

Data Warehouse

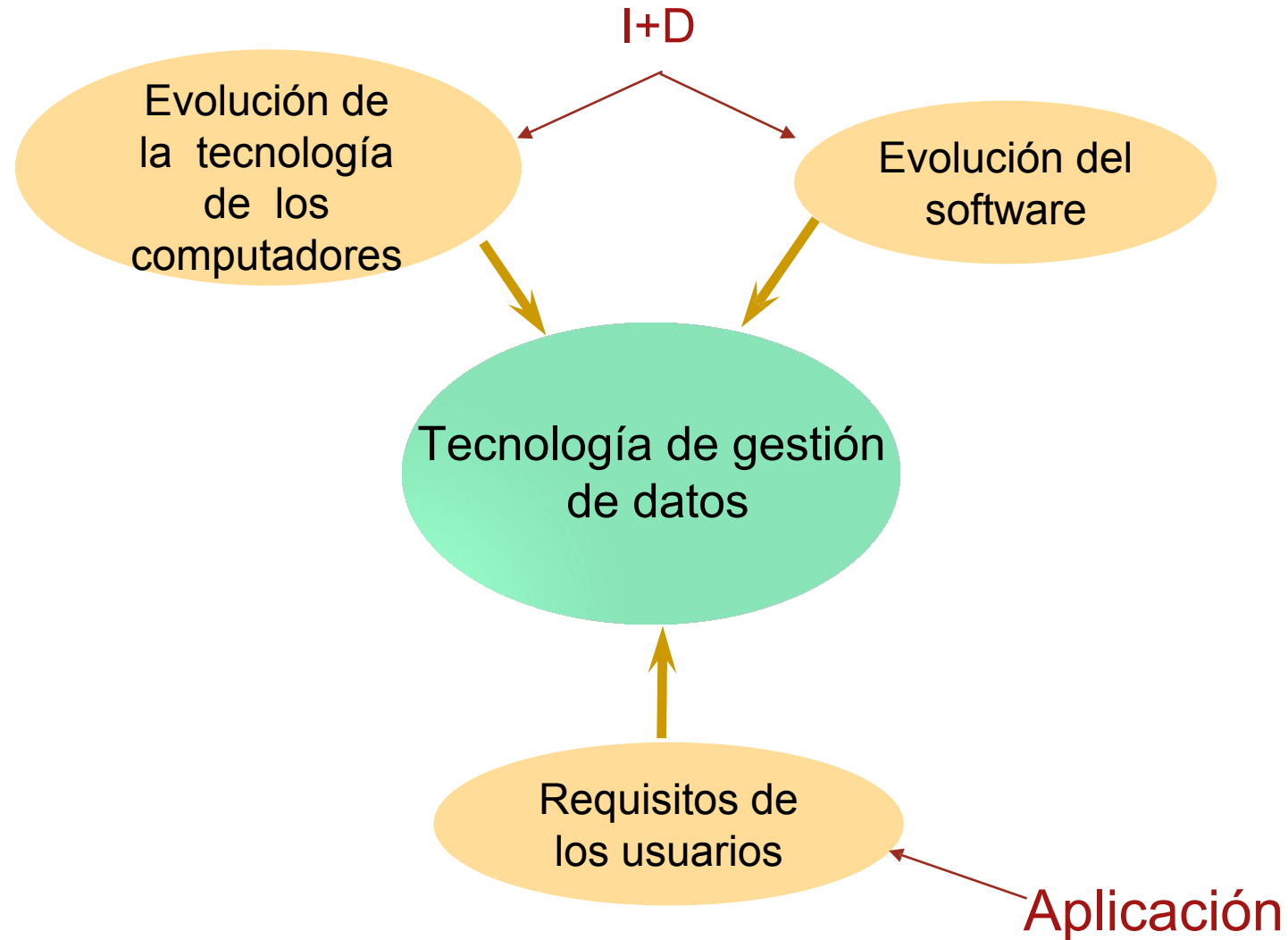
Almacenes de datos (Data Warehouse)

1. Introducción a los almacenes de datos: motivación definición y características.
2. Arquitectura de un sistema de almacén de datos.
3. Explotación de un almacén de datos: herramientas OLAP

Almacenes de datos (Data Warehouse)

1. Introducción a los almacenes de datos: motivación definición y características.

1. Introducción a los almacenes de datos



1. Introducción a los almacenes de datos

Situación actual de la tecnología de gestión de datos.



- SGBD eficientes
- SGBD robustos
- lenguajes y herramientas de uso de alto nivel

Situación actual: uso extendido de los SGBD

- ✓ BD son el soporte del Sistema de Información de las organizaciones
- ✓ BD son diseñadas para dar soporte (eficiente) a las funciones básicas de la organización (ventas, producción, personal...)

SISTEMAS OPERACIONALES (OLTP On Line Transaction Processing)

- ✓ las organizaciones almacenan grandes volúmenes de datos con información histórica

1. Introducción a los almacenes de datos

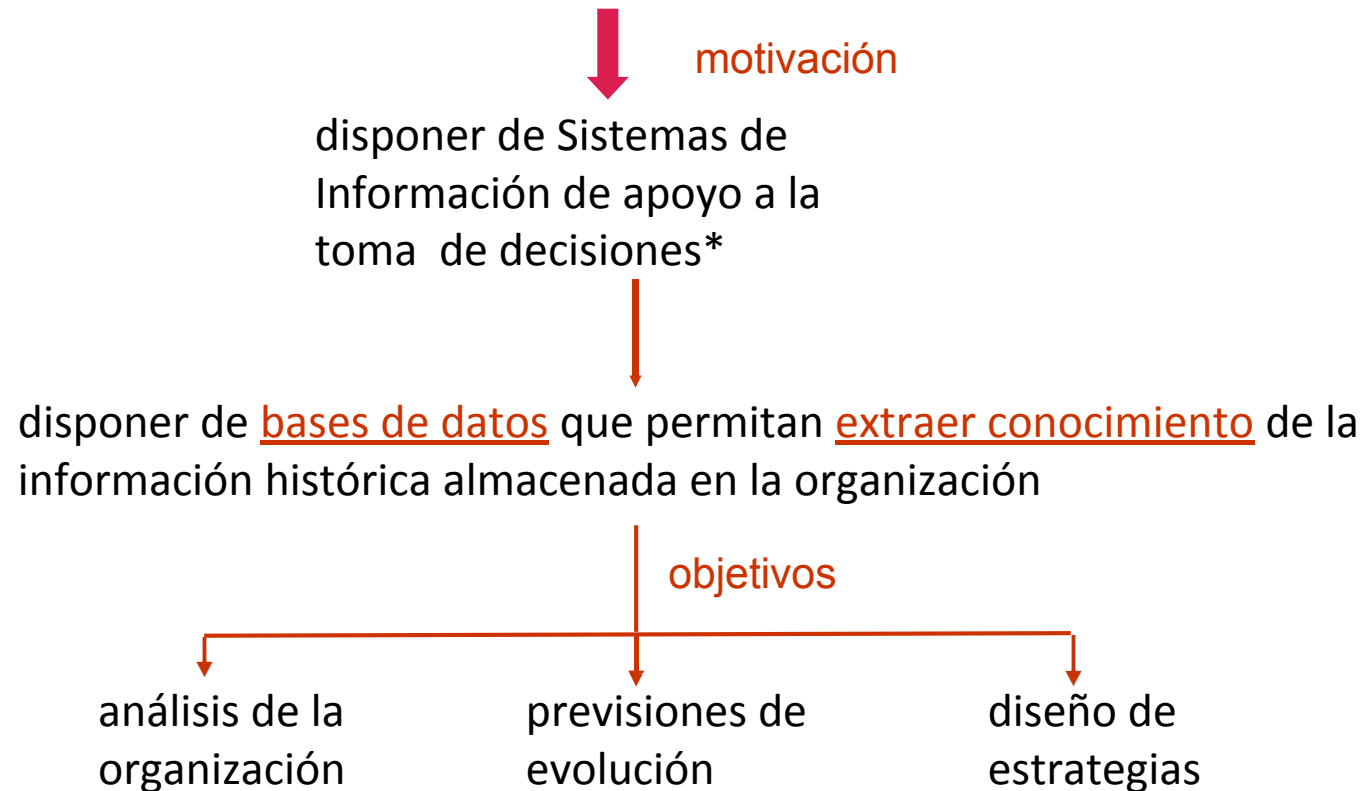
Una vez satisfecha la necesidad de tener un soporte informático para los procesos básicos de la organización (**sistemas de información para la gestión**).



La organizaciones exigen nuevas prestaciones de los sistemas de información (**sistemas de información para la toma de decisiones**).

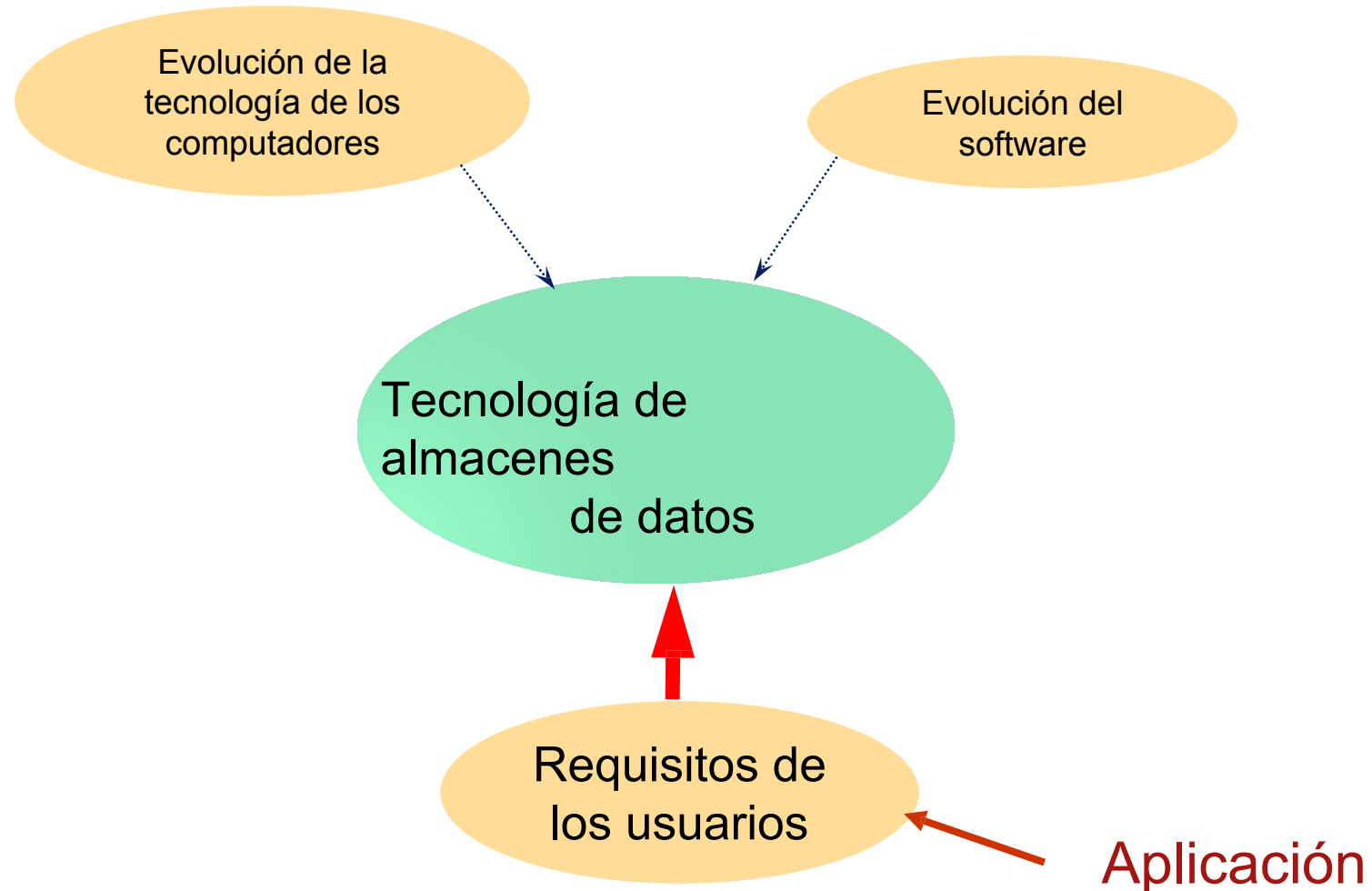
1. Introducción a los almacenes de datos

Almacenes de datos (AD) (data warehouse)



* DSS: Decision Support Systems

1. Introducción a los almacenes de datos

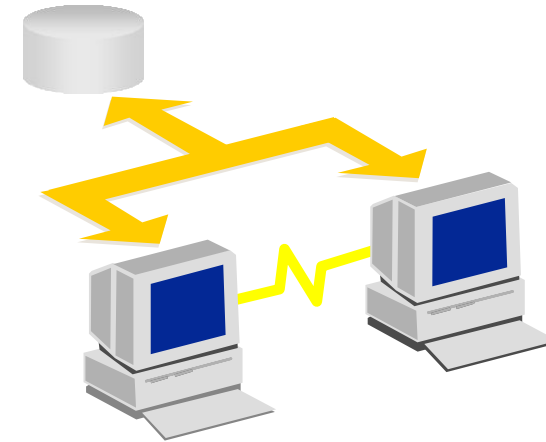
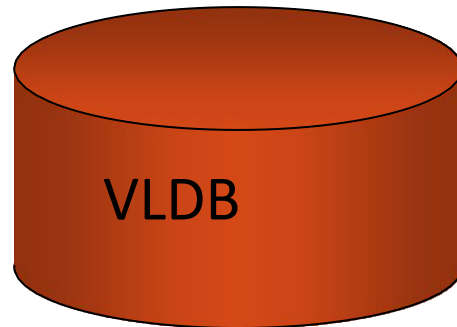


1. Introducción a los almacenes de datos

Avances tecnológicos que han favorecido el desarrollo de la tecnología de almacenes de datos

— Paralelismo

- Hardware
- Sistemas Operativos
- Bases de Datos
- Consultas
- Índices



- VLD (bases de datos muy grandes)
- Arquitecturas de 64 bit
- Técnicas de indexación
- Sistemas abiertos
- Herramientas y sistemas para DW
- Herramientas de análisis de usuario final

1. Introducción a los almacenes de datos

Almacén de datos

(Data

Warehouse)



definición

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

1. Introducción a los almacenes de datos

Según Oracle: “Un data warehouse es una colección de datos orientada a temas específicos, integrada no volátil, y variante en el tiempo, organizada para apoyar las necesidades de administración

Según Informix: “data warehouse no es un producto que puede ser comprado en el mercado, sino mas bien es un concepto que debe de ser construido”.

1. Introducción a los almacenes de datos

Almacén 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

1. Introducción a los almacenes de datos

Almacén de datos



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

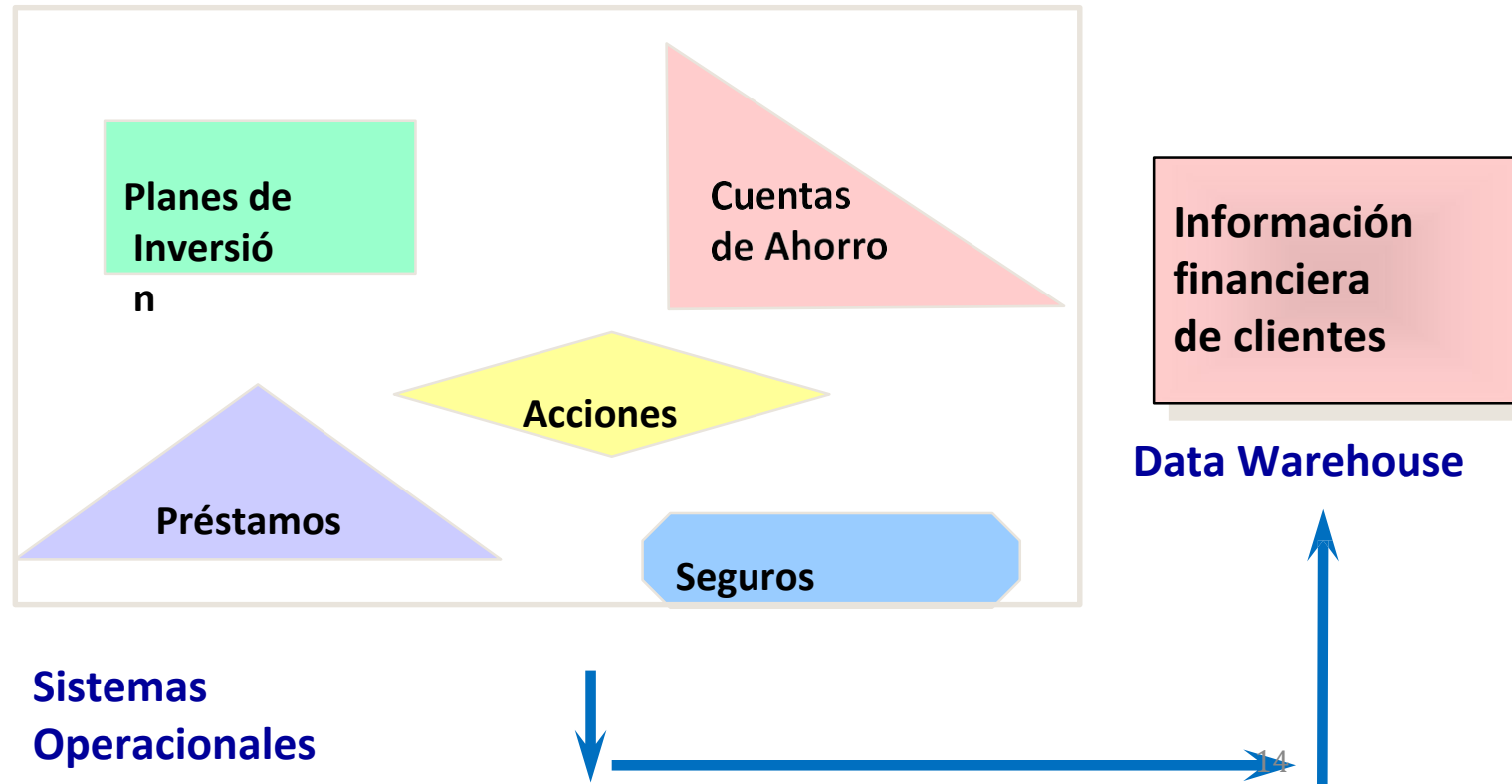
** subject oriented, not process oriented*

1. Introducción a los almacenes de datos

Orientada hacia la
información
relevante de la
organización

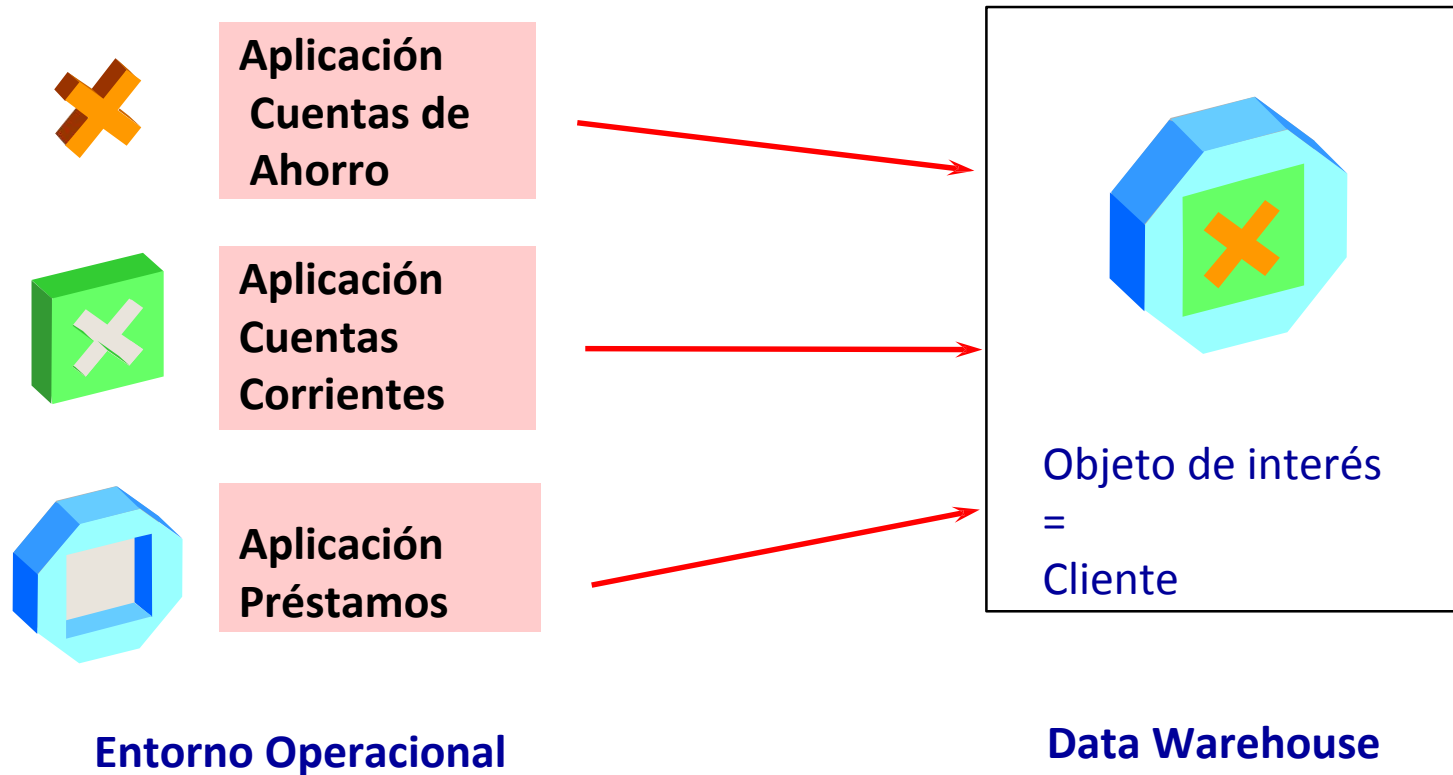


se diseña para consultar eficientemente información
relativa a las actividades (ventas, compras, producción,
...) básicas de la organización, no para soportar los
procesos que se realizan en ella (gestión de
pedidos, facturación, etc).



1. Introducción a los almacenes de datos

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



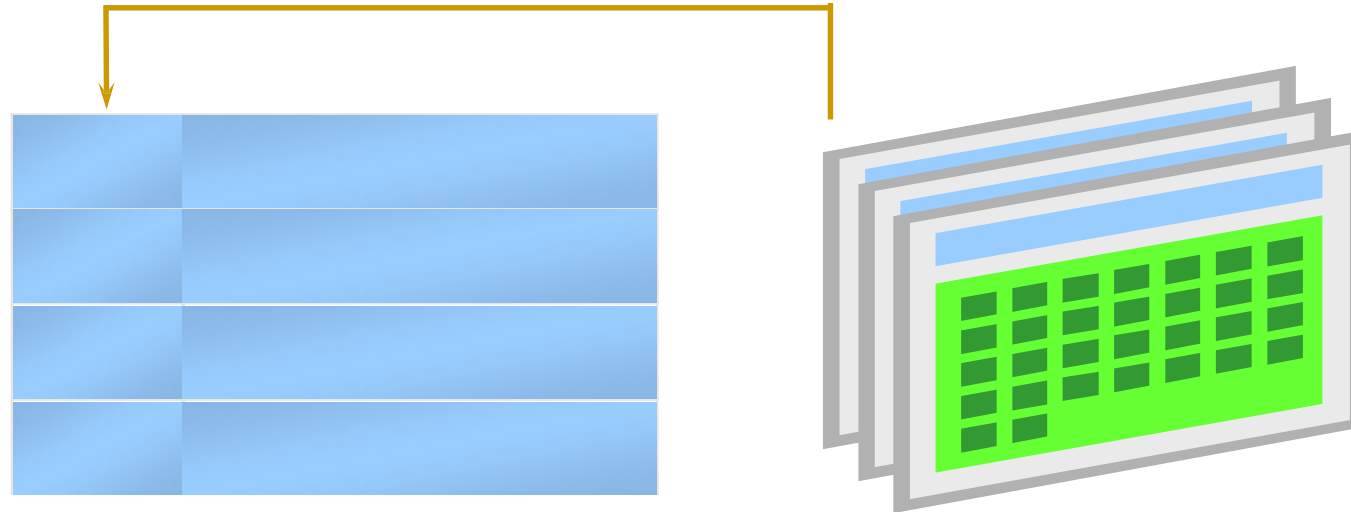
1. Introducción a los almacenes de datos

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.

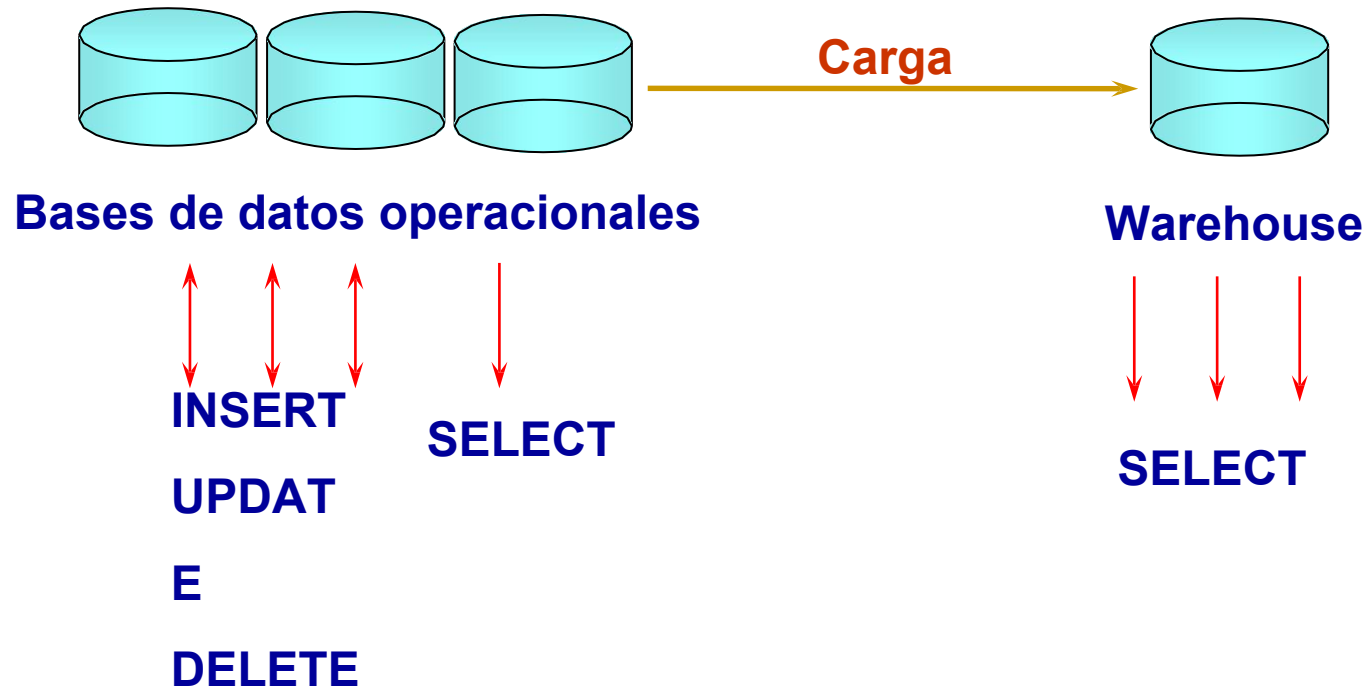


**Data
Warehous
e**

Nota: El periodo de tiempo cubierto por un DW varia entre 2 y 10 años.

1. Introducción a los almacenes de datos

No volátil → los datos almacenados no son actualizados, sólo son incrementados.



1. Introducción a los almacenes de datos

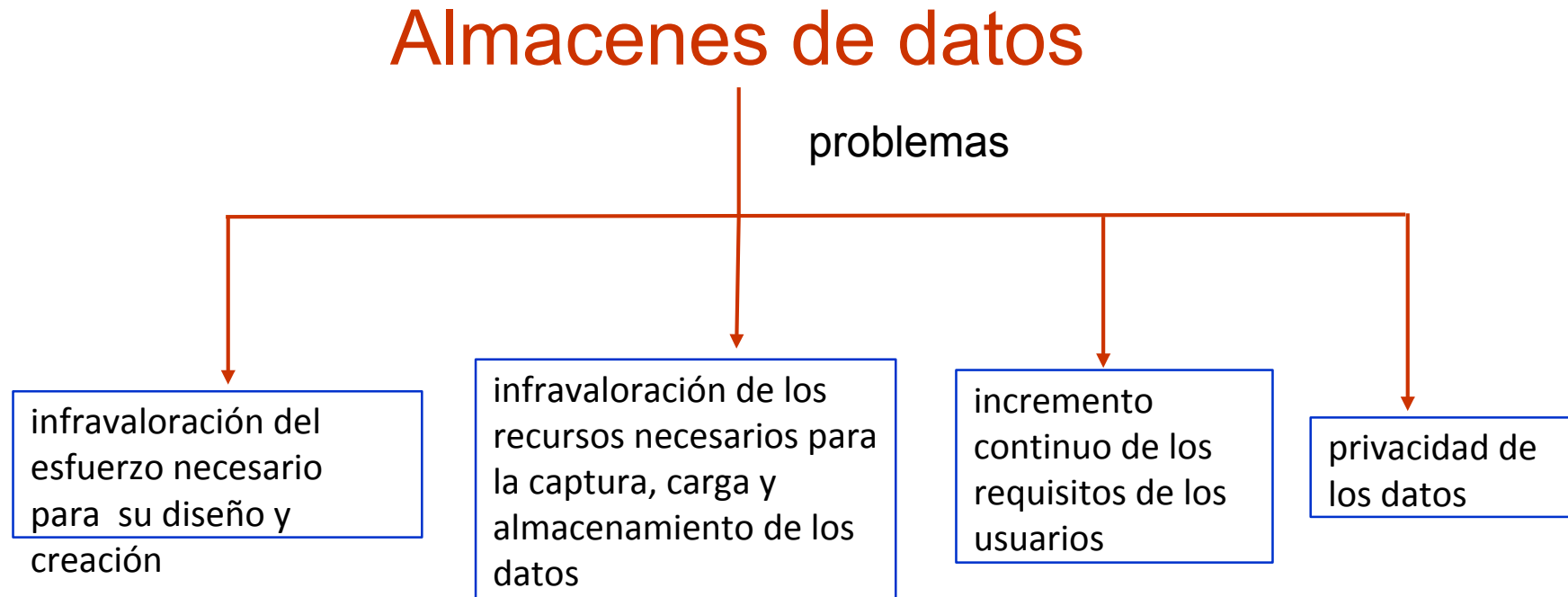
- El **objetivo** último de un **almacén de datos** es integrar datos corporativos, residentes en bases de datos operacionales de la organización, en un único repositorio sobre el cual los usuarios pueden realizar consultas o informes y hacer análisis de datos.

➡ La **tecnología de almacenes de datos** integra las técnicas de bases de datos y las técnicas de análisis de datos.

1. Introducción a los almacenes de datos



1. Introducción a los almacenes de datos



1. Introducción a los almacenes de datos

Sistema Operacional (OLTP)

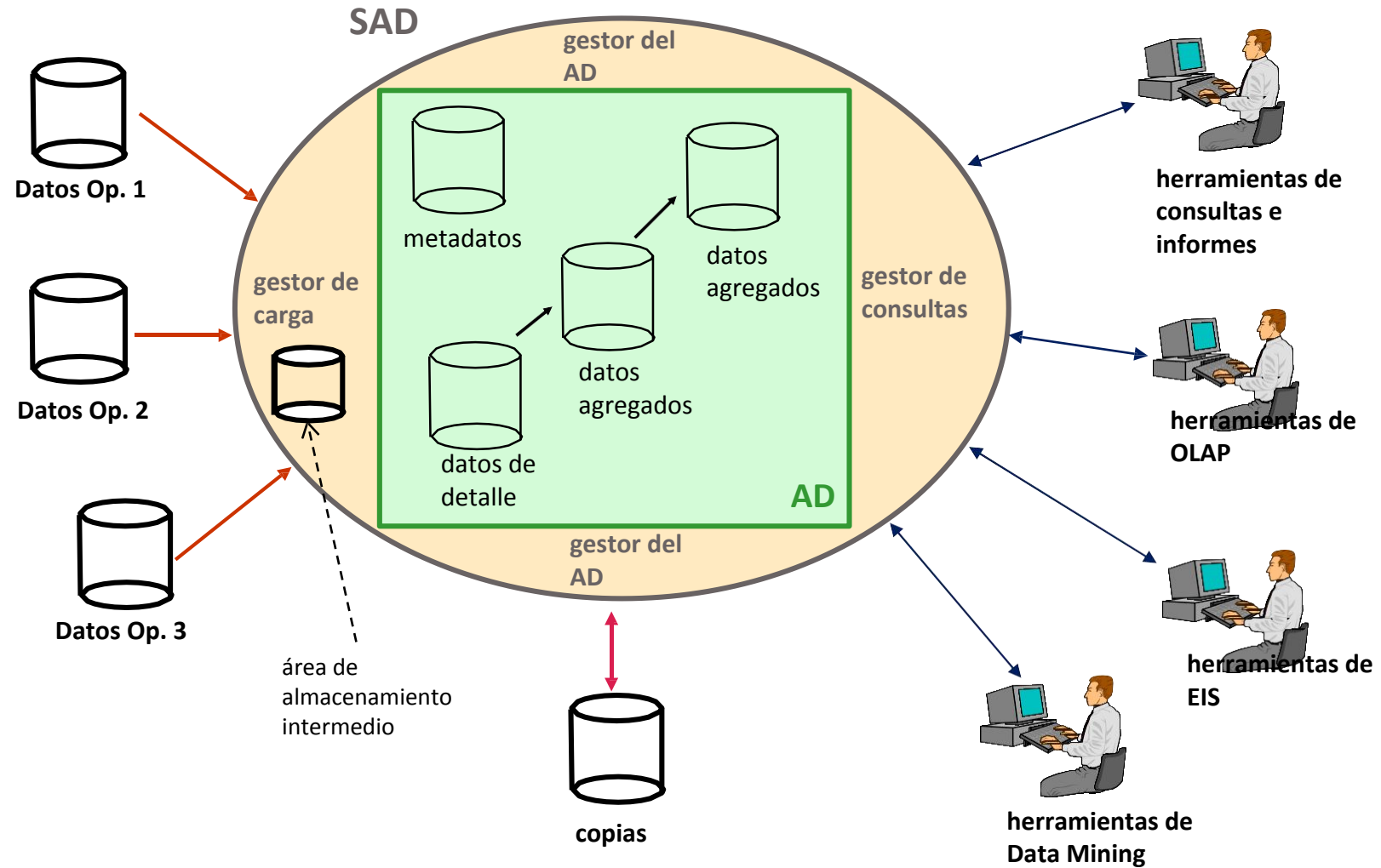
- almacena datos actuales
- almacena datos de detalle
- bases de datos medianas
- los datos son dinámicos (actualizables)
- los procesos (transacciones) son repetitivos
- el número de transacciones es elevado
- tiempo de respuesta pequeño (segundos)
- dedicado al procesamiento de transacciones
- orientado a los procesos de la organización
- soporta decisiones diarias
- sirve a muchos usuarios (administrativos)

Almacén de datos (DW)

- almacena datos históricos
- almacena datos de detalle y datos agregados a distintos niveles
- bases de datos grandes
- los datos son estáticos
- los procesos no son previsibles
- el número de transacciones es bajo o medio
- tiempo de respuesta variable (segundos-horas)
- dedicado al análisis de datos relevante
- orientado a la información
- soporta decisiones estratégicas
- sirve a técnicos de dirección

2. Arquitectura de un sistema de almacén de datos.

2. Arquitectura de los almacenes de datos



2. Arquitectura de los almacenes de datos

Componentes:

- **datos operacionales:** el origen de los datos puede ser: bases de datos operacionales de la organización, bases de datos privadas, bases de datos públicas, etc.
- **gestor de carga:** permite realizar las funciones de extracción de datos de las fuentes externas, transformación (limpieza, consolidación, ...) y la carga del AD, utiliza un almacenamiento intermedio y realiza las siguientes operaciones:
 - extracción de los datos.
 - transformación de los datos: limpieza, estandarización, etc.
 - carga inicial del almacén: ordenación, agregaciones, etc.
 - refresco del almacén: operación periódica que propaga los cambios de las fuentes operacionales al almacén de datos

(herramientas* del fabricante o programas de la organización).

- ETT (extracción, transformación y transporte)
- ETL (extracción, transformación y carga (load))

2. Arquitectura de los almacenes de datos

Componentes:

- **gestor (servidor) del AD:** permite realizar todas las funciones de definición y mantenimiento del almacén de datos: definición, agregación de datos, vistas, creación de índices, copias, etc. (herramienta del fabricante).
- **gestor de consultas:** ejecución de consultas. (herramienta del fabricante).
- **metadatos:** documentación sobre los datos (origen, descripción, nivel de agregación, almacenamiento, etc).
- **herramientas de consulta:** herramientas para diseñar consultas e informes, herramientas de desarrollo de aplicaciones de usuario final, herramientas de análisis de datos (OLAP), herramientas de minería de datos (DATA MINING), herramientas dirigidas a ejecutivos (EIS). (herramientas de diferentes fabricantes).

2. Arquitectura de los almacenes de datos

El servidor (gestor del almacén de datos) puede estar construido usando:

- ✓ tecnología multidimensional (sistemas MOLAP): sistemas de gestión de bases de datos contruidos específicamente para el análisis de datos (estructuras de almacenamiento, optimizadores de consultas, etc.). (Express de ORACLE)
- ✓ tecnología relacional (sistemas ROLAP): SGBD relacionales con ciertas extensiones. Sobre estos sistemas relacionales se acoplan herramientas de OLAP.
 - MicroStrategy: herramienta OLAP que trabaja sobre ACCESS, ORACLE, SQL Server, ...
 - Discoverer: herramienta OLAP de ORACLE.

2. Arquitectura de los almacenes de datos

Data mart  subconjunto de un almacén de datos

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

En la **construcción de un data mart** se siguen dos aproximaciones:

- ✓ definir previamente el almacén de datos de la organización
y
posteriormente definir sobre él/los *data marts*.
- ✓ definir previamente los *data marts* de departamentos y
posteriormente integrarlos en un almacén de datos para
la organización

3. Explotación de un almacén de datos: herramientas OLAP.

3. Explotación de un almacén de datos: herramientas OLAP

Las herramientas (OLAP) 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.

3. Explotación de un almacén de datos: herramientas OLAP

EJEMPLO

Organización: Cadena de supermercados.

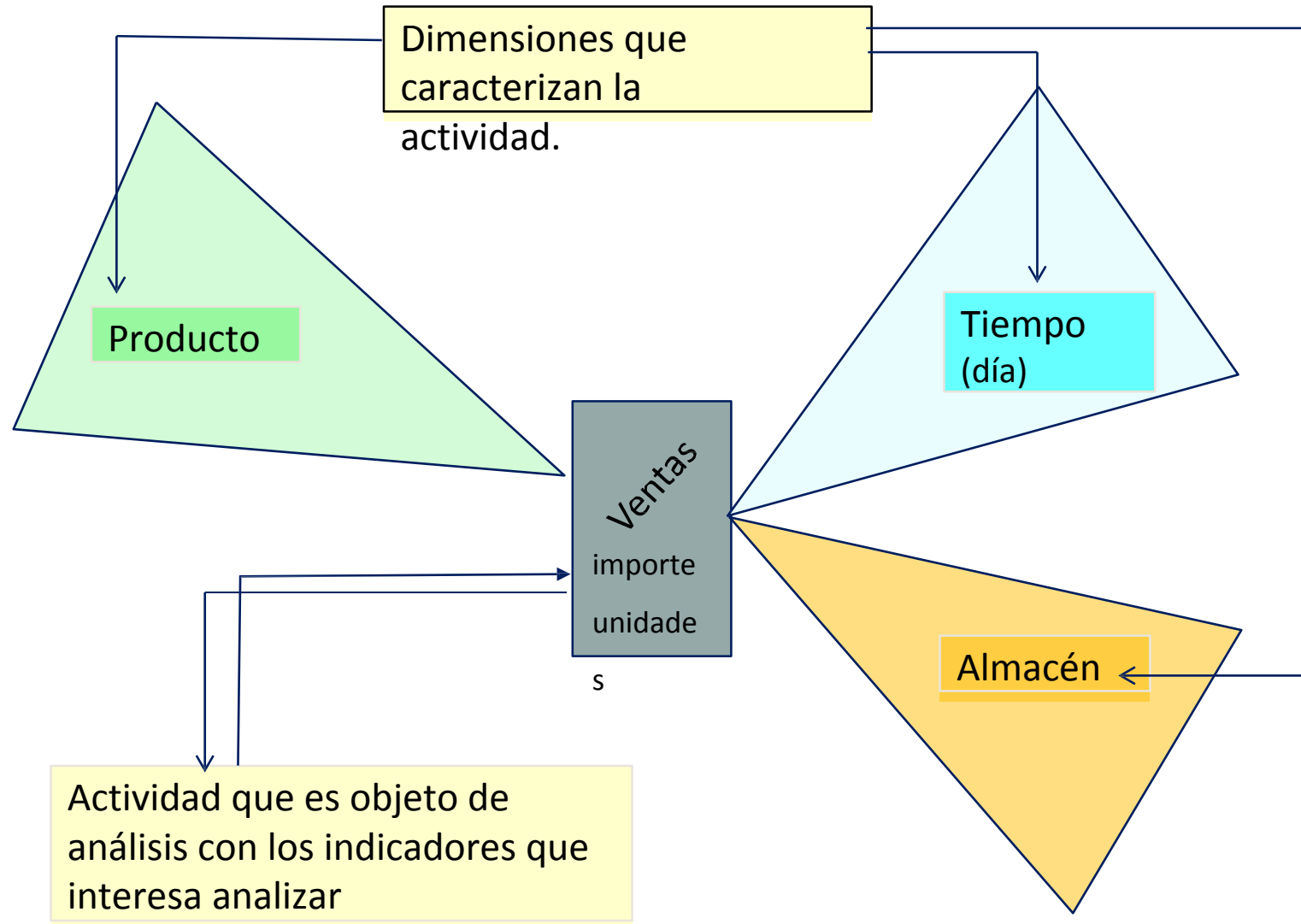
Actividad objeto de análisis: ventas de productos.

Información registrada sobre las ventas: "ventas diarias de productos en los supermercados de la cadena".

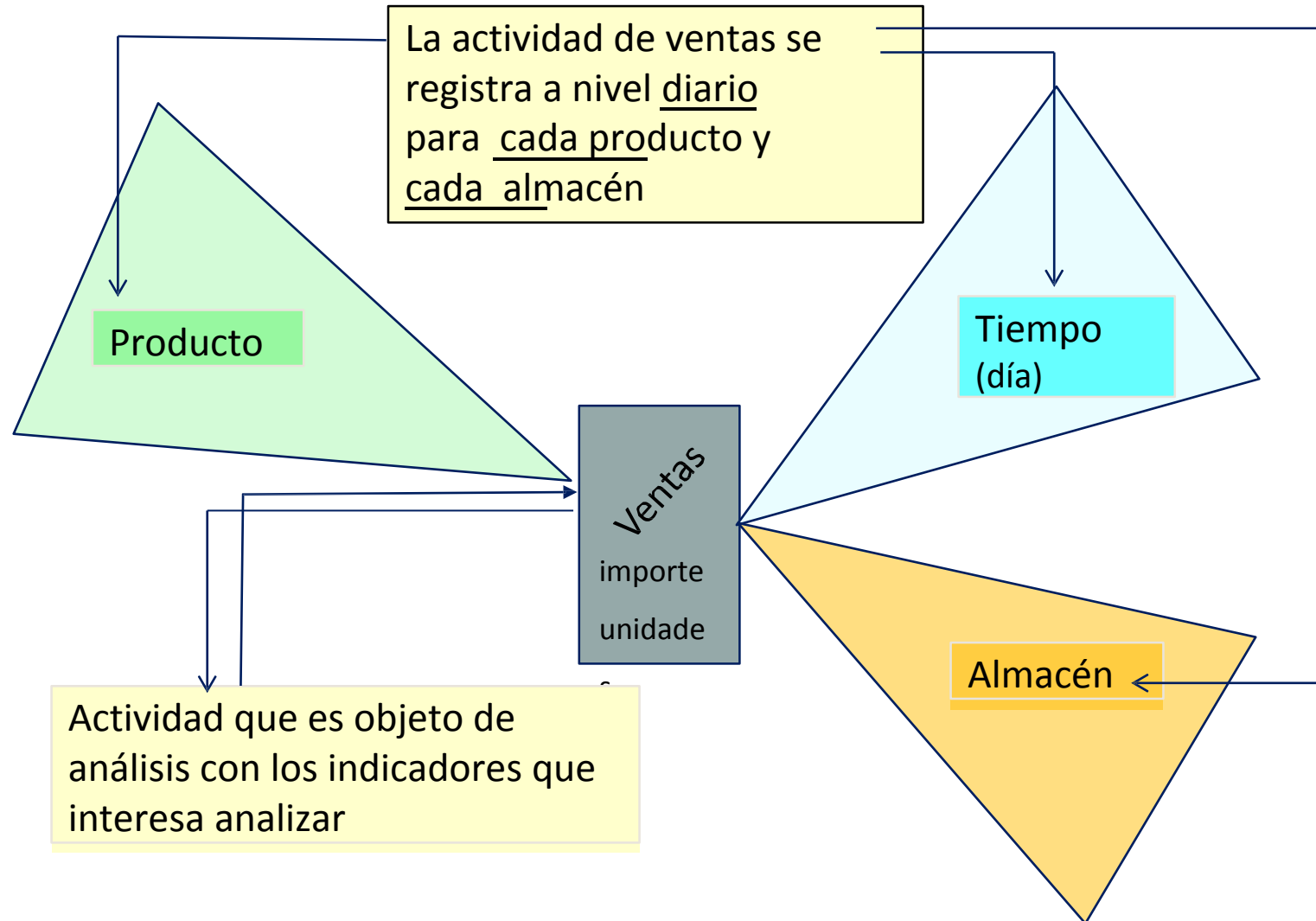
Ejemplo: "del **producto** "Coca-Cola 33cl" se han vendido en el **almacén** "Almacén nro.1" el **día** 12/01/1999, 50 **unidades** por un **importe** de 70€."

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

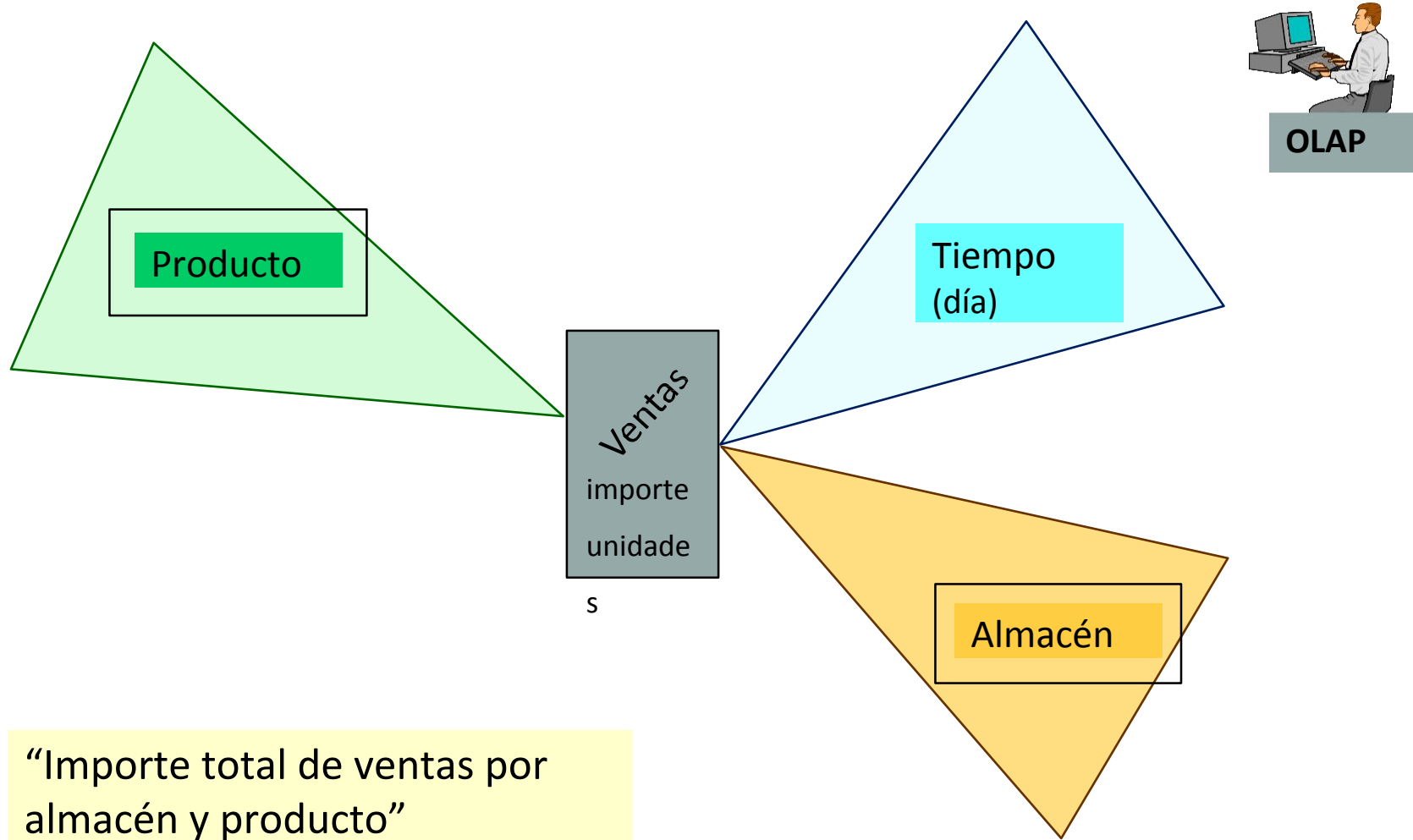
3. Explotación de un almacén de datos: herramientas OLAP



3. Explotación de un almacén de datos: herramientas OLAP



3. Explotación de un almacén de datos: herramientas OLAP



3. Explotación de un almacén de datos: herramientas OLAP



Producto Almacén	Leche	Vino	Jabón	Agua
Almacén1	2000000	1000000	3000000	2000000
Almacén2	1000000	1500000	8000000	2400000

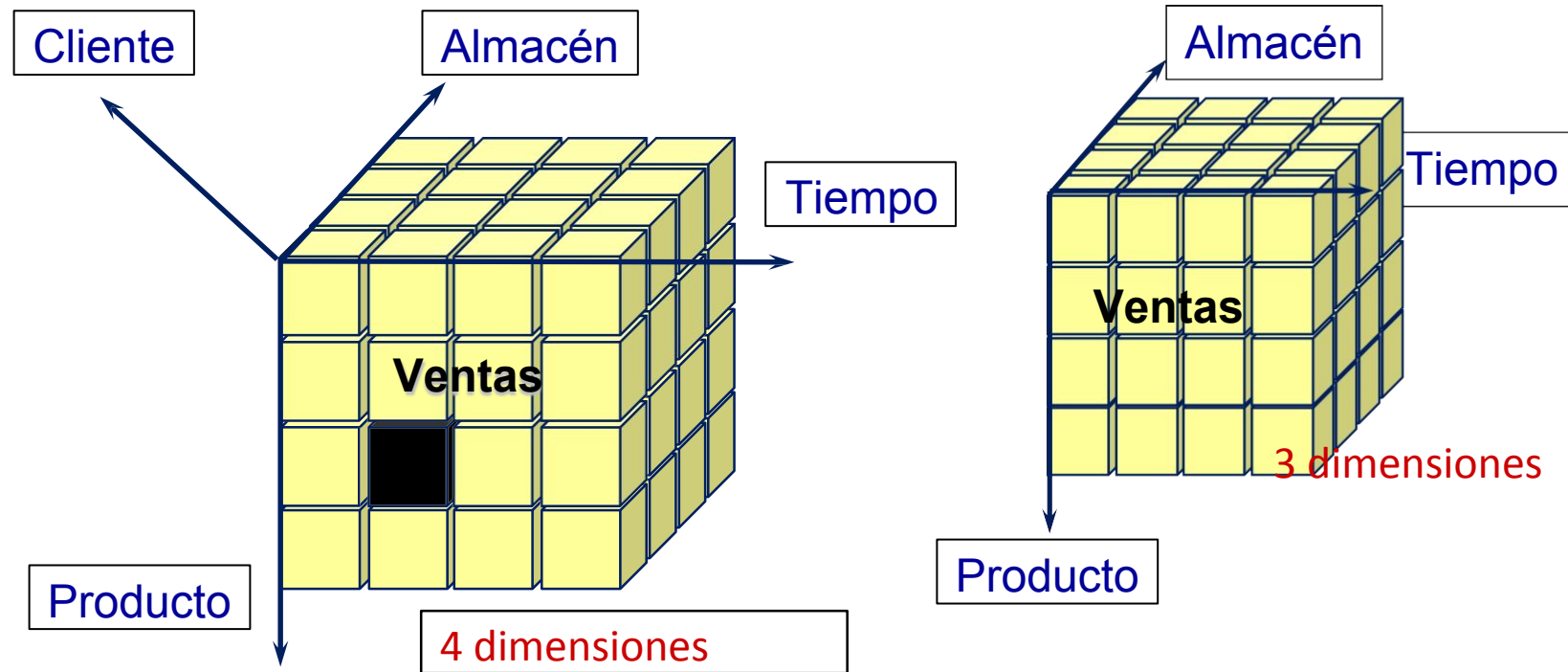
Tabla
multidimensional

3. Explotación de un almacén de datos: herramientas OLAP

Presentaciones del esquema multidimensional
en una herramienta OLAP:

- representación en estrella
- representación en cubo de datos

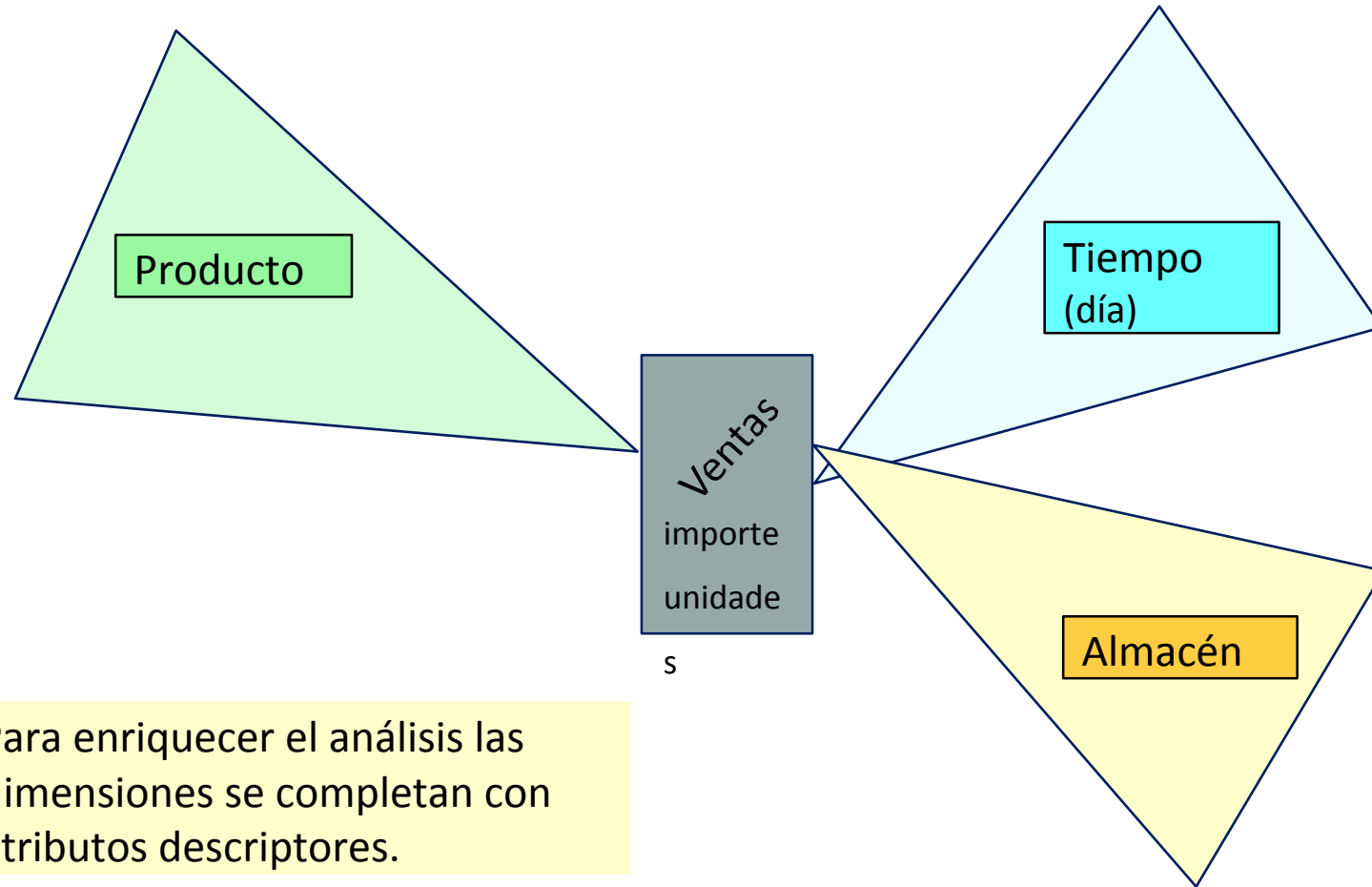
3. Explotación de un almacén de datos: herramientas OLAP



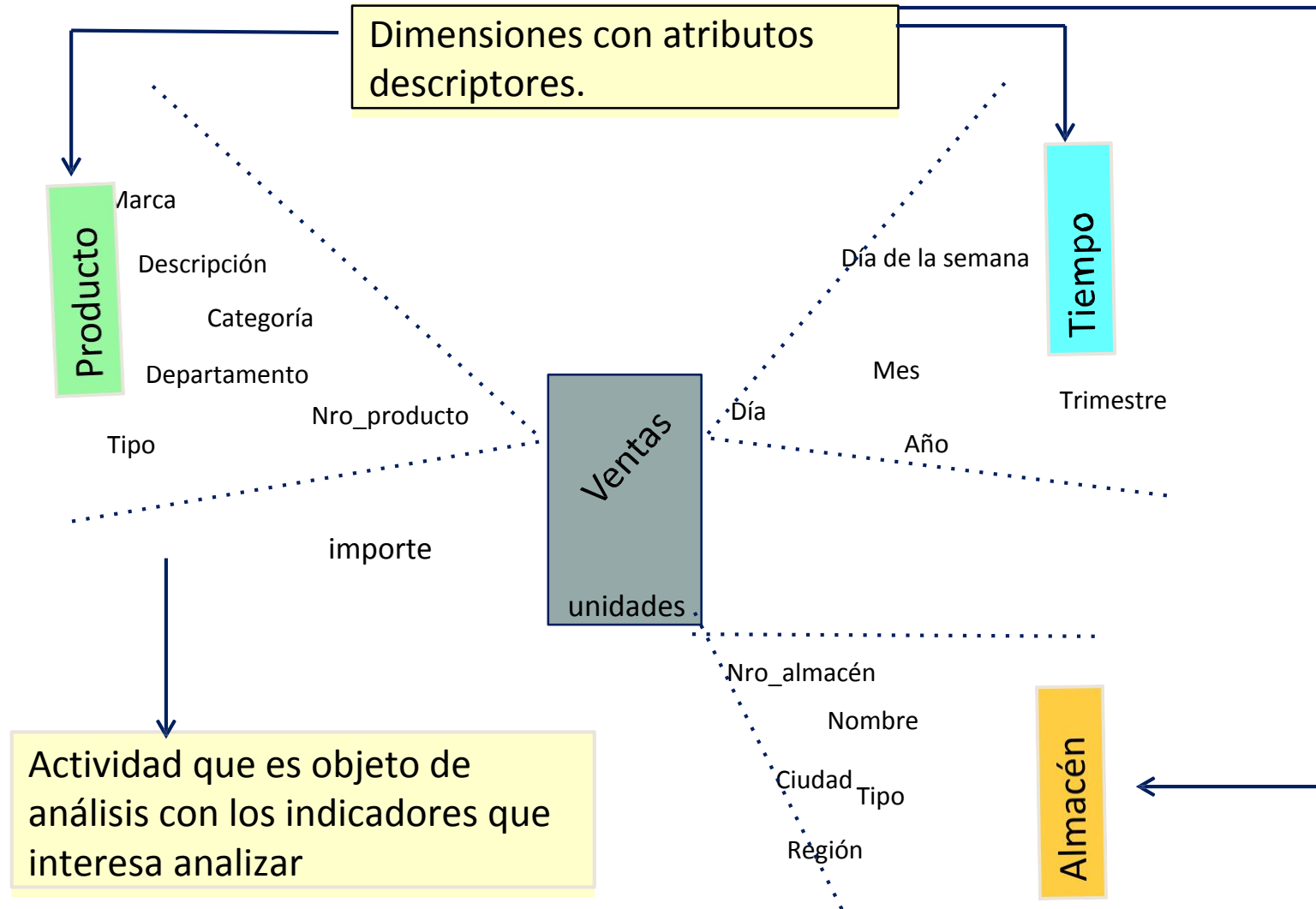
Modelo en cubo de datos:

- actividad (celda)
- dimensiones (ejes)

3. Explotación de un almacén de datos: herramientas OLAP



3. Explotación de un almacén de datos: herramientas OLAP

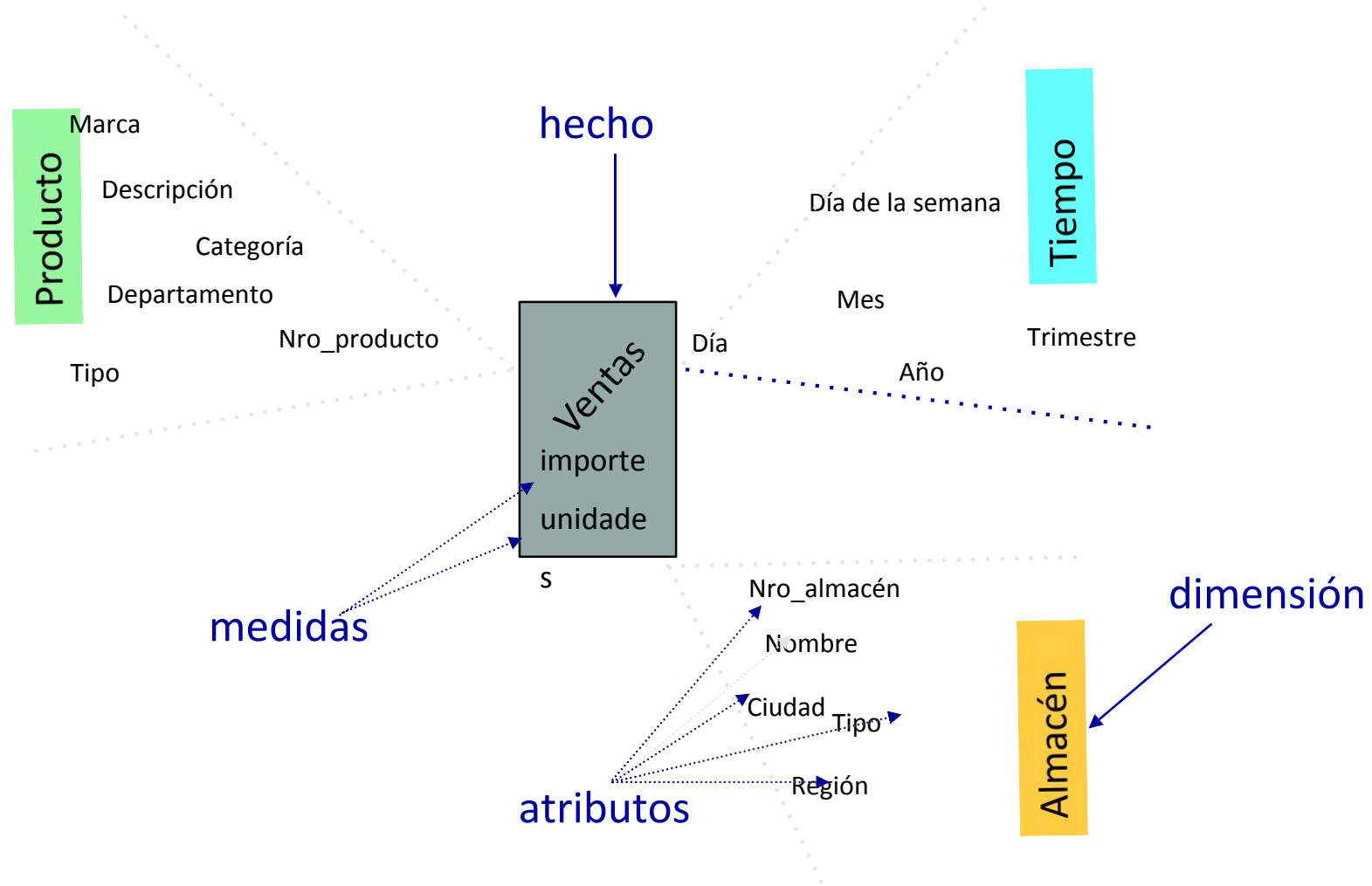


3. Explotación de un almacén de datos: herramientas OLAP

Modelo multidimensional:

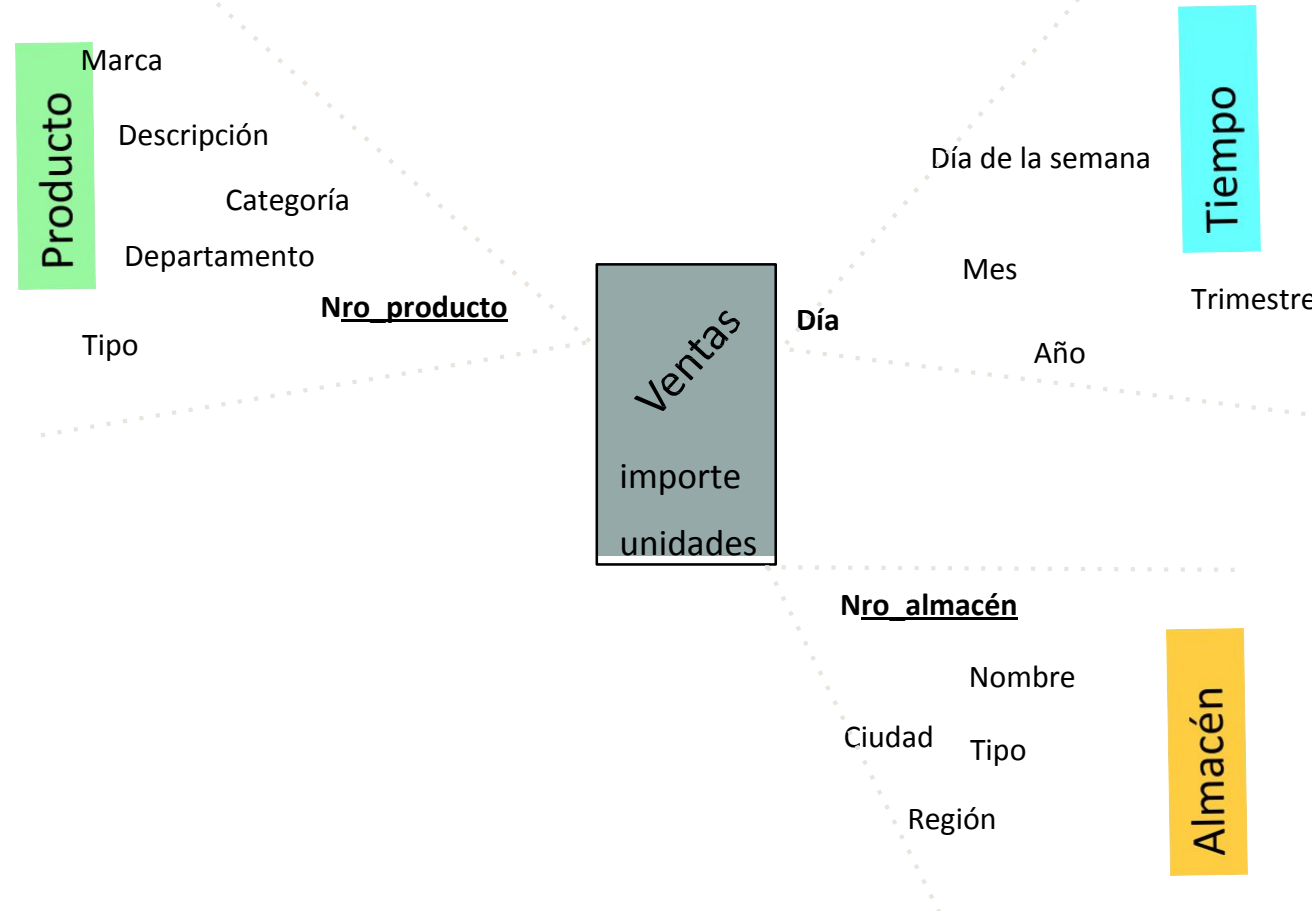
- ✓ en un esquema multidimensional se representa una actividad que es objeto de análisis (**hecho**) y las dimensiones que caracterizan la actividad (**dimensiones**).
- ✓ la información relevante sobre el **hecho** se representa por un conjunto de indicadores (**medidas o atributos de hecho**).
- ✓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

3. Explotación de un almacén de datos: herramientas OLAP



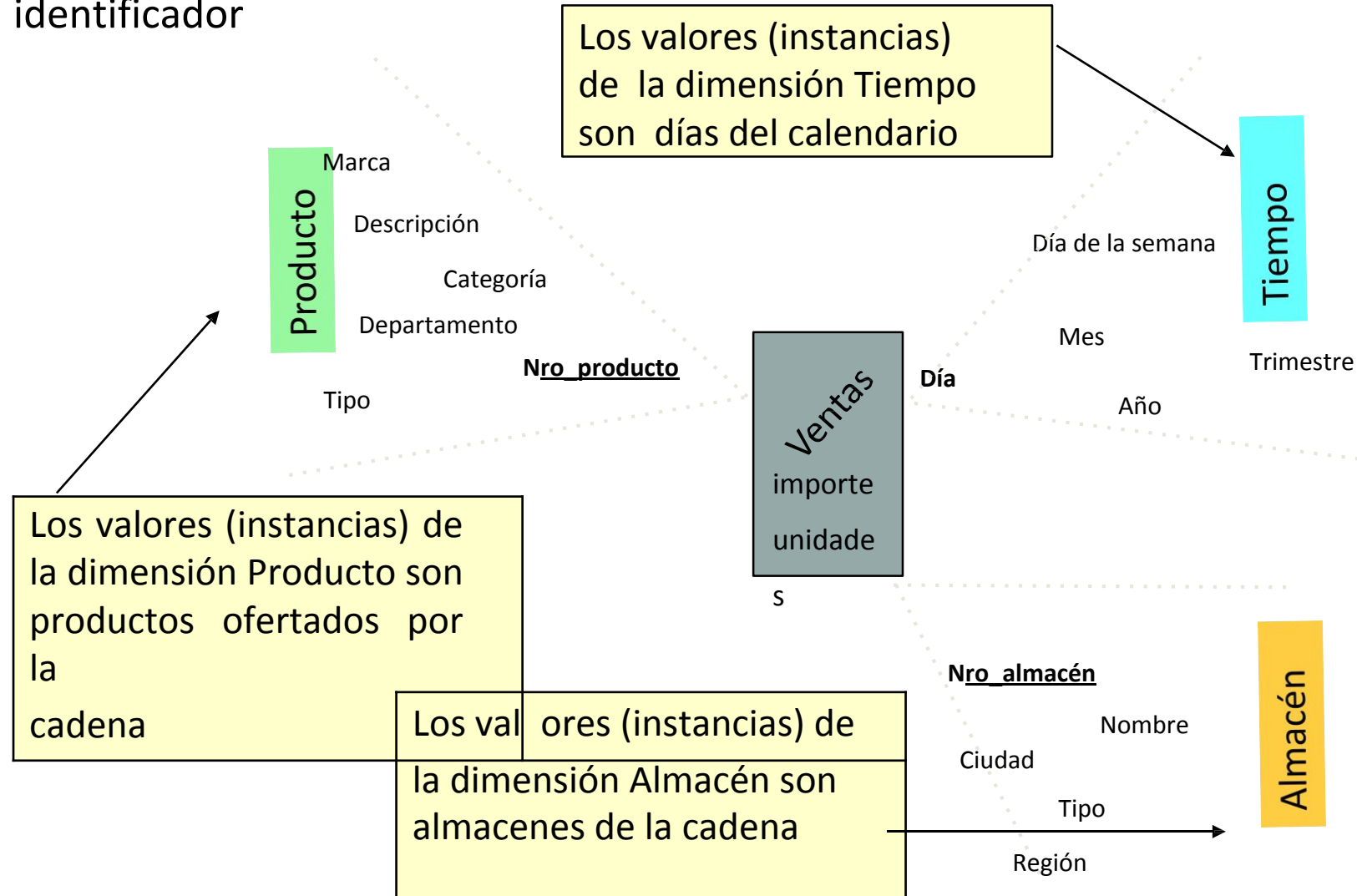
3. Explotación de un almacén de datos: herramientas OLAP

Entre los atributos de una dimensión existe uno que hace la función de identificador



3. Explotación de un almacén de datos: herramientas OLAP

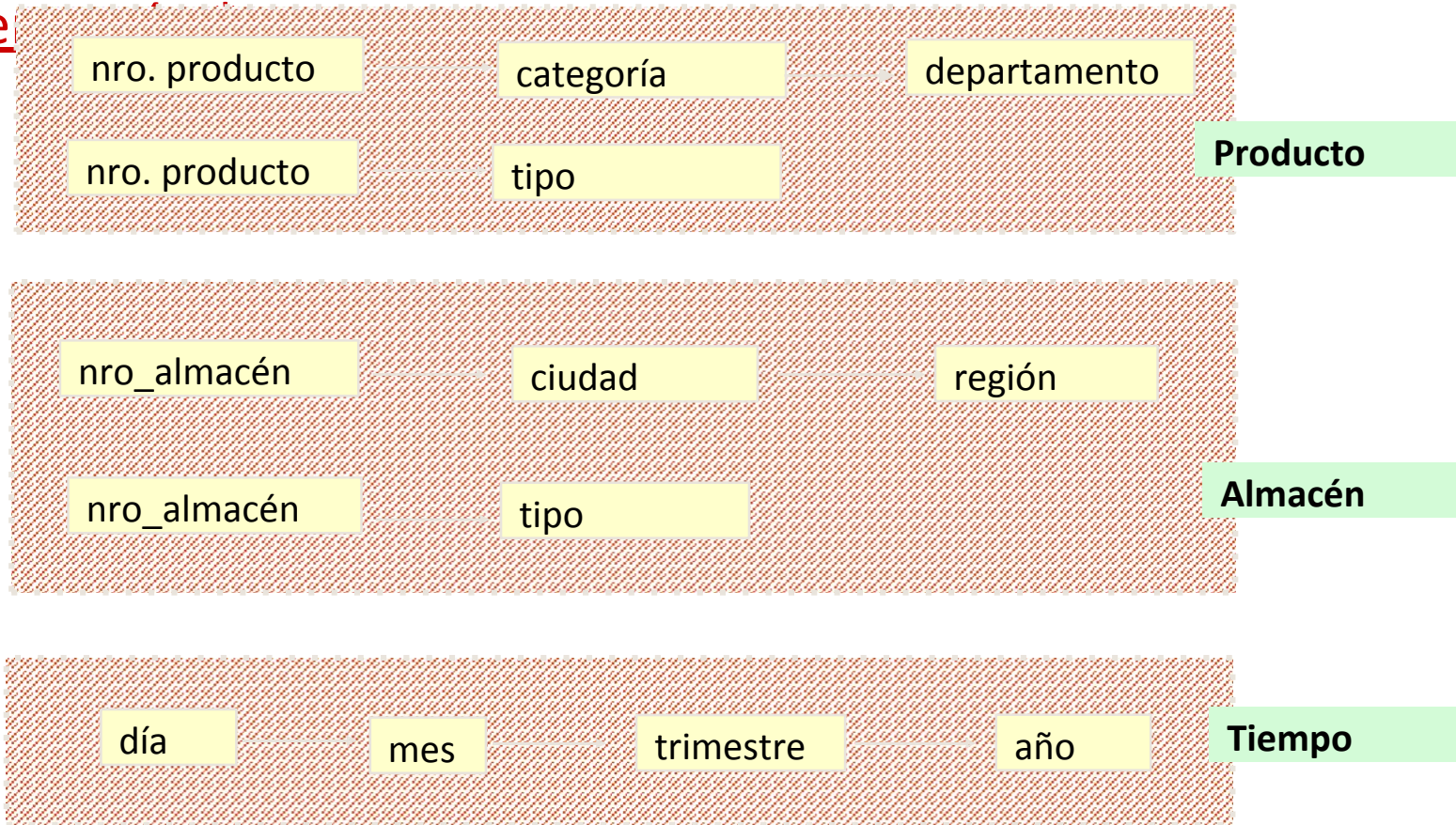
Entre los atributos de una dimensión existe uno que hace la función de identificador



3. Explotación de un almacén de datos: herramientas OLAP

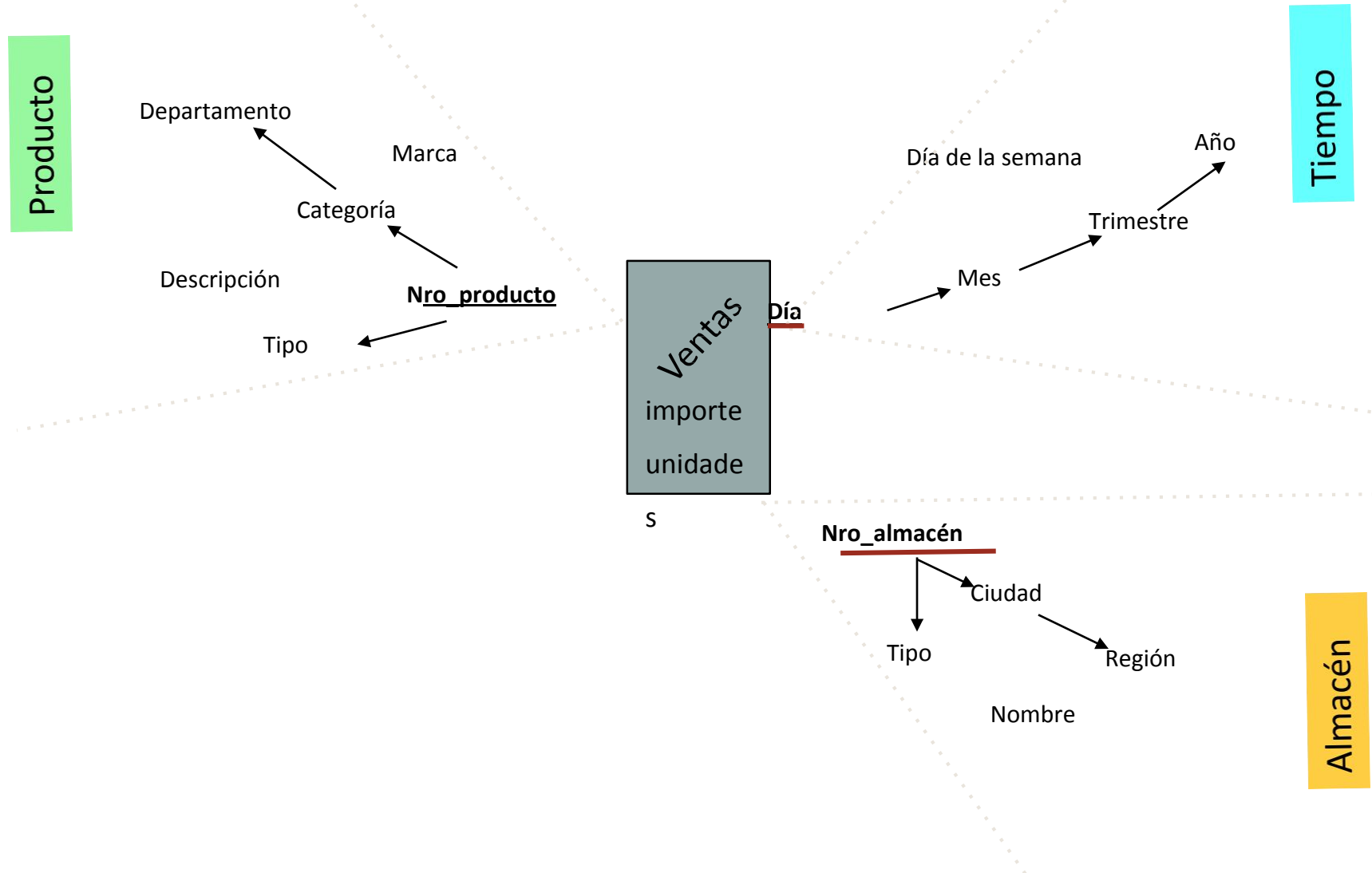
Entre los atributos de una dimensión existen

je



* Jerarquías basadas generalmente en dependencias funcionales entre los atributos de la dimensión.

3. Explotación de un almacén de datos: herramientas OLAP

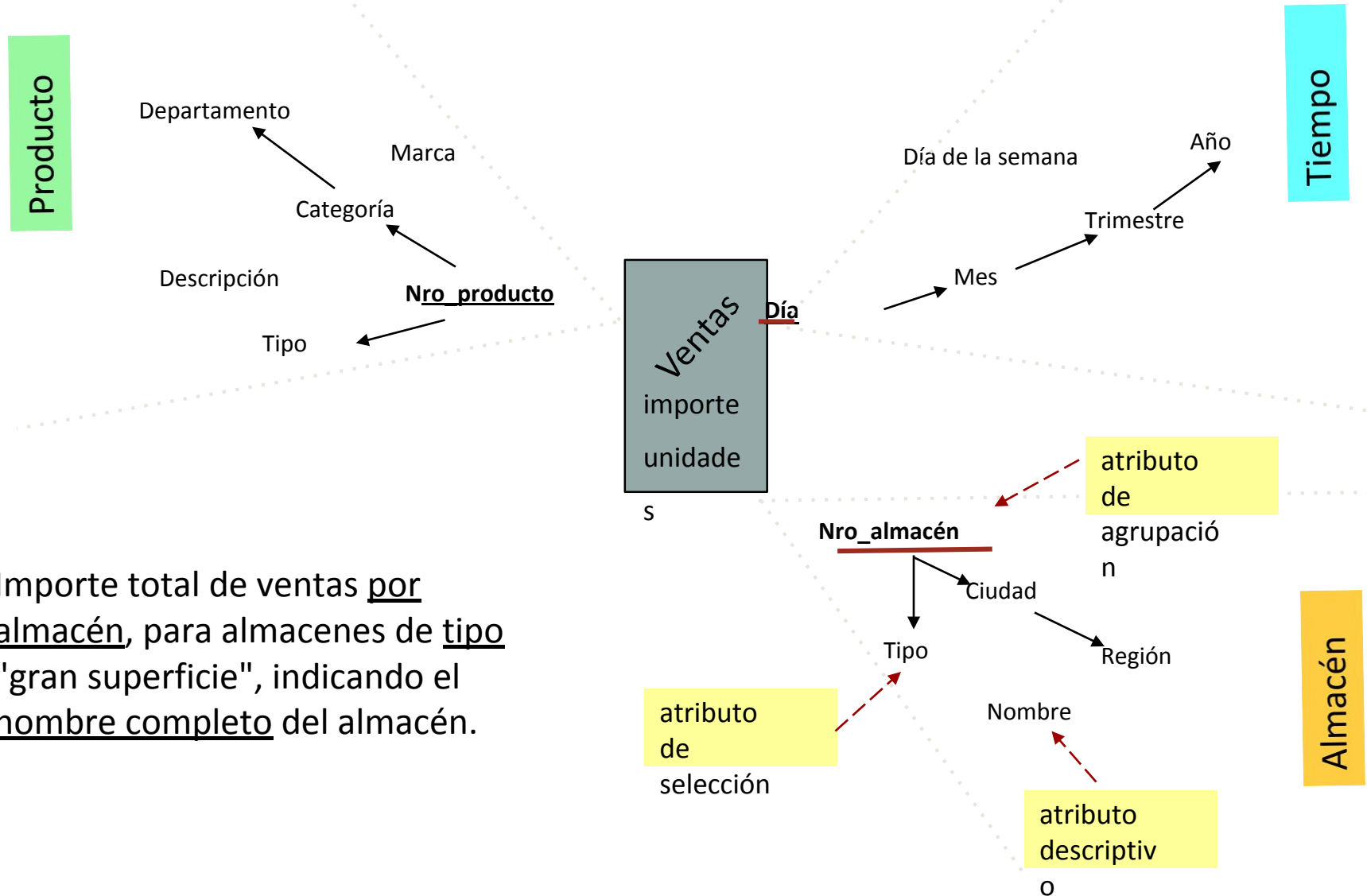


3. Explotación de un almacén de datos: herramientas OLAP

Los atributos de las dimensiones van a servir para:

- ✓ expresar condiciones que restringen el subconjunto de datos del AD que se desea consultar
- ✓ definir los parámetros de la consulta: nivel de detalle (o agregación) al que se desean presentar los datos seleccionados
- ✓ añadir información descriptiva a los elementos de la dimensión
- ✓ las jerarquías definidas entre los atributos de las dimensiones son una guía para "navegar" por los datos seleccionados, cambiando el nivel de agregación con el que son presentados

3. Explotación de un almacén de datos: herramientas OLAP



3. Explotación de un almacén de datos: herramientas OLAP

Consulta de un AD con una herramienta OLAP

- ✓ Las herramientas de OLAP presentan al usuario un 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 de base de datos) del almacén de datos.
- ✓ La herramienta OLAP genera la correspondiente consulta y la envía al gestor de consultas del sistema (sentencia SELECT en un sistema ROLAP).