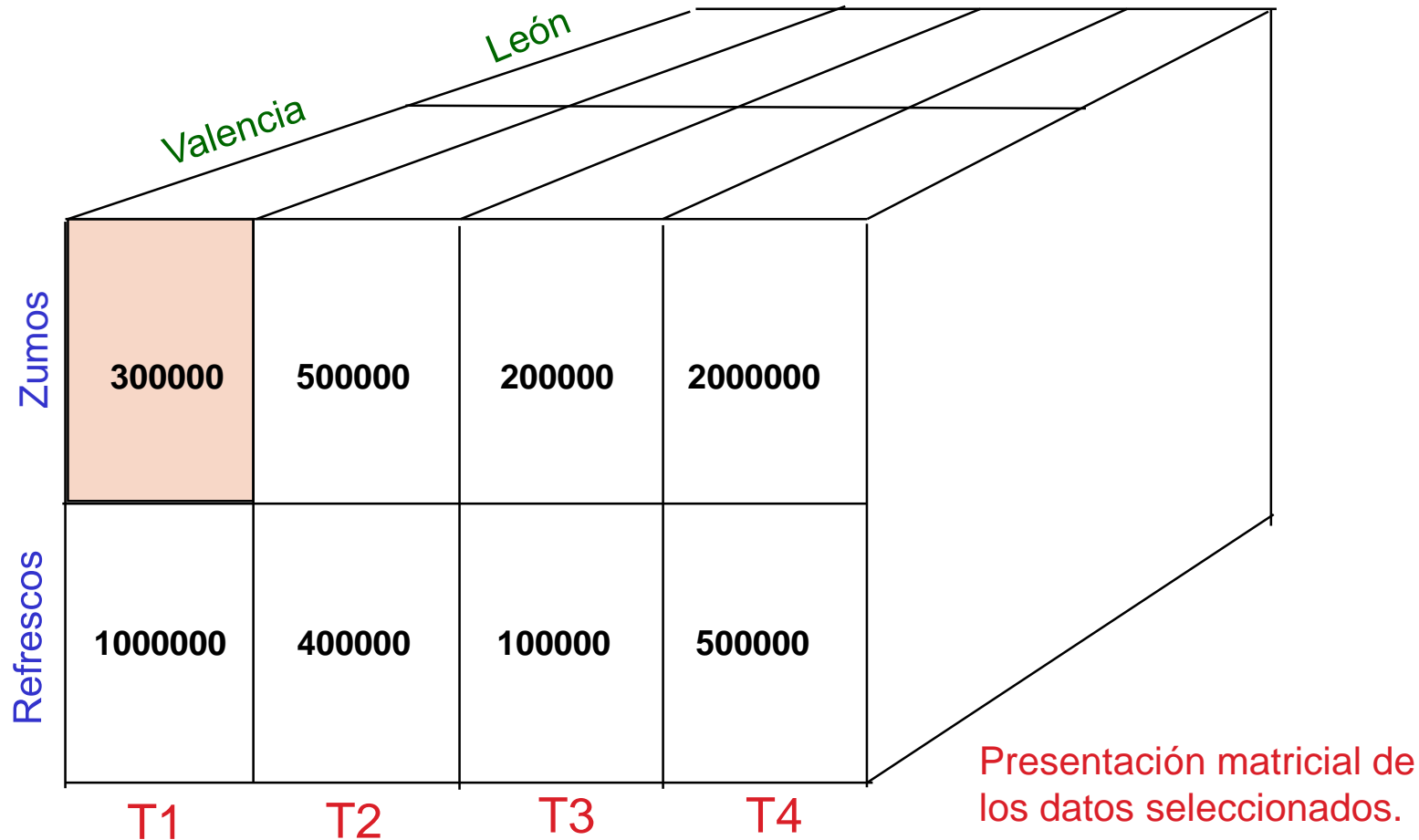
drill-across

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.

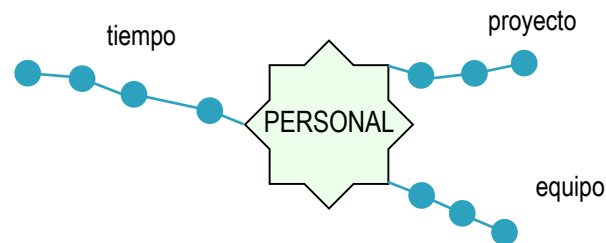
# CUBOS OLAP



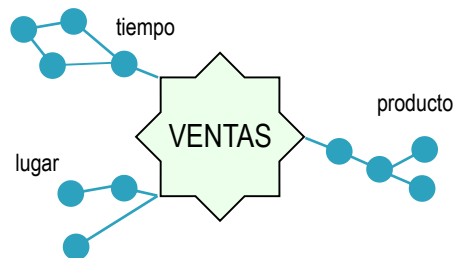
# Arquitectura de un DW

Este esquema multidimensional recibe varios nombres:

❖ **estrella**: si la jerarquía de dimensiones es lineal

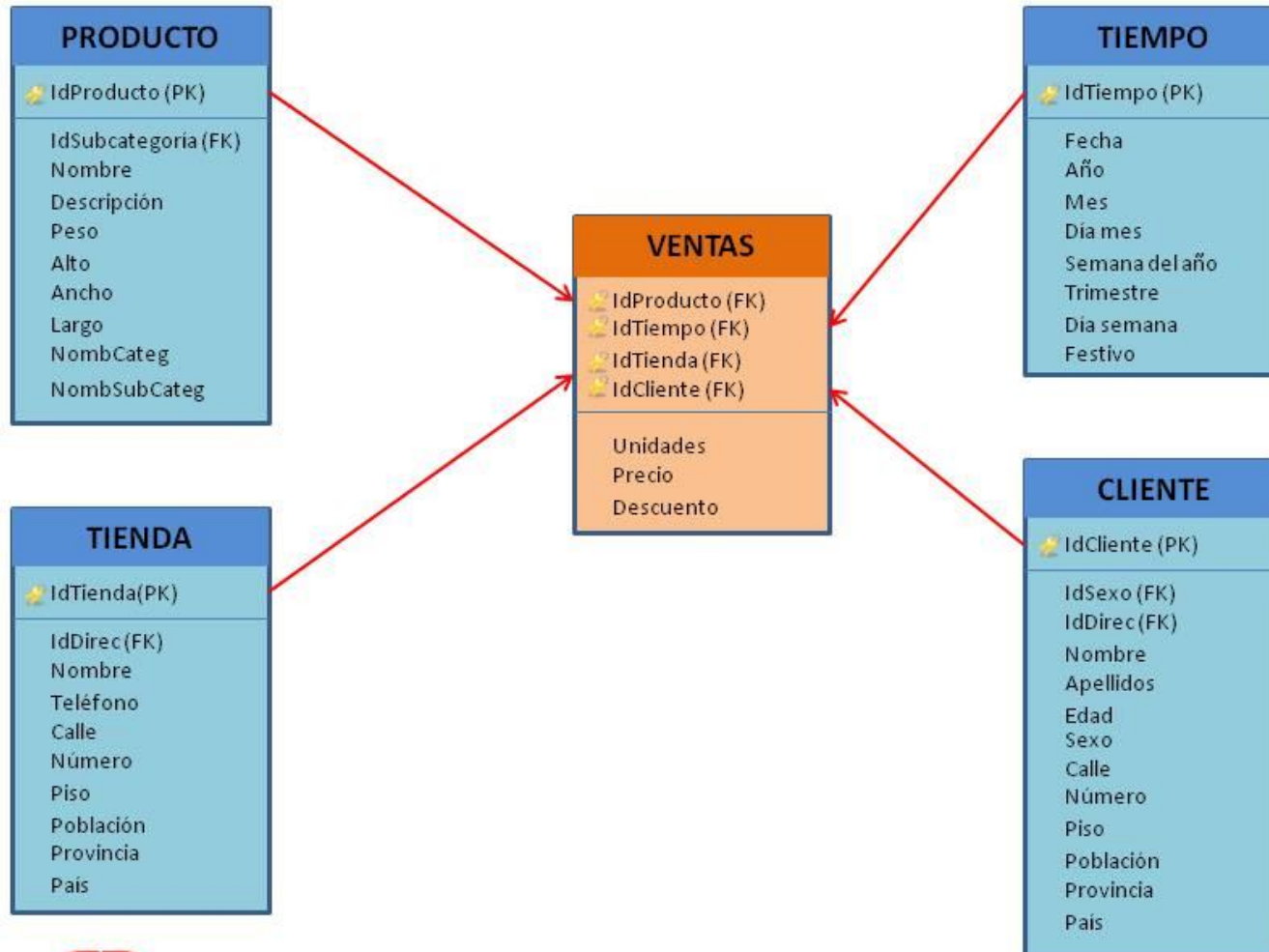


❖ **estrella jerárquica o copo de nieve**: si la jerarquía no es lineal.



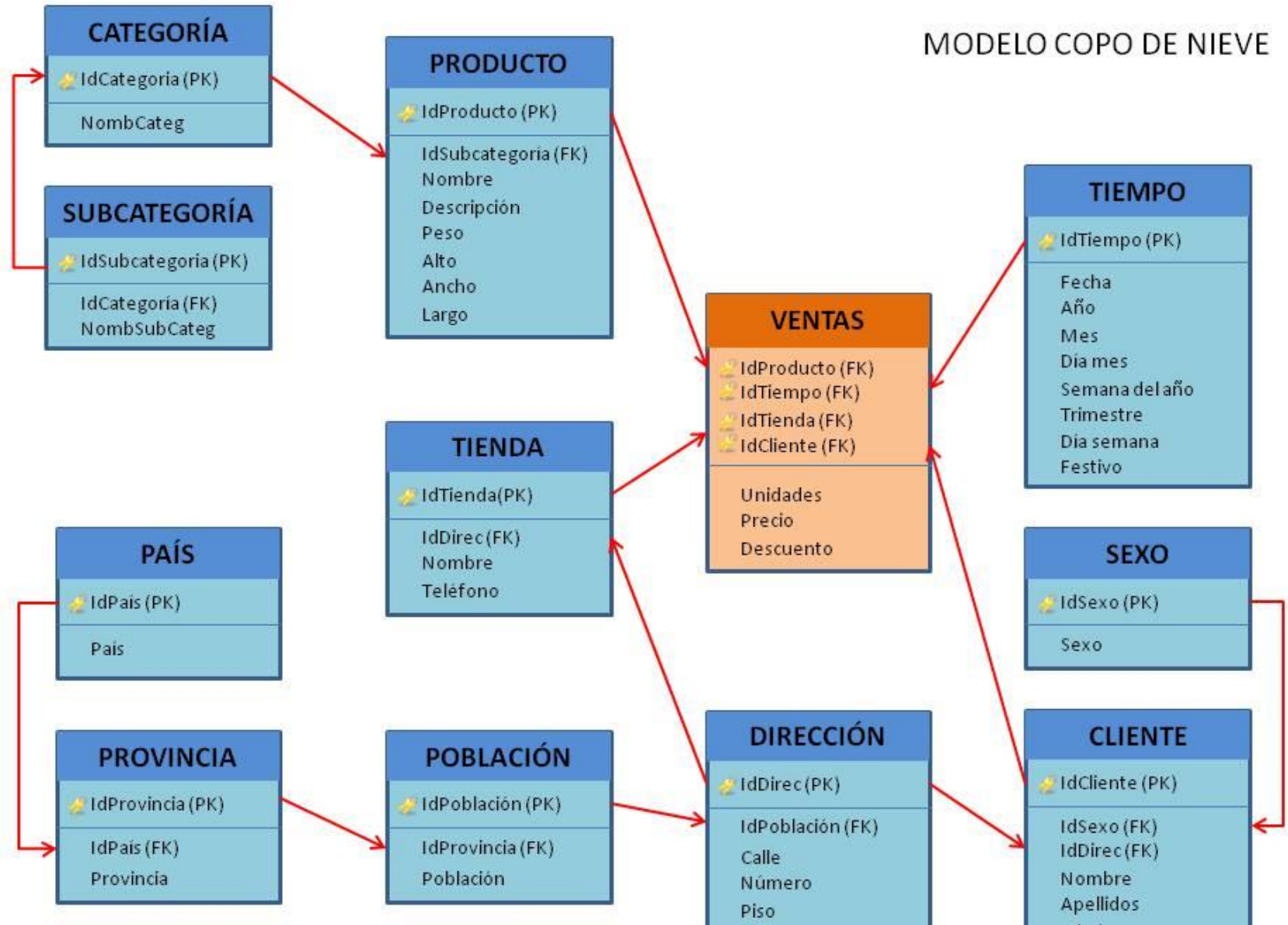
# Arquitectura de un DW

## MODELO ESTRELLA



# Arquitectura de un DW

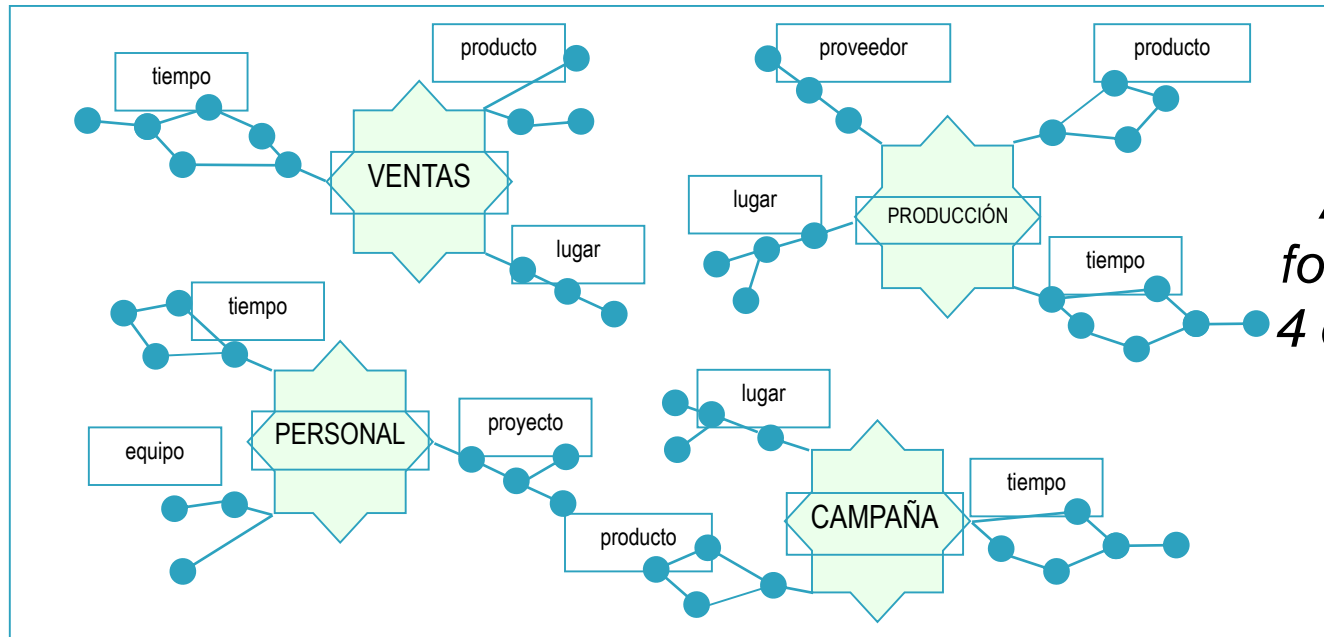
MODELO COPO DE NIEVE



# Arquitectura de un DW

Cada uno de los esquemas del que esta compuesto un DW se denomina un Datamart.

Datamart: subconjunto de un almacén de datos, generalmente en forma de estrella o copo de nieve.



*Almacén  
formado por  
4 datamarts.*



# Procesamiento Analítico en Línea. OLAP(On-Line Analytical Processing )

- 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).

# 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

# Herramientas OLAP

