



Gestión del Dato: Modelado, Data warehouse

Jose Aguilar
CEMISID, Escuela de Sistemas
Facultad de Ingeniería
Universidad de Los Andes
Mérida, Venezuela

Introducción a Data Warehouse

Analizando la información de una empresa

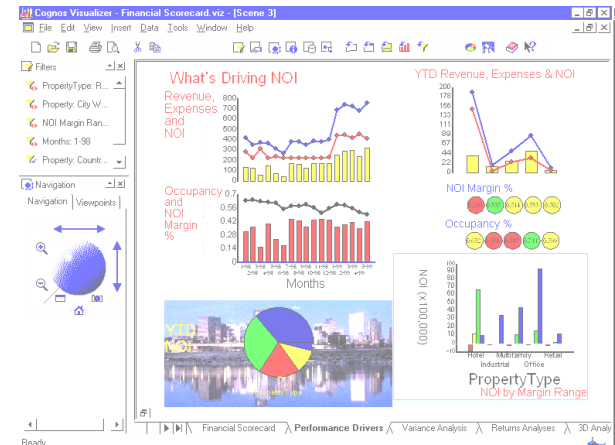
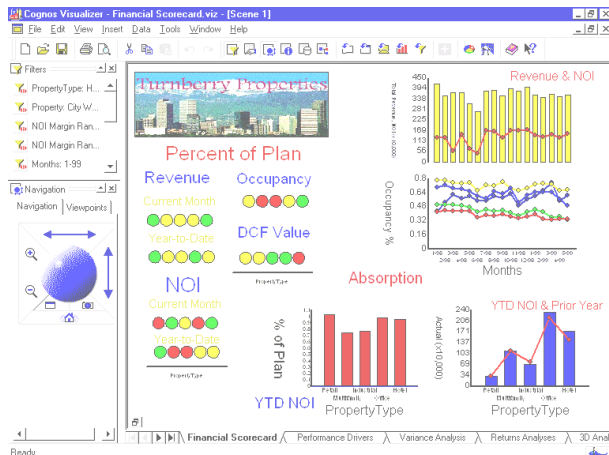
- ✓ Información periódica de las ventas
- ✓ Información del esfuerzo comercial
- ✓ Información sobre los pedidos a los proveedores

Por qué no integrarla y cruzarla para obtener:

- ✓ ¿En qué zonas se está vendiendo más de cada línea de productos?
- ✓ ¿Quiénes son los clientes más rentables?
- ✓ ¿Cuál es la relación entre el esfuerzo comercial y las operaciones cerradas?
- ✓ ¿De qué proveedores se está comprando la mayor parte de los productos vendidos ?

Introducción a Data Warehouse

- ✓ Se necesita entender no solo **QUÉ** está pasando, sino **CUÁNDO, DÓNDE, QUIÉN, CÓMO Y POR QUÉ**.
- ✓ Requerimientos de información con **OPORTUNIDAD**.
- ✓ **ESCALAR, ENRIQUECER Y COMPARTIR** a todos los usuarios en la organización



Introducción a Data Warehouse

Ventas

- Número de pedidos
- Productos pedidos
- Clientes que realizaron los pedidos

- Clientes más rentables
- Pedidos más frecuentes
- Productos más rentables
- % de nuevos clientes

Servicio al Cliente

- Datos de llamadas de nuestros clientes
- Información sobre el log de nuestra página web

- ¿Qué clientes visitan nuestra página web ?
- % pedidos realizados por nuestros canales de ventas
- ¿Qué consulta es más frecuente?

Marketing

- Número de campañas realizadas y características de cada una

- % éxito de las campañas
- ¿Qué tipo de clientes han respondido mejor a cada una de las campañas realizadas ?

Distribución

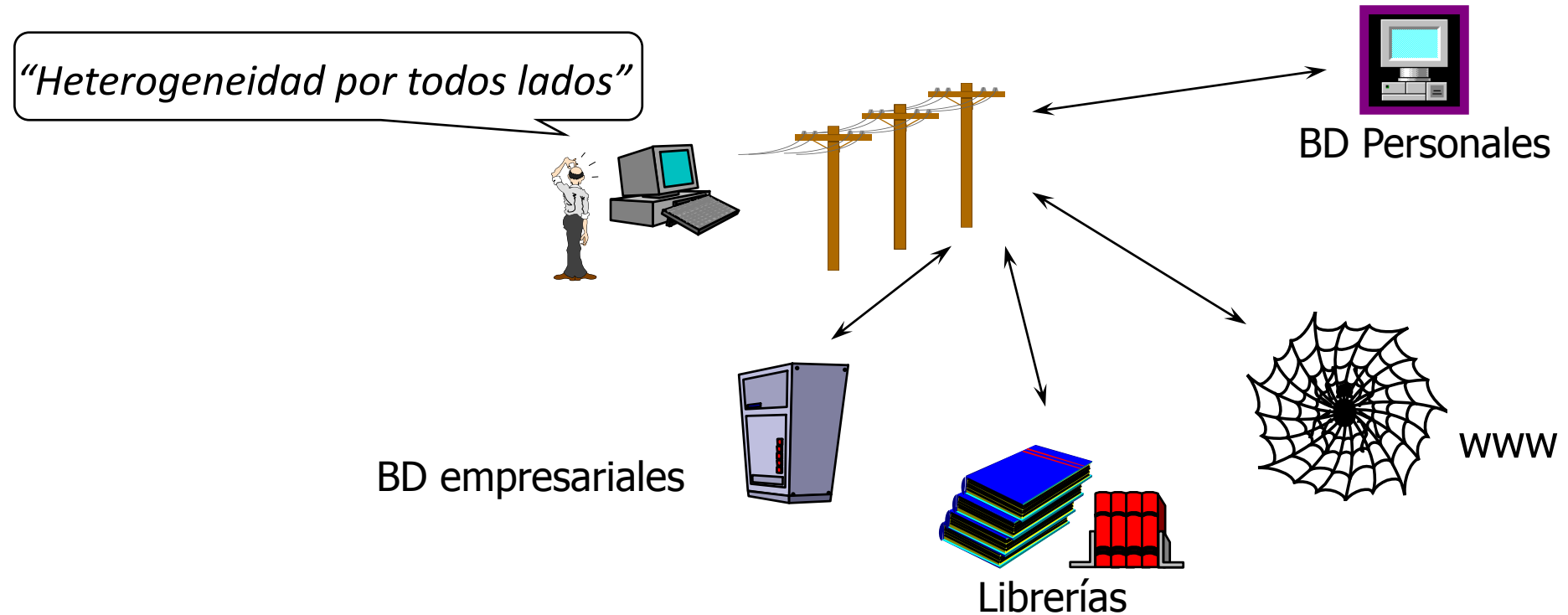
- Productos que salen diariamente del almacén
- Tiempo teórico de entrega

- Número de pedidos retrasados
- Distribuidor que tiene el mayor número de retrasos
- Tiempo medio de entrega

DATO

INFORMACION

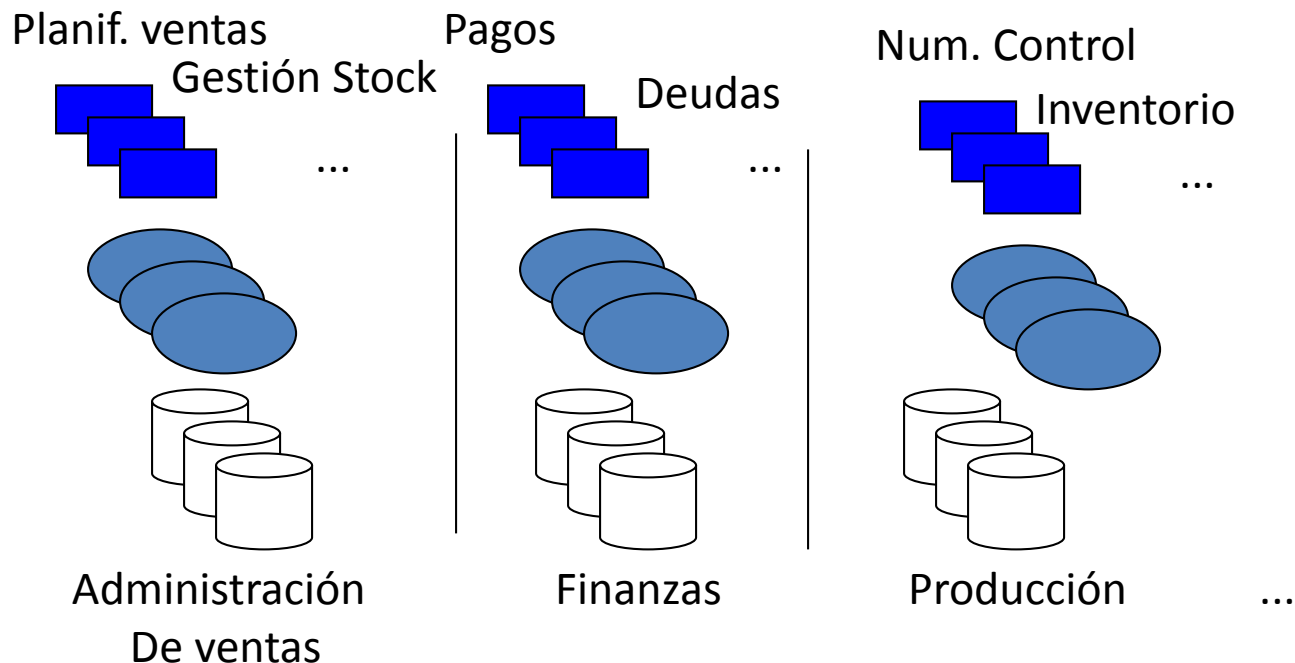
Problemas: Heterogeneidad de las fuentes de Información



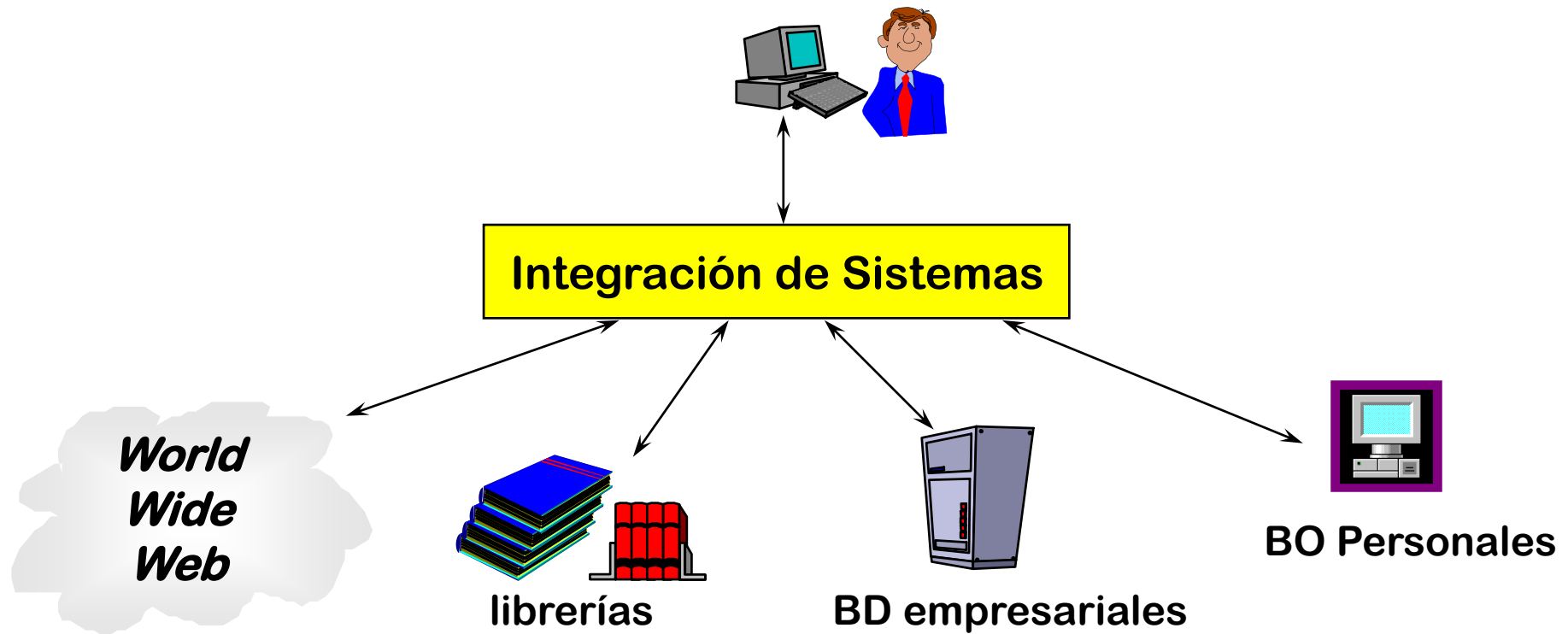
- Diferentes interfaces
- Diferentes representaciones de datos
- Información duplicada e inconsistente

Problema: Gestión de datos en grandes empresas

- Fragmentación vertical de los sistemas de información
- Desarrollo de las aplicaciones guiadas por los sistemas operativos



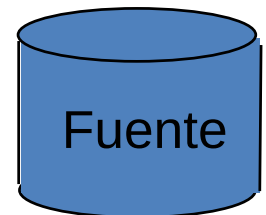
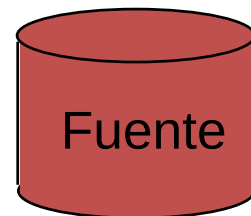
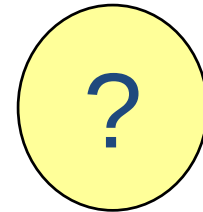
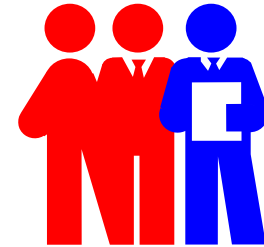
Objetivo: Unificar Acceso a los Datos



- Recopilar y combinar la información
- Proporcionar visión integrada, en una interfaz de usuario uniforme
- Soportar el intercambio

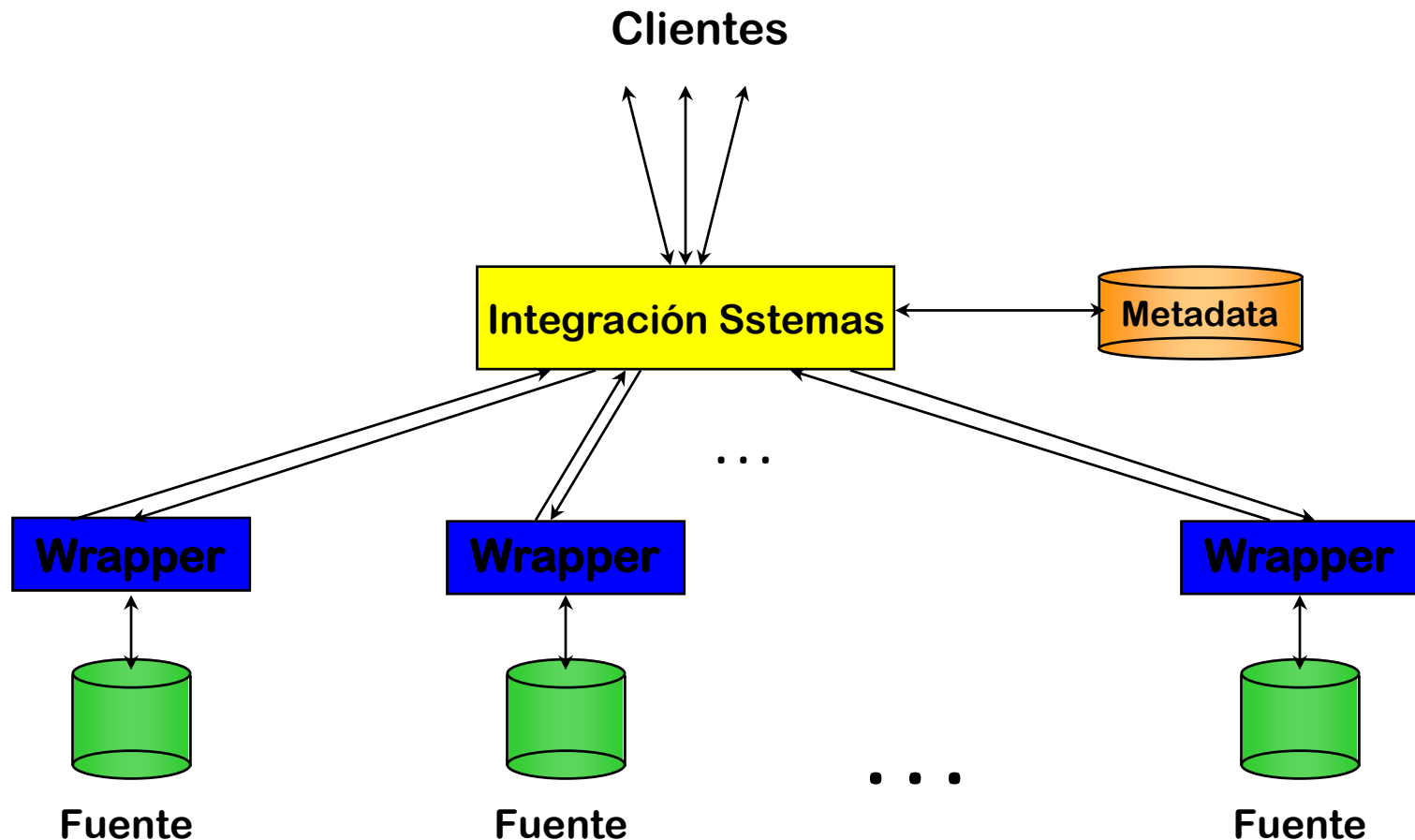
Objetivo: Unificar Acceso a los Datos

- Dos enfoques:
 - Guiado por la consulta (perezoso)
 - Warehouse (ansioso)



Enfoque tradicional

Guiado por la consulta (perezoso, on-demand)

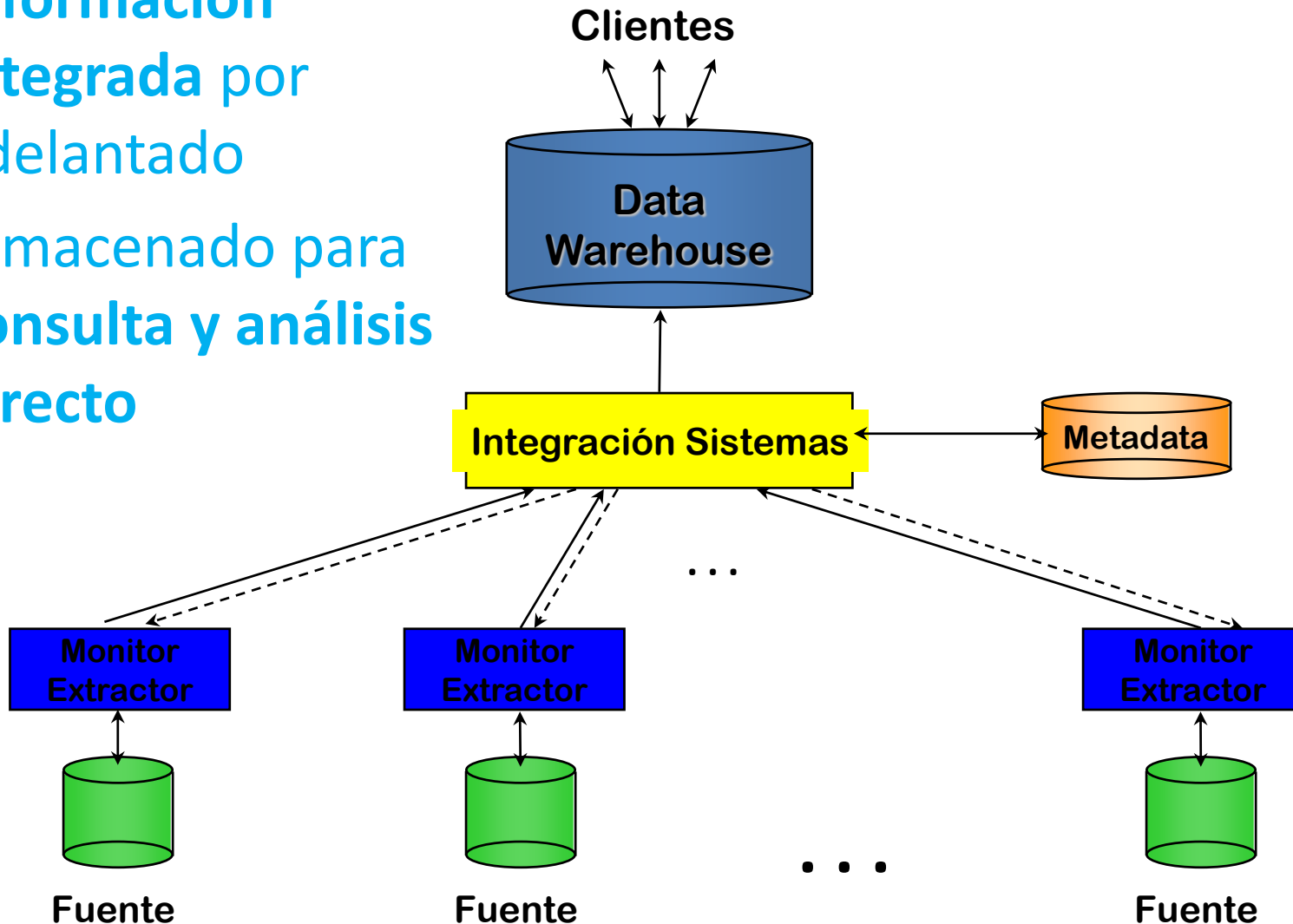


Problema del Enfoque tradicional

- ♦ El **retraso** en el procesamiento de consultas
- ♦ Fuentes de información **lentas o no disponibles**
 - ♦ Filtrado e integración son complejas
 - ♦ Ineficiente y potencialmente costoso para las consultas frecuentes
- ♦ **Compite con el procesamiento local** en el sitio fuente

Enfoque Warehousing

- Información integrada por adelantado
- Almacenado para consulta y análisis directo



Ventaja Enfoque Warehousing

- **Alto rendimiento de la consulta**
 - Pero no necesariamente la información más actualizada
- **No interfiere con el procesamiento local en el sitio origen**
 - Las consultas complejas en warehouse
 - OLTP en las fuentes de información
- **Información copiada en el almacén**
 - Se puede modificar, anotar, resumir, reestructurar, etc.
 - Puede almacenar información histórica
 - Seguridad, sin auditoría
- **Usada en la industria**

Concepto Data Warehousing

Es un **gran almacén de datos** para consultar

Es un **repositorio de datos** de muy **fácil acceso**, alimentado de **numerosas fuentes**, transformadas en grupos de información sobre **temas específicos de negocios**, para permitir **nuevas consultas, análisis, toma de decisiones**.

Se trata, de un **expediente completo de una organización**, más allá de la información transaccional y operacional, almacenado en una base de datos diseñada para favorecer el **análisis y la divulgación eficiente de datos**.

Tiene **gran capacidad de almacenamiento**, pues los datos pueden ser de grandes periodos de tiempo.



Concepto Data Warehousing (Bodegas de Datos)

Conjunto de **datos integrados y orientados a una materia, varían con el tiempo, soportan el proceso de toma de decisiones de la administración** y esta orientada al manejo de **grandes volúmenes** de datos provenientes de **diversas fuentes** o diversos tipos.

Estos datos **cubren largos períodos de tiempo**, lo que trae consigo que se tengan diferentes esquemas de los datos. **Previo a su utilización**, se debe aplicar procesos de análisis, selección y transferencia de datos seleccionados desde diversas fuentes.

Concepto Data Warehousing (Bodegas de Datos)

Colección de datos que verifican las siguientes propiedades:

- Está orientado a un tema
- Datos integrados
- No volátiles
- Variante en el tiempo

surgieron como una herramienta de soporte para la toma de decisiones a nivel gerencial

Orientado hacia temas



Los datos se almacenan y agrupan por temas de interés

Se agrupa por **temas orientados a la organización**, tales como :

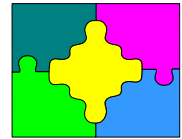
- Clientes
- Productos
- venta

en lugar de las transacciones individuales.

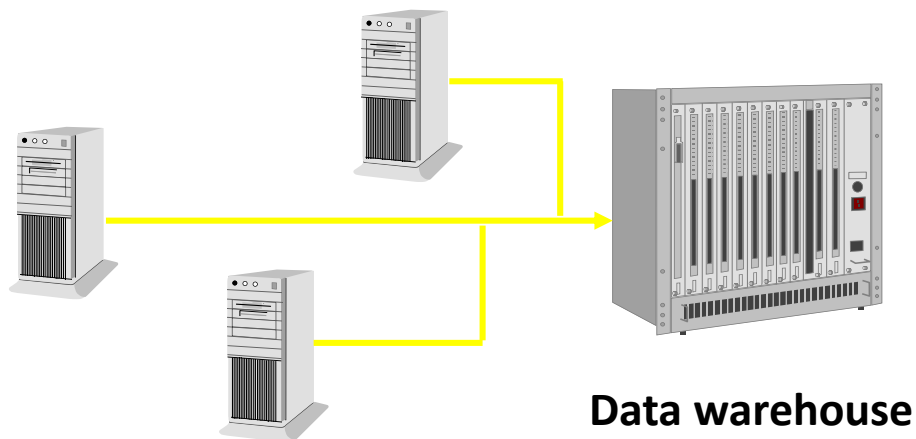
La normalización no es relevante

- **Centrándose en el modelado y análisis de datos** para la toma de decisiones, no en las operaciones diarias o el procesamiento de transacciones.
- Proporcionar una **visión sencilla y concisa en torno a cuestiones temáticas particulares** mediante la exclusión de los datos que no son útiles en el proceso de apoyo a la decisión.

Integración de Datos

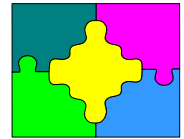


- El almacén de datos **integra datos** que provienen de varias fuentes. Parte de una base de datos operacional, y mediante un proceso de **carga de datos** hace el Datawarehouse.
- El proceso de carga es lo más complicado por problemas de **preparación de los datos**.



Los datos en el almacén deben ser:

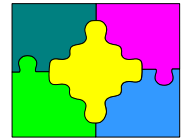
- Limpios
- Validados
- Adecuadamente integrados



Integración de Datos

- Integración de **múltiples fuentes de datos heterogéneas**,
 - bases de datos relacionales, archivos planos, registros de transacciones en línea
- Se aplican **técnicas de limpieza y de integración de datos**.
- Garantiza la **coherencia** en los convenios de denominación, estructuras de codificación, medidas, etc., entre las diferentes fuentes de datos
 - Por ejemplo, el precio del hotel: moneda, impuestos, desayuno, etc.
- Cuando los datos se mueven a la bodega, **se convierten** al estándar usado.

Ejemplo Integración de Datos



Sistema de Cuenta de cheques

Jane Doe (name)
Female (gender)
Bounced check #145 on 1/5/95
Opened account 1994

← *Datos operacionales*

Sistema de cuentas de ahorro

Jane Doe
F (gender)
Opened account 1992

cliente

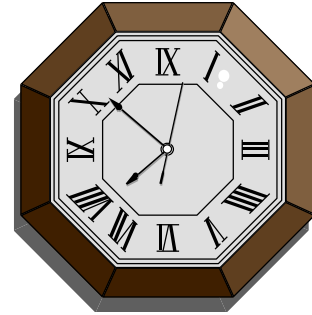
Jane Doe
Female
Bounced check #145
Married
Owns 25 Shares Exxon
Customer since 1992

Sistema de inversiones de clientes

Jane Doe
Owns 25 Shares Exxon
Opened account 1995

↑ *data warehouse*

... variante en el tiempo ...



- Todos los datos en el almacén de datos tienen una **marca de tiempo** en el momento de entrada en el almacén o cuando se usan en el almacén.
- Esta grabación cronológica de datos ofrece posibilidades **históricas y análisis de tendencias**.
- Normalmente, en las BD operativas **los datos se sobrescribe**, ya que los valores anteriores no son de interés.

TIEMPO

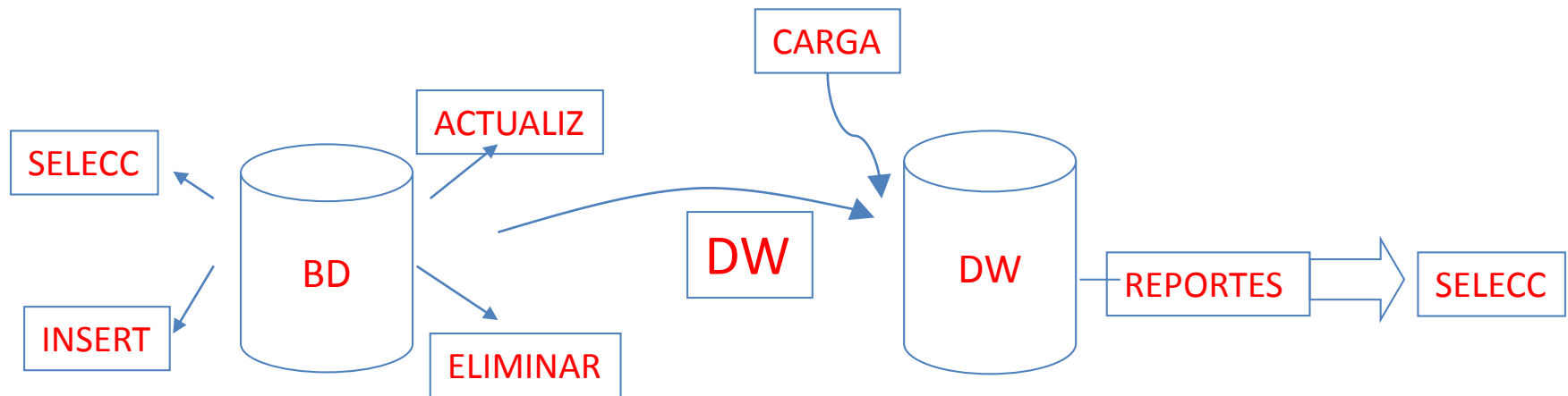
id_tiempo

* periodo

No volátiles



Son estables, una vez almacenados los datos **no se modifican**.



Datos actúan como un recurso estable para informes coherentes y análisis comparativo.

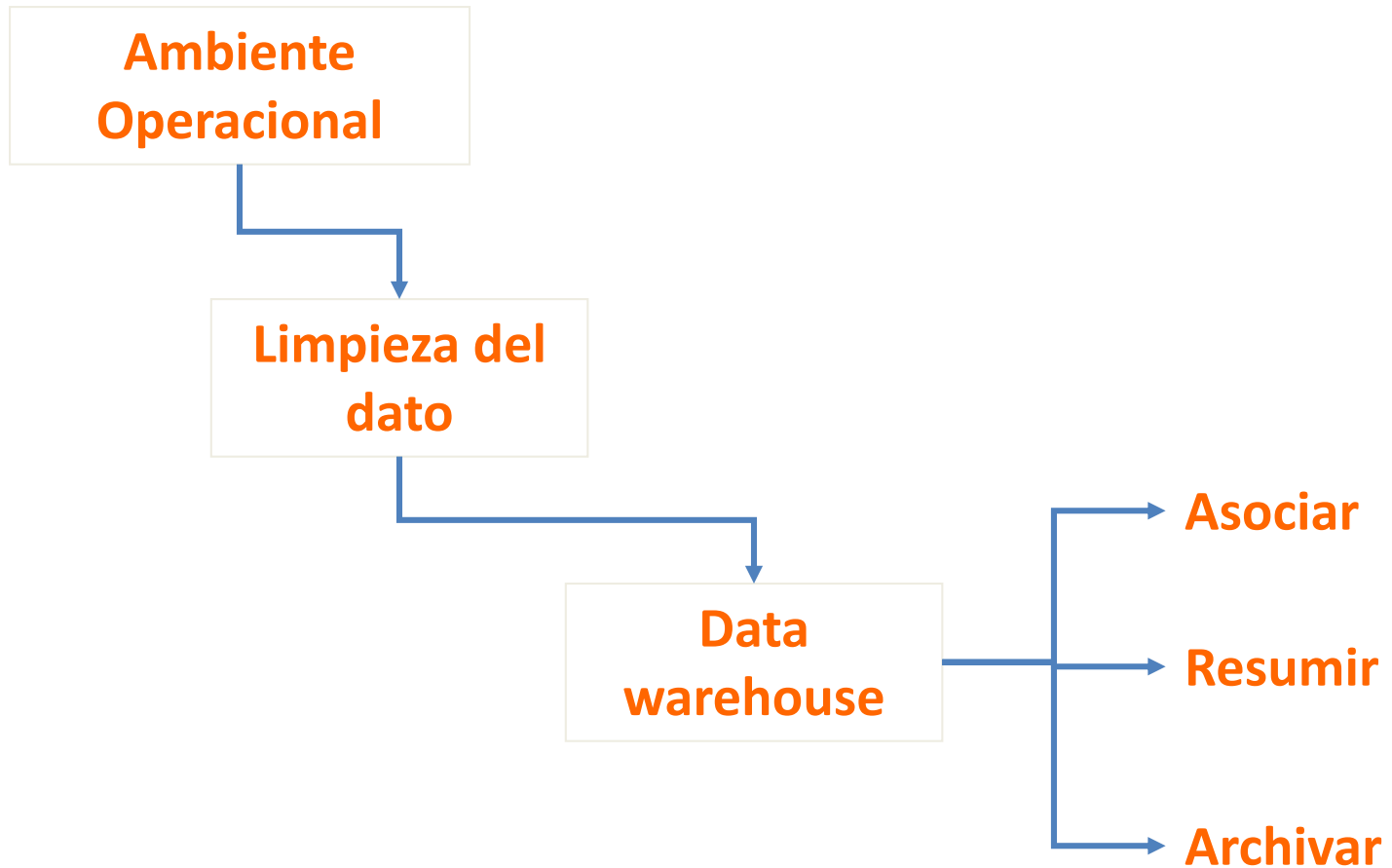
Por el contrario, los datos operativos se actualizan (insertar, eliminar, modificar, etc.)

No volátiles



- **Físicamente separado** de los datos del entorno operacional.
- **Actualización de los datos no se produce** en el entorno de almacén de datos.
- **No requiere el procesamiento de transacciones**, ni mecanismos de recuperación y control de concurrencia
 - Requiere sólo dos operaciones en el Acceso a los datos:
carga inicial de datos y el acceso a los datos.

Flujo del Dato



RESUMEN DIFERENCIAS

BD OPERACIONAL

- Datos operacionales
- Orientado a aplicaciones
- Datos Actuales
- Datos Detallados
- Datos en continuo cambio

DATAWAREHOUSE

- Datos de negocio
- Orientado al sujeto
- Actuales + Histórico
- Datos Resumidos
- Datos Estables

Por qué Data Warehouse?

Diferentes funciones y datos:

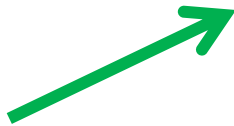
- **Datos que faltan:** apoyo a las decisiones requiere datos históricos que BDs operacionales no suelen tener
- **Consolidación de datos:** Se requiere de la consolidación (agregación, resumen) de los datos de fuentes heterogéneas
- **Calidad de los datos:** las diferentes fuentes suelen utilizar representaciones de datos inconsistentes, códigos y formatos que deben ser conciliados, etc.

Funcionamiento de Data warehouse

Tres funciones esenciales:

1. Recopilación de los datos desde Bases de Datos
2. Gestión de los datos en el almacenamiento
3. Análisis de datos para toma de decisiones

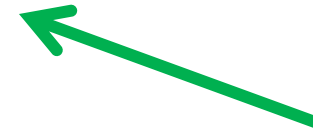
Sistemas de
Soporte a
Decisiones
(DSS)



herramientas para
hacer consultas o
informes



Sistemas de
información
ejecutiva (EIS)



Elementos que integran un almacén de datos

Metadatos

"datos acerca de los datos".

Su función es recoger la descripción de la estructura del almacén de datos:

Tablas

Columnas en tablas

Jerarquías y Dimensiones de datos

Relaciones entre tablas

Entidades y Relaciones

Elementos que integran un almacén de datos

Metadato también cuenta con

- **Meta-datos de los datos Operacionales**
 - Historia de los datos (migraciones, transformaciones),
 - Estado de los datos (activo, archivados, o limpiado),
 - Información de seguimiento (estadísticas de uso de almacén, informes de errores, trazas de auditoría)
- **Los algoritmos utilizados para resumir**
- **El mapeo del entorno operativo al almacén de datos**
- **Los datos relacionados con el rendimiento del sistema**
- ...

12 reglas de un Data Warehouse

1. Data Warehouse y los entornos operativos estan **separados**
2. Los datos se **integran**
3. Contiene **datos históricos** durante un largo período de tiempo
4. Datos **son capturados en un punto dado en el tiempo**
5. Datos son **orientados a temas**

12 reglas de un Data Warehouse

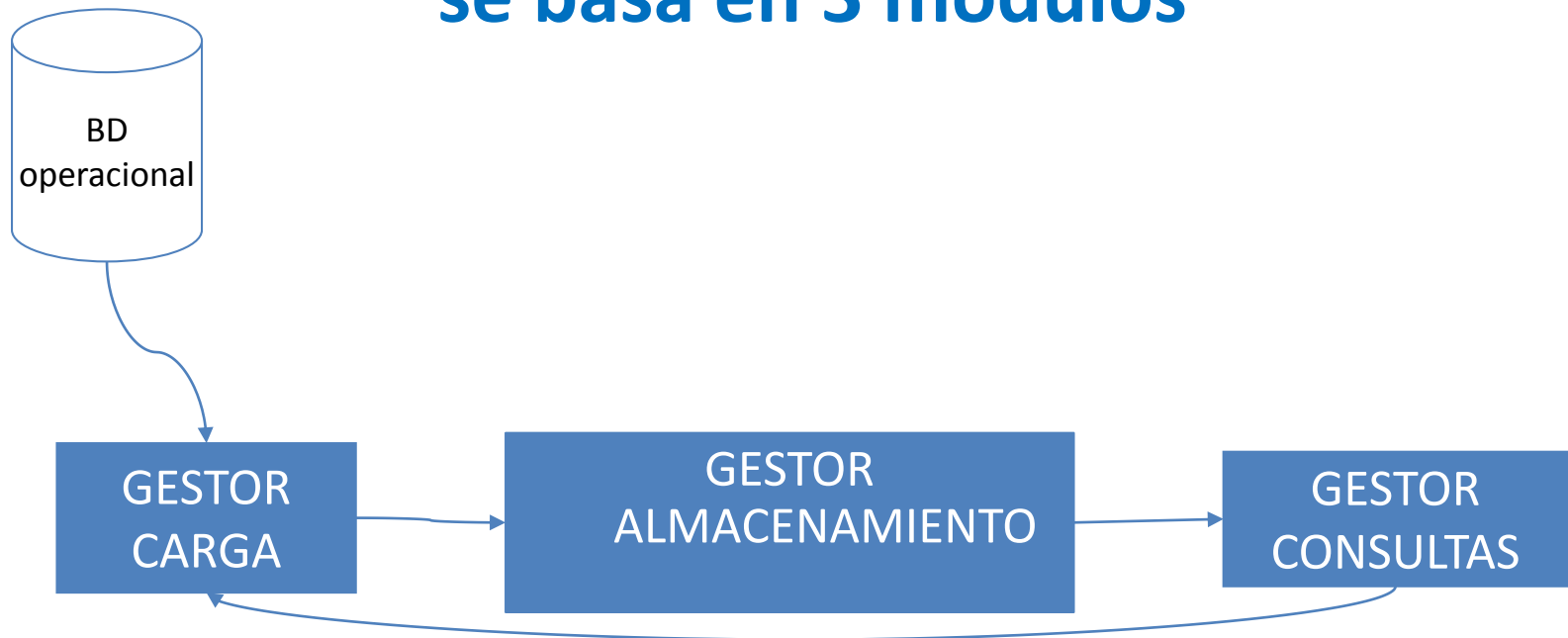
6. Principalmente **es usado sólo de lectura**, con **actualizaciones periódicas por lotes**
7. Ciclo de Vida de Desarrollo sigue **enfoque dirigido por los datos** frente al enfoque tradicional basado en procesos
8. Datos contiene **varios niveles de detalle**: actual, viejo, ligeramente resumido, muy resumido

12 reglas de un Data Warehouse

9. Medio Ambiente se caracteriza por **transacciones de sólo lectura a conjuntos de datos muy grandes**
10. Sistema **rastrea las fuentes de los datos**, las transformaciones y el almacenamiento
11. **Los metadatos** son un componente crítico
12. Contiene mecanismos para el **uso óptimo de los recursos**

ARQUITECTURA PLANA DW

se basa en 3 módulos



GESTOR DE CARGA

- Permite hacer la carga.
- Dificultades:
 - La integración de los datos
 - La elección del momento de la carga
 - Minimizar el tiempo de carga
 - Buen diccionario de datos o METADATA (para evitar cometer errores en la carga)
 - Diseño de procedimientos SQL

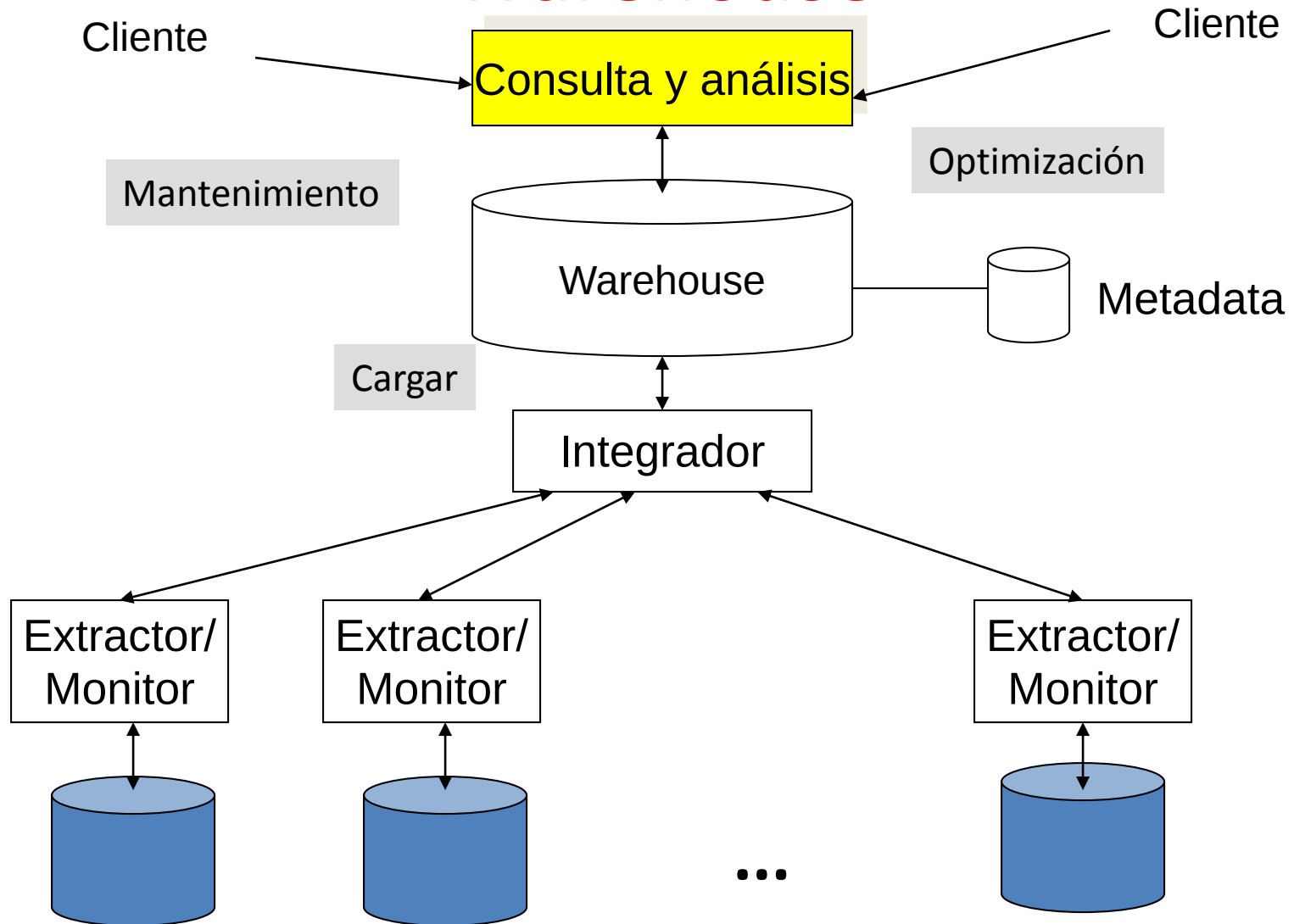
GESTOR DE ALMACENAMIENTO

- Se encarga del almacenamiento, de la estructura,....
- Se basa en modelos multidimensionales

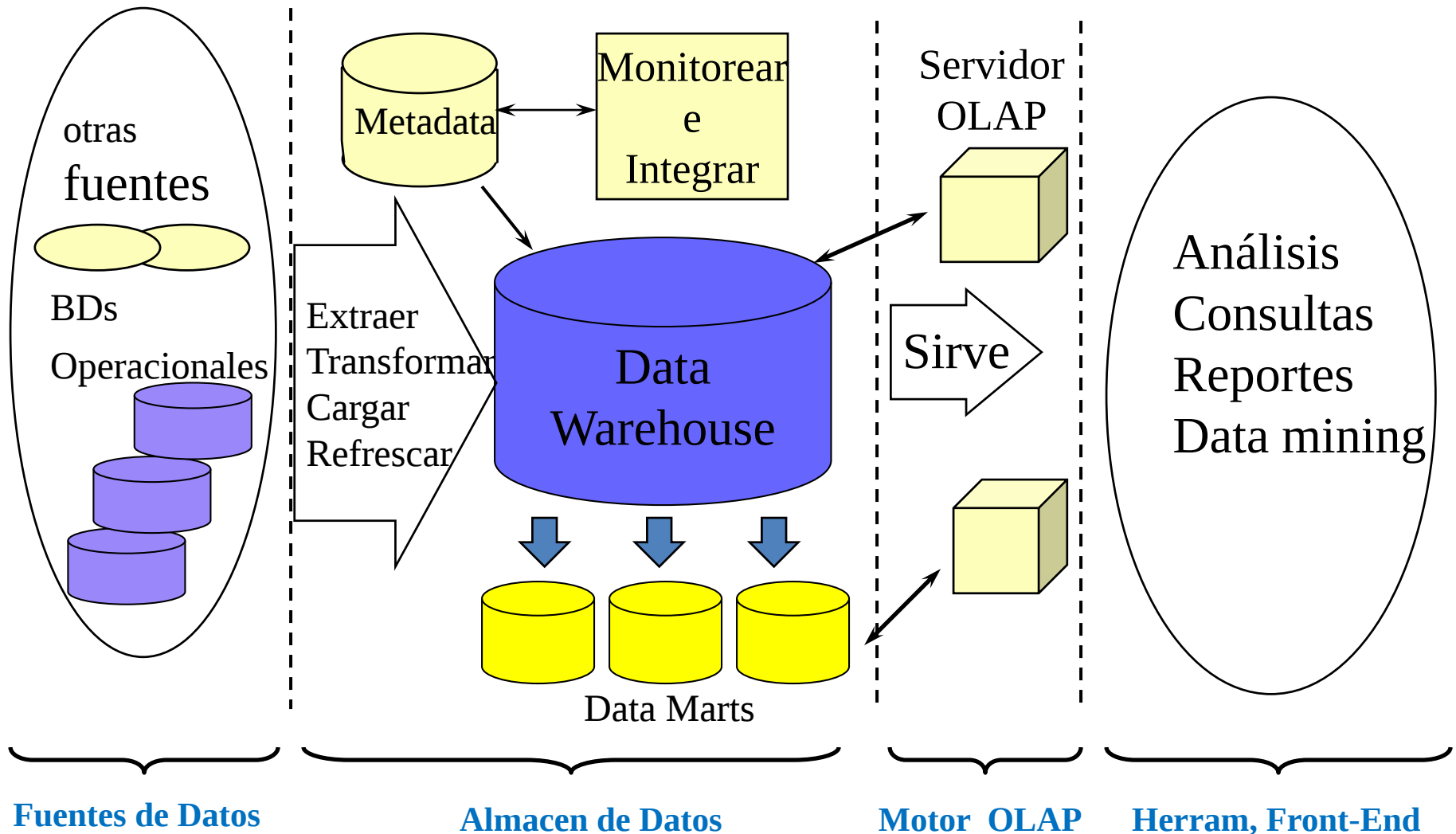
GESTOR DE CONSULTAS

- Las consultas se hacen sobre la tabla FACT.
- También define perfiles de consultas, pues ellas son diferentes dependiendo del usuario y sus necesidades.

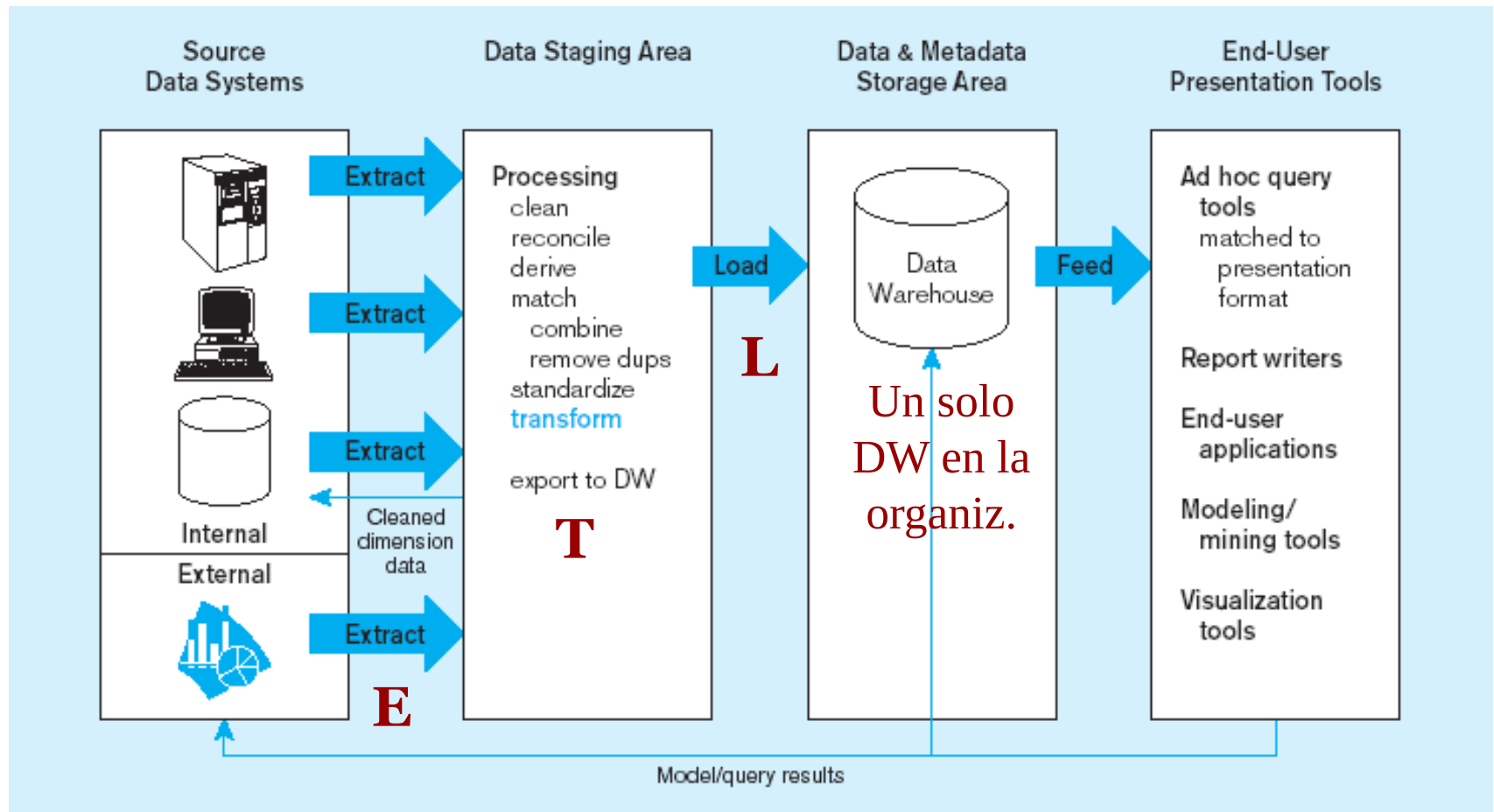
Arquitectura genérica Data Warehouse



Arquitectura varios niveles



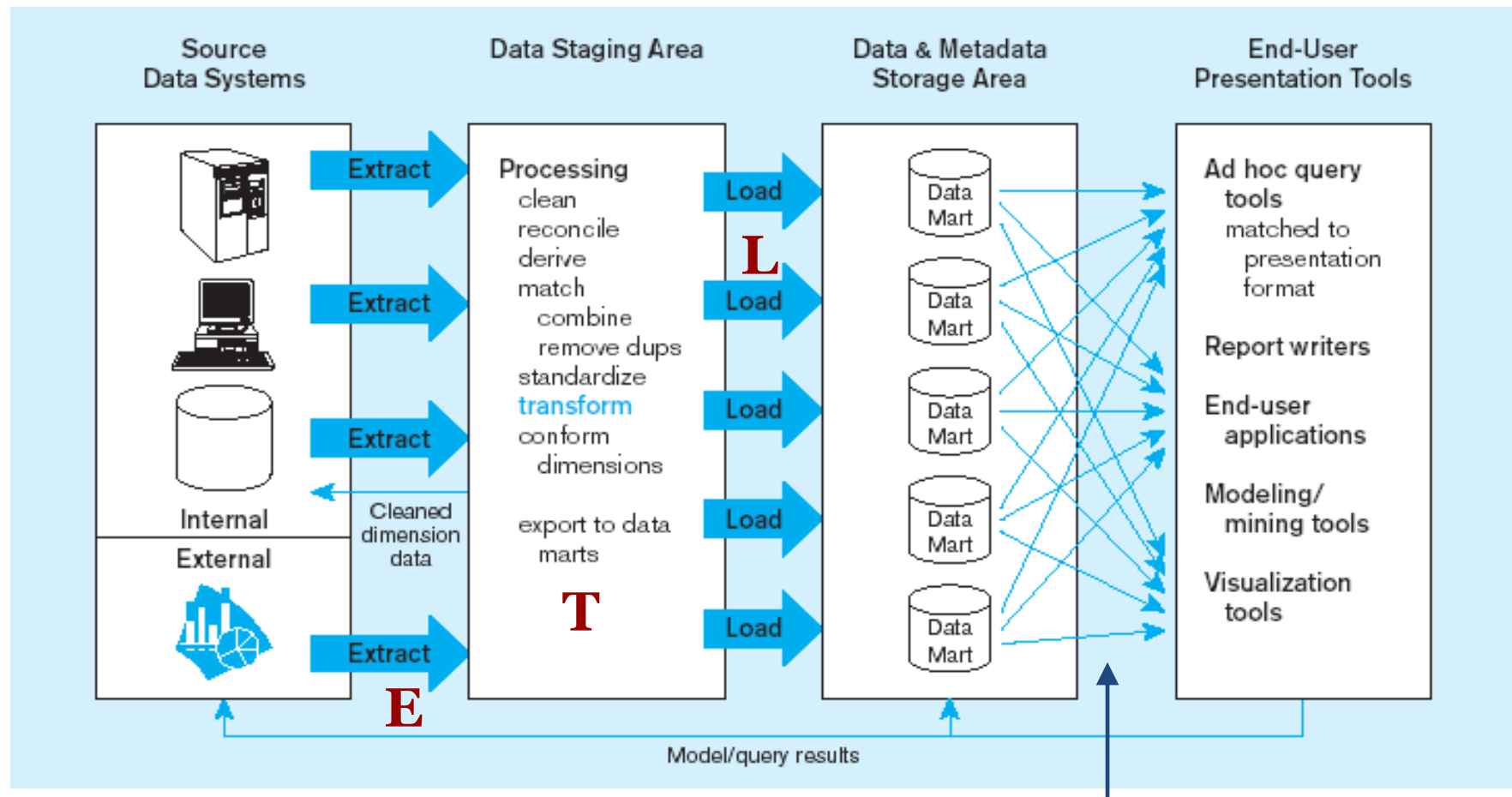
Genérica arquitectura de niveles de DW



Centralizado

Extracción de datos periódica
Datos no están completamente al día

Arquitectura independientes Data mart

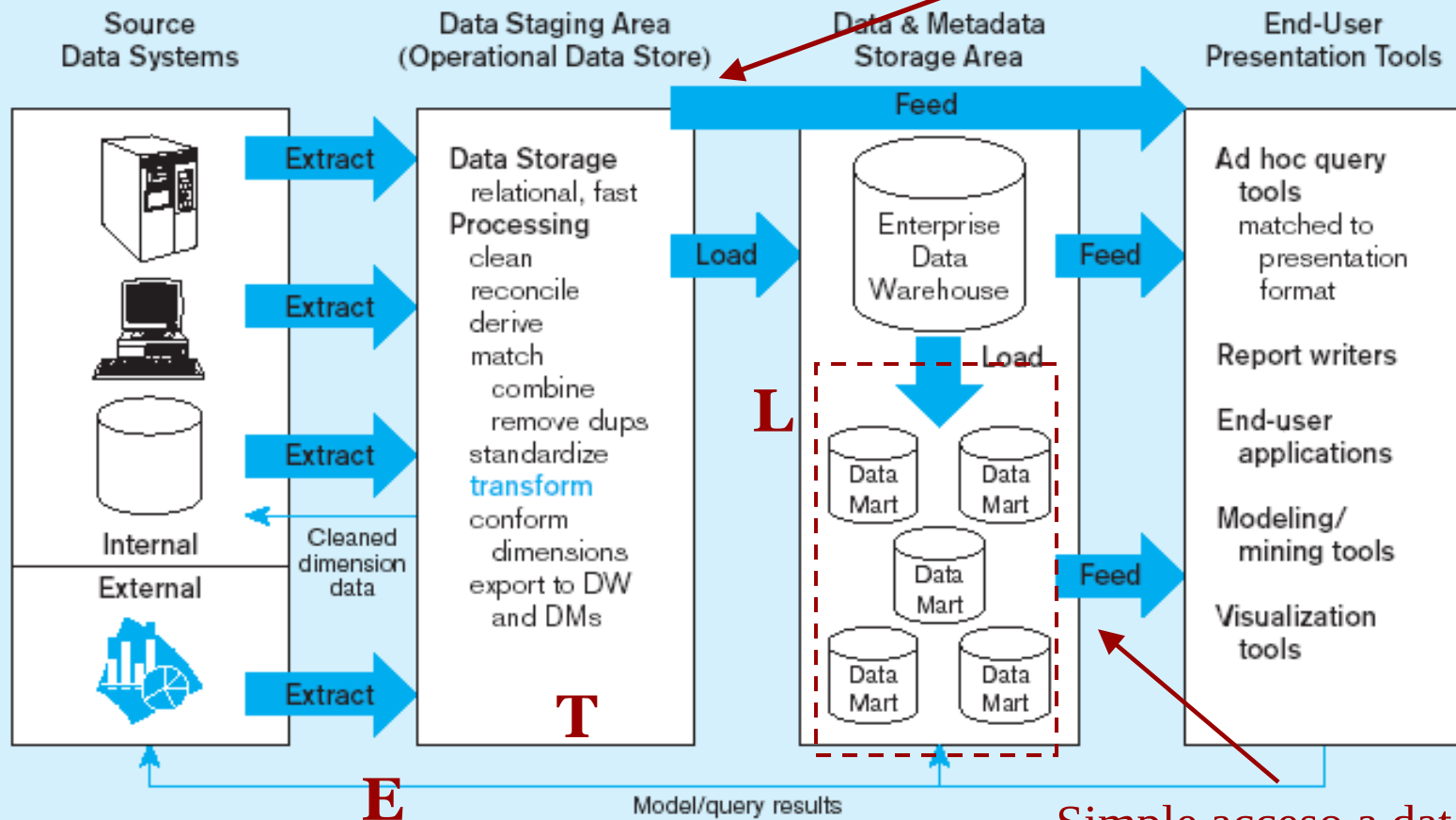


Separado ETL para c/data mart

Acceso datos por
*multiple*s data marts

Arquitectura Híbrida

Se pueden obtener
actuales datos



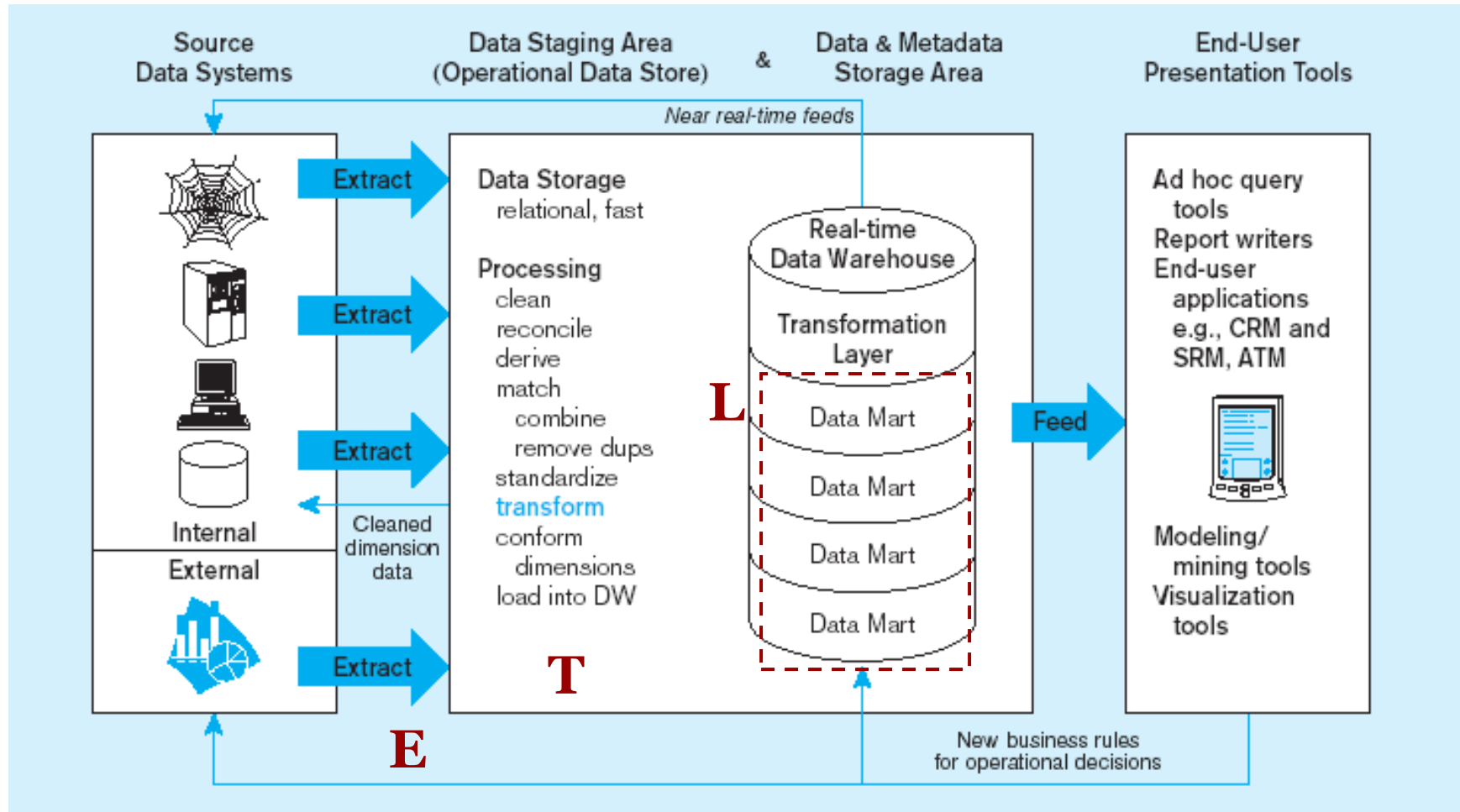
Simple acceso a datos

Un solo ETL para
enterprise data warehouse
(EDW)

Data marts *dependientes*
del EDW

Federado

Data mart Lógico y DW-RT



ETL RT para
Data Warehouse

Data marts no separado de BDs, vistas lógicas de los datos de data warehouse
➔ Fácil crear nuevos data marts



UNIVERSIDAD
DE LOS ANDES
MÉRIDA VENEZUELA

Modelado de Datos

Jose Aguilar
CEMISID, Escuela de Sistemas
Facultad de Ingeniería
Universidad de Los Andes
Mérida, Venezuela

Consideraciones para el Diseño Data warehouse

Para abordar un proyecto de data warehouse es necesario hacer un estudio de algunos temas generales de la organización:

- **Situación actual:** Cualquier solución propuesta de data warehouse debe estar muy orientada por las necesidades del negocio, debe ser compatible con la arquitectura técnica existente y planeada de la compañía.
- **Tipo y características del negocio:** Tener el conocimiento exacto sobre el tipo de negocios de la organización y el soporte que representa la información dentro de todo su proceso de toma de decisiones.

Consideraciones para el Diseño Data warehouse

Para abordar un proyecto de data warehouse es necesario hacer un estudio de algunos temas generales de la organización:

- **Entorno técnico:** hardware (servidores, redes,...) así como aplicaciones y herramientas. Se dará énfasis a los Sistemas de Soporte a Decisiones (DSS).
- **Expectativas de los usuarios:** Es una forma de vida de las organizaciones y como tal, tiene que contar con el apoyo de todos los usuarios y su convencimiento sobre su bondad.

Modelos dimensionales

Es una técnica de **diseño lógico** comúnmente utilizada para Data Warehouses, que busca presentar los datos en una arquitectura estándar y permita una **alta performance de acceso** a los usuarios finales.

El modelo se basa en **esquemas estrella**, conformados por **Tablas de Hechos** y **Tablas Dimensionales** (p.ej. cubos).

Modelos dimensionales

- Un **modelo relacional desnormalizado**
 - Compuesto por tablas con atributos
 - Las relaciones son definidas por claves nuevas y claves externas
- Organizado para **la comprensibilidad y facilidad de presentación** de informes en lugar de facilitar la actualización
- Consultado y mantenido por herramientas especiales de **gestión analítica**

Diseño de Esquemas



- **Los datos se organizan por temas importantes:**

Los clientes, los productos, las ventas, ...

- **Tema = datos + dimensiones**

1. **Recopilación de datos útiles sobre un tema**

Ejemplo: ventas

2. **Sintetizar una visión única de los temas a analizar**

Ejemplo: Ventas (producto, período, tienda, número)

3. **Detallar la vista según dimensiones**

Ejemplo:

Productos (IDprod, descripción, color, tamaño ...)

Tiendas (IDmag, nombre, ciudad, departamento, país)

Periodo (IDper, año, trimestre, mes, día)

Diseño de Esquemas

Los tipos de Esquema

- En estrella
- Constelación
- Copo de nieve

1. Aislar Datos a tener en cuenta

- Esquemas de las Tablas de hechos

2. Definir las dimensiones

- Ejes de análisis

3. Estandarizar dimensiones

- Dividir en varias tablas unidas por referencias

4. Integrar todo

- Varias tablas de hechos comparten algunas tablas de dimensiones (constelación de la estrella)



Esquema en estrella

- Modelado relacional actual no satisface las necesidades actuales
- Representaciones de datos multidimensionales
- Optimizar las operaciones de consulta de datos en lugar de las operaciones de actualización de datos
- Los datos no son usados para realizar transacciones del negocio.
- Los datos pueden obtenerse mediante cálculos o agregaciones.

Esquema en estrella

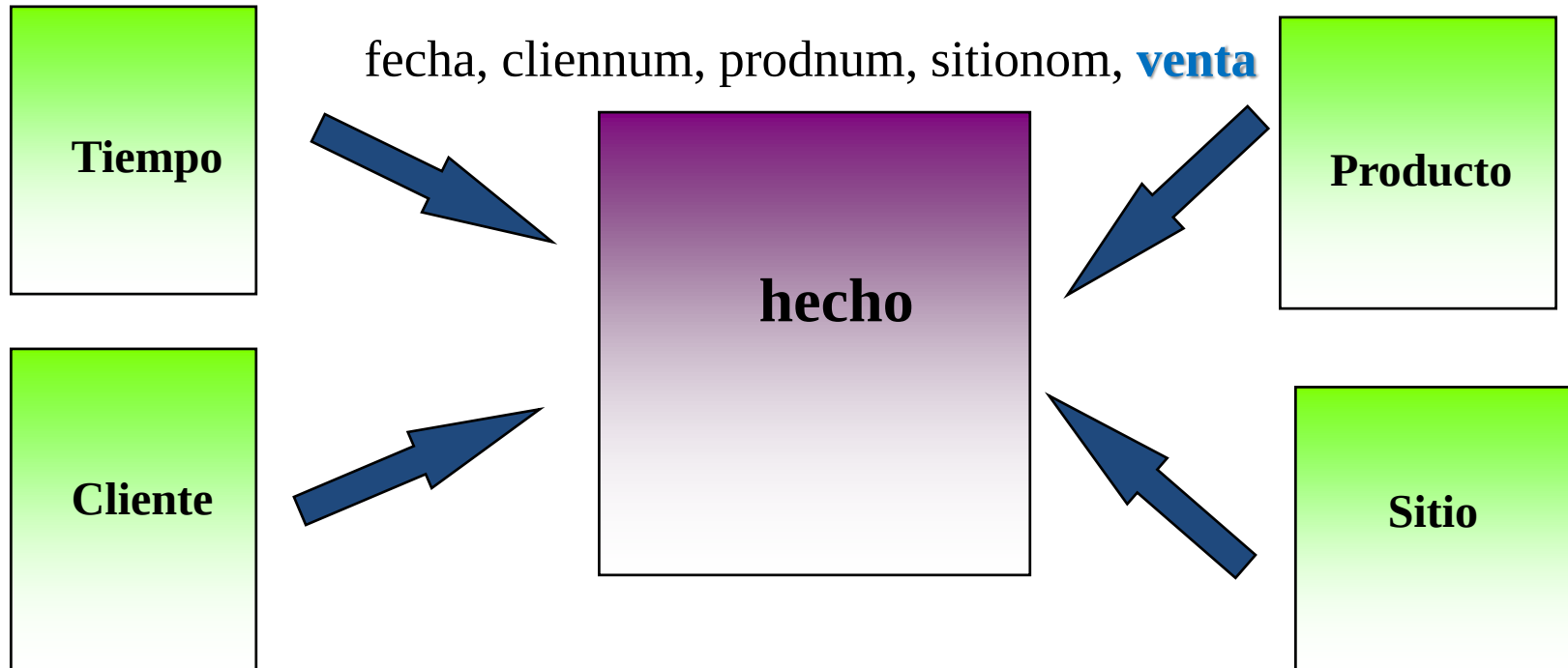
- El modelo estrella es una representación de una vista de la organización.
 - Ventas
 - Mercadeo
- El modelo estrella consolida hechos en relación a dimensiones o filtros.
- Esquema en estrella
 - Hecho rodeado de varias dimensiones (4-15)
 - Las dimensiones se de-normalizan
- Una tabla de hechos en el medio conectado a un conjunto de tablas de dimensiones

Esquema en estrella: Componentes

- Datos (hechos)
- Dimensiones
- Atributos
- Jerarquías de atributos

Esquema en estrella

- Una sola tabla de hechos y para cada dimensión una tabla de dimensiones
- No captura jerarquías directamente



Esquema en estrella

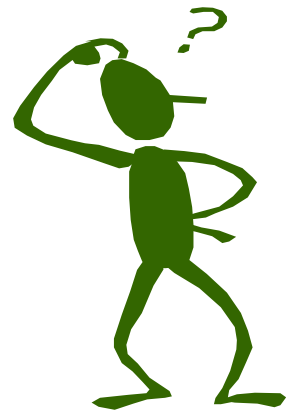
Existe una tabla llamada FACT (Hechos) y unas tablas llamadas dimensiones o tablas dimensionales.

- Entre la tabla FACT y las tablas dimensionales suele haber relaciones 1:N
- Este modelo tiene forma de estrella, por eso se denomina MODELO ESTRELLA

Tablas de hechos

Contienen los hechos que serán utilizados por los analistas para apoyar el proceso de toma de decisiones.

- Toma los datos desde los sistemas transaccionales
- Realiza las transformaciones requeridas en los datos
- Enlaza dimensiones a través de sus claves



Esquema en estrella: Hechos

- Mediciones numéricas (valores) que representan un aspecto del negocio o actividad específica
- Almacenado en una tabla de hechos en el centro del esquema de estrella
- Contiene hechos caracterizados a través de sus dimensiones
- Se pueden calcular o derivar en tiempo de ejecución
- Actualizado periódicamente con los datos de las bases de datos operacionales

Esquema en estrella: Tabla de Hechos

Tabla central

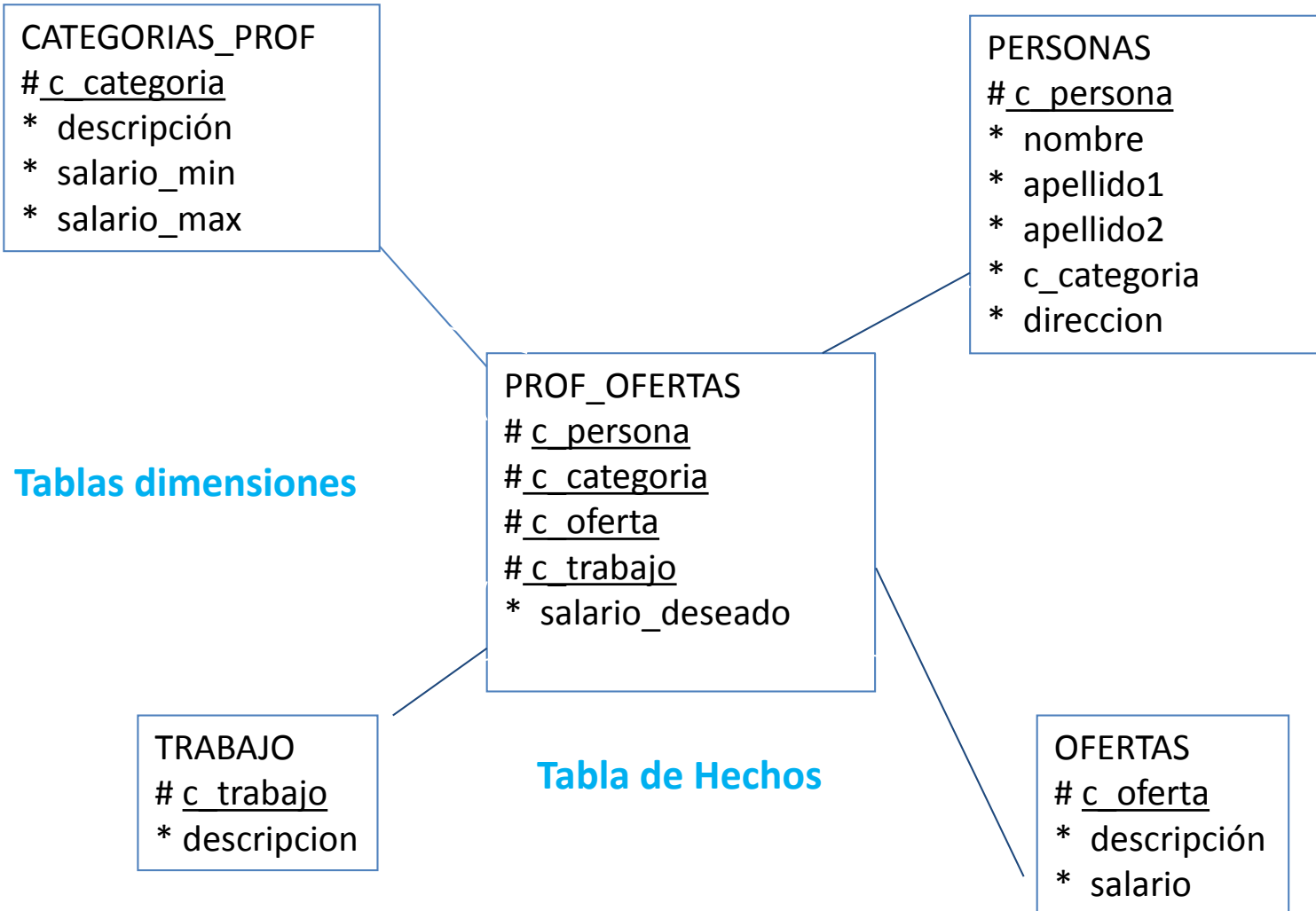
- Representa un proceso o reporta el entorno que es de valor para la organización
- Especifica exactamente lo que representa.
- Por lo general, corresponden a una entidad asociativa en el modelo ER
- Guarda Medidas de interés del negocio
- Varían bastante sus datos

Esquema en estrella: Tabla de Hechos

Tabla central

- Gran número de filas (millones a un mil millones)
- Algunas columnas como máximo
- Acceso por dimensiones: Enlaces directos a las dimensiones
 - Contiene dos o más claves foráneas
- Clave principal de varias partes

Esquema en estrella



Tablas de dimensión

Definen como están los datos organizados lógicamente y proveen el medio para analizar el contexto organizacional.

- Toma los datos desde los sistemas transaccionales
- Depura los valores de los atributos para incorporarlos al modelo dimensional
- Mantiene las claves
- Mantiene la tabla de referencias cruzadas



Esquema en estrella: Dimensiones

- Características cualitativas que proveen perspectivas adicionales a un hecho
- Las dimensiones se **almacenan en tablas de dimensiones**
- Dimensiones comunes:
períodos de tiempo, áreas geográficas (mercados, ciudades), productos, clientes, vendedores, etc.
- **Típicamente contienen atributos para consultas**

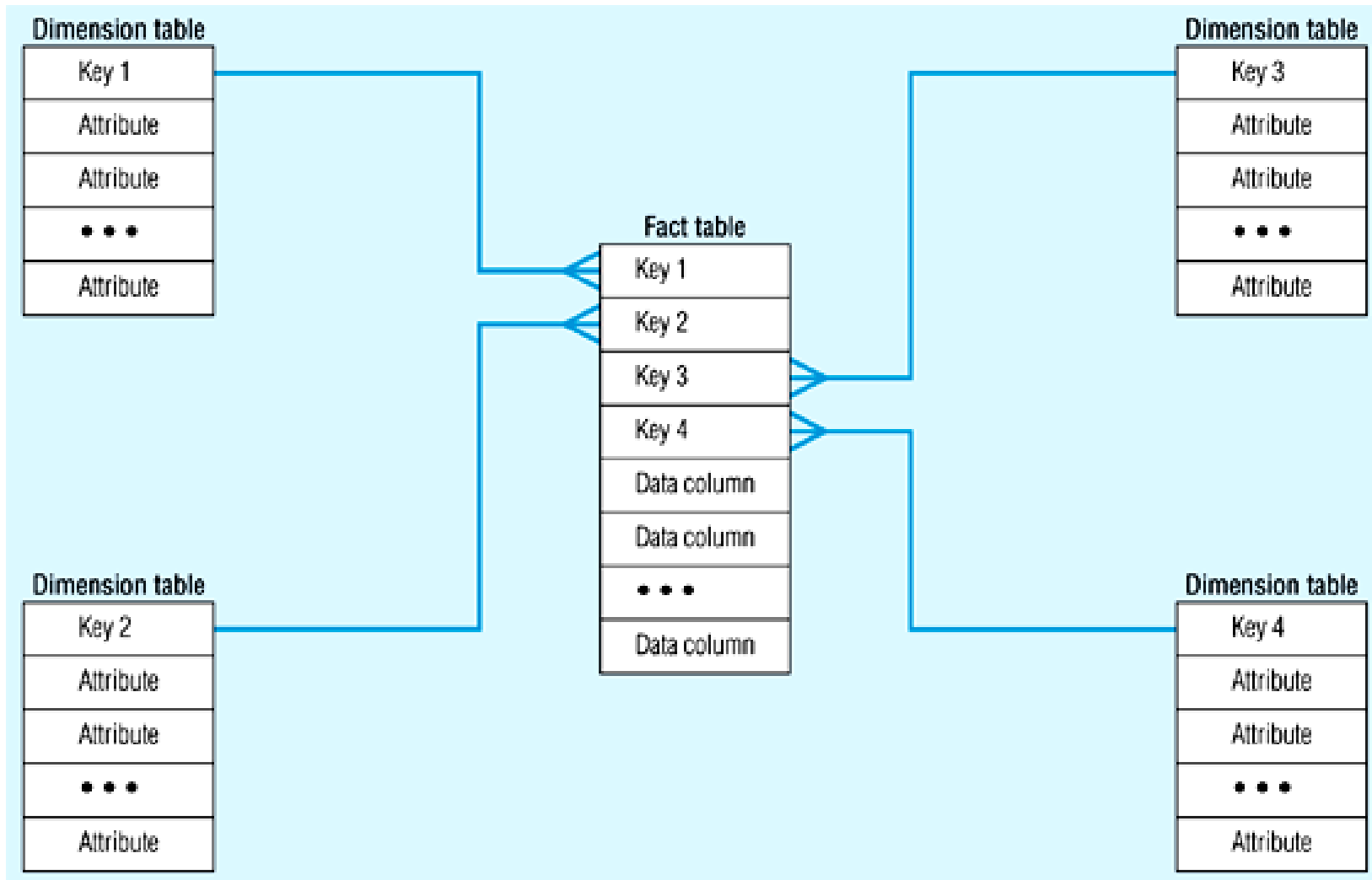
Esquema en estrella: Tabla de Dimensiones

- Se enlaza a la tabla de hechos (clave primaria única)
- Guarda los Atributos del negocio
- Más o menos constante los datos
- Contiene información textual descriptiva
- Filas anchas (muchos campos, incluso descriptivos)
- Tablas pequeñas (alrededor de un millón de filas)
- Ingresó a la tabla de hechos mediante una clave externa
- Fuertemente indexados

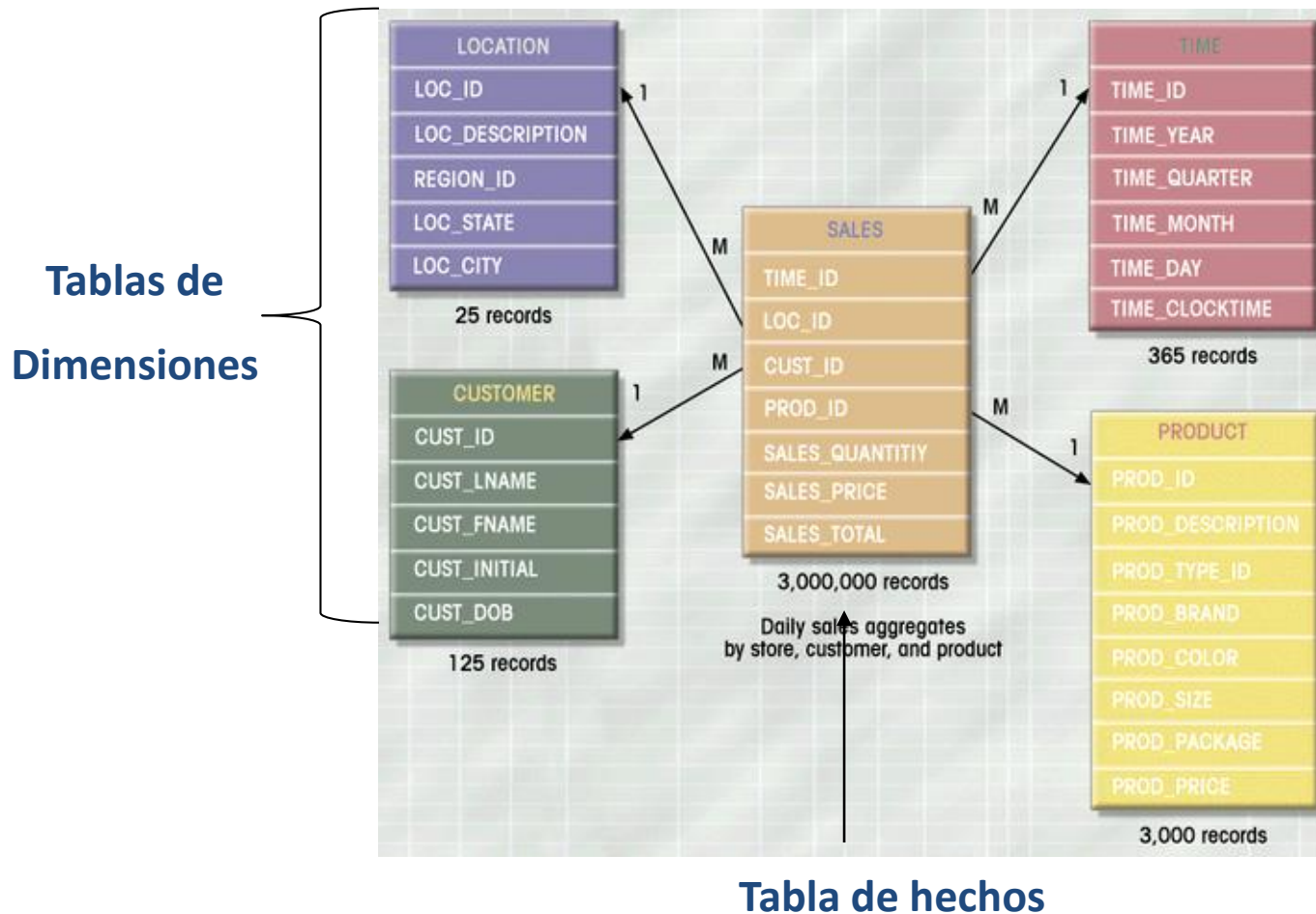
Esquema en estrella: Atributos

- Tablas de dimensiones contienen atributos
- Los atributos se utilizan para buscar, filtrar o clasificar los hechos
- Dimensiones proporcionan características descriptivas acerca de los hechos a través de sus atributos
- Debe definir los atributos comunes que se utilizará para reducir la búsqueda, agrupar información, o describir las dimensiones (por ejemplo, tiempo/lugar/producto)
- Sin límite matemático para el número de dimensiones (3D hace que sea fácil modelar)

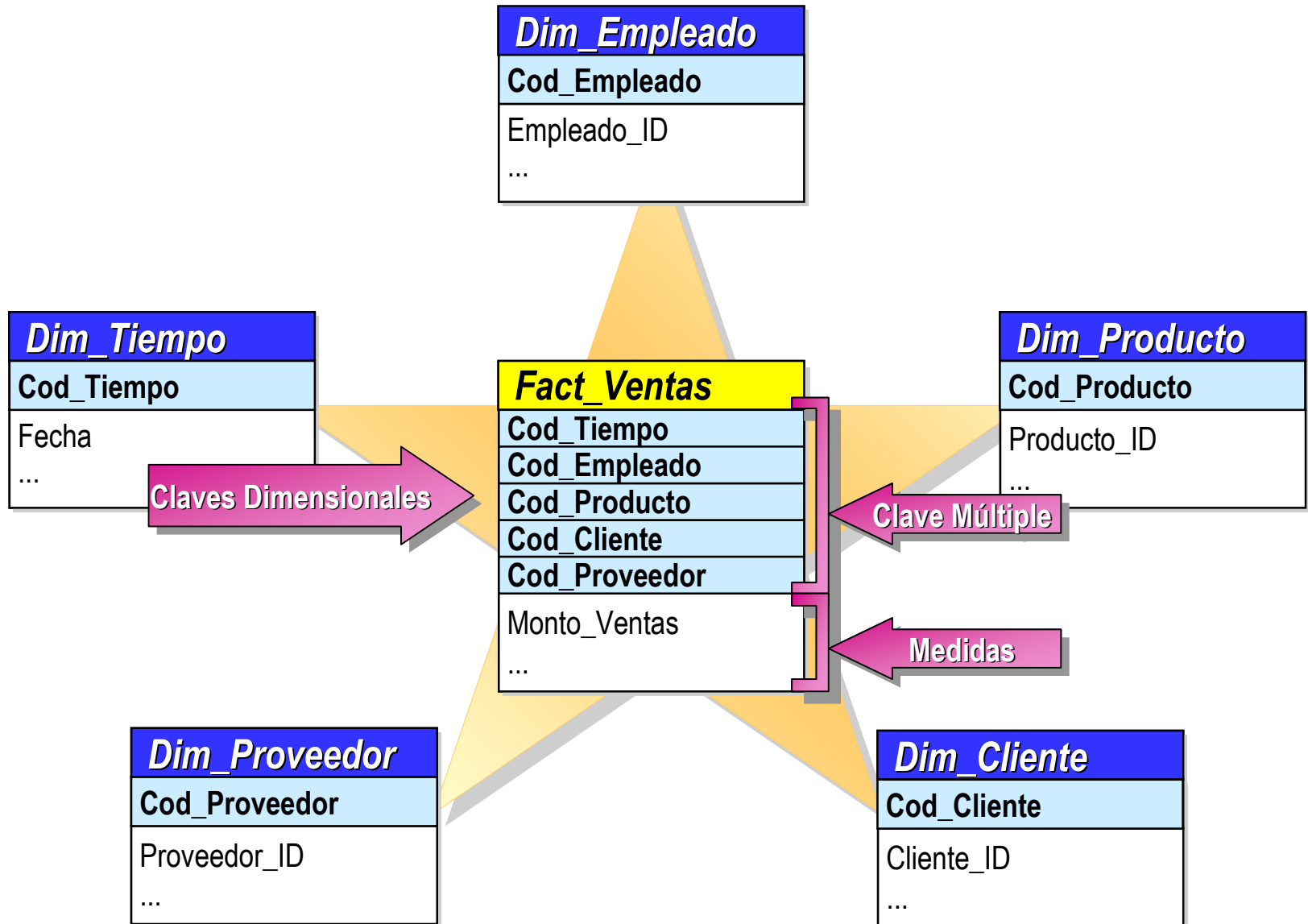
Esquema en estrella: Atributos



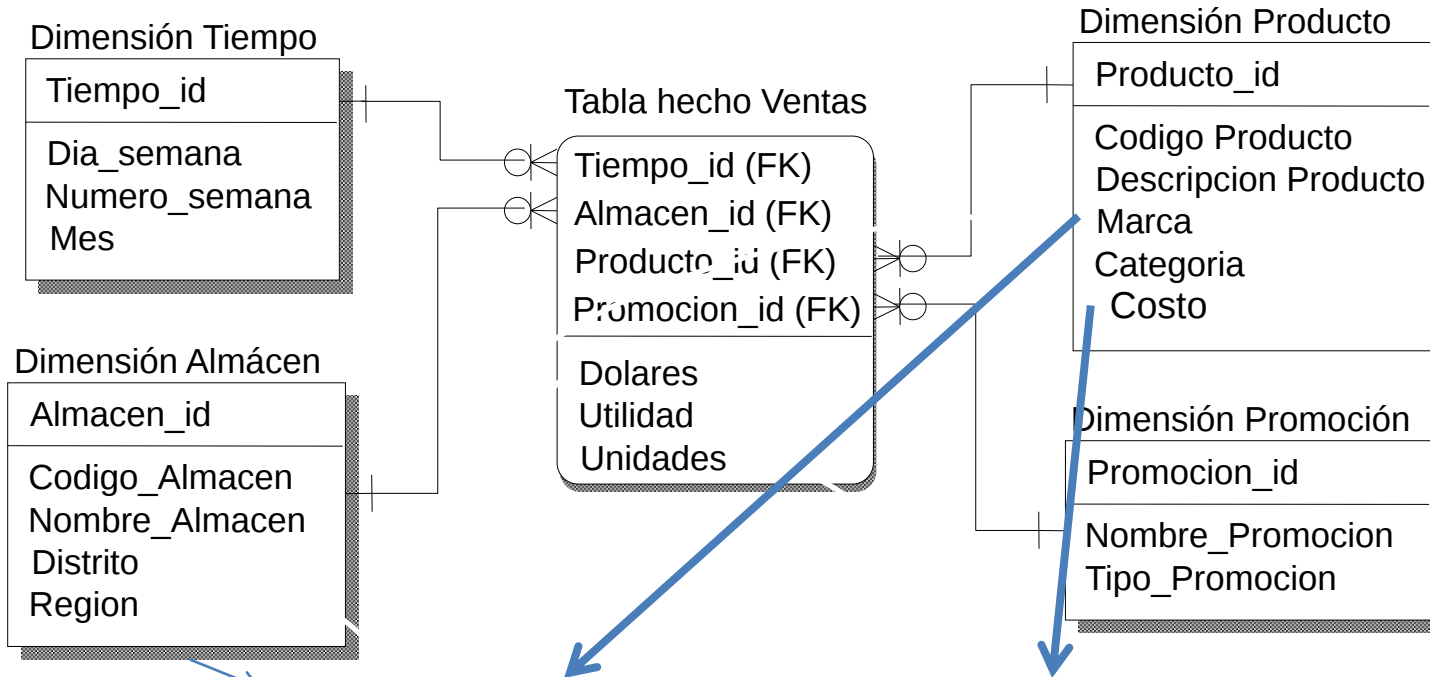
Ejemplo de esquema en estrella para ventas



Ejemplo de esquema en estrella para ventas



Ejemplo de esquema en estrella para Ventas



Distrito	Marca	Total Dolares	Total Costo	Utilidad	Unidades
Atherton	Clean Fast	\$ 1,233	\$ 1,058	\$ 175	10
Atherton	More Power	\$ 2,239	\$ 2,200	\$ 39	2
Atherton	Zippy	\$ 848	\$ 650	\$ 198	4
Belmont	Clean Fast	\$ 2,097	\$ 1,848	\$ 249	6
Belmont	More Power	\$ 2,428	\$ 2,350	\$ 78	3
Belmont	Zippy	\$ 633	\$ 580	\$ 53	5

Conclusiones Esquema en Estrella

- Las tablas de hechos están relacionados a cada tabla de dimensión en una **relación Muchos a Uno**
- Tabla de hechos está **relacionado con muchas tablas** de dimensiones
- La clave principal de la tabla de hechos es compuesta de las **claves principales de las tablas de dimensiones**
- Cada tabla de hecho está diseñada para **responder a una pregunta específica de IN**

Esquema Copo de nieve

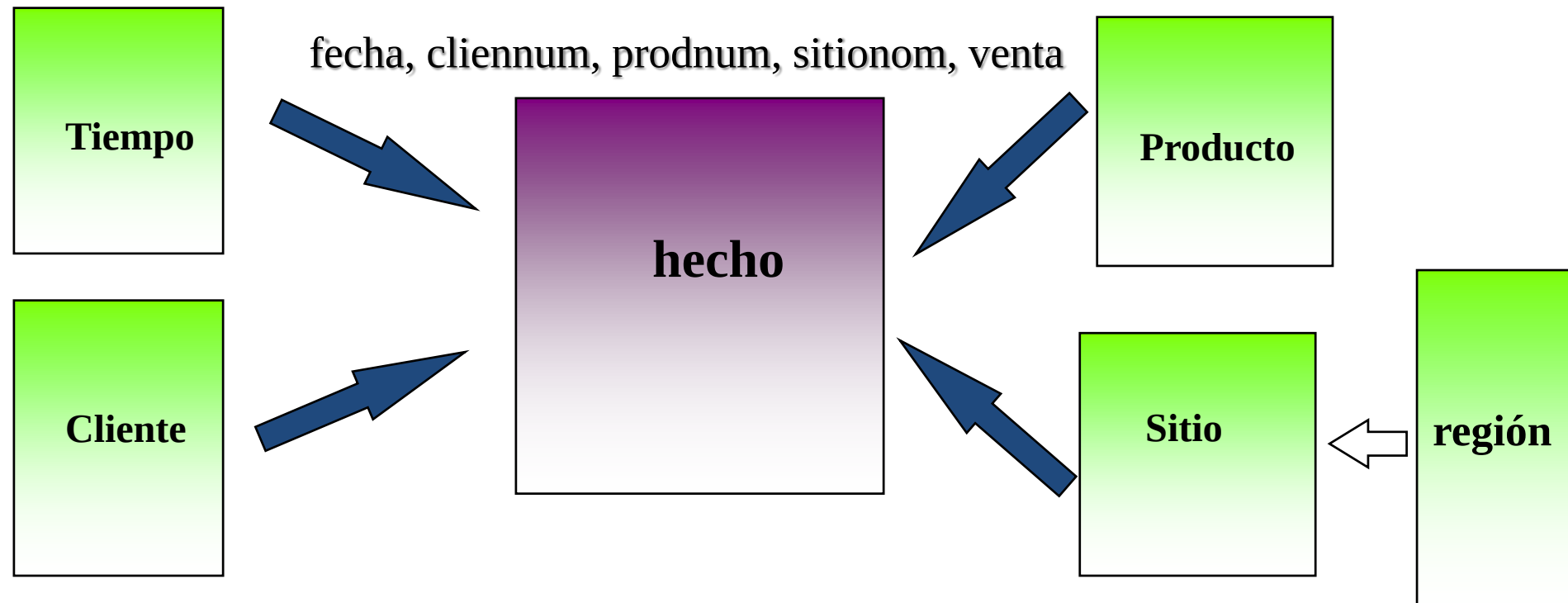
Un refinamiento del esquema en estrella donde alguna jerarquía dimensional se despliega en un conjunto de tablas de dimensiones más pequeñas

- Forma:
 - Esquema en estrella con dimensiones secundarias
 - Fácil de mantener y ahorra almacenamiento

Copo de nieve cuando las dimensiones tienen muchos atributos

Esquema Copo de nieve

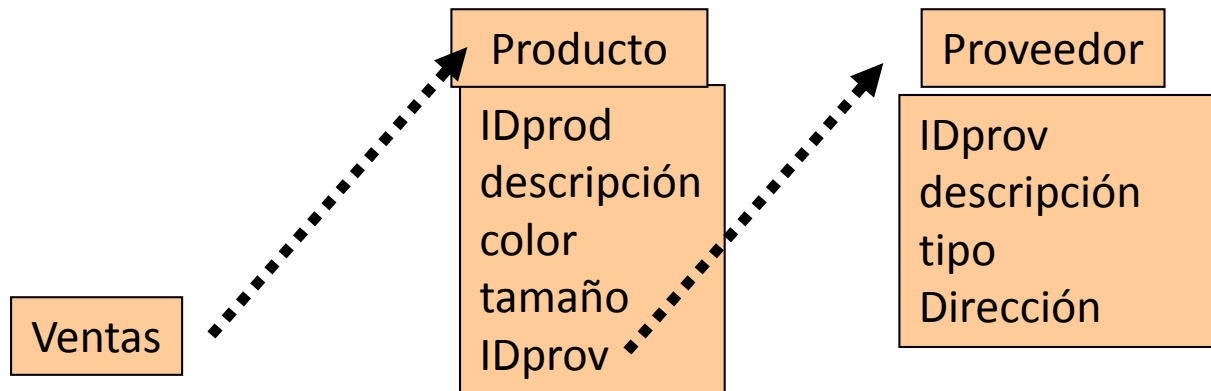
Representa jerarquía dimensional



Esquema Copo de nieve

Beneficios

- Evita la duplicación
- Conduce a las constelaciones (varias tablas de hechos con dimensiones compartidas)



Esquema Copo de nieve

Dimensión del almacén

Clave Almacen
Descripción
Ciudad
Estado
ID Distrito

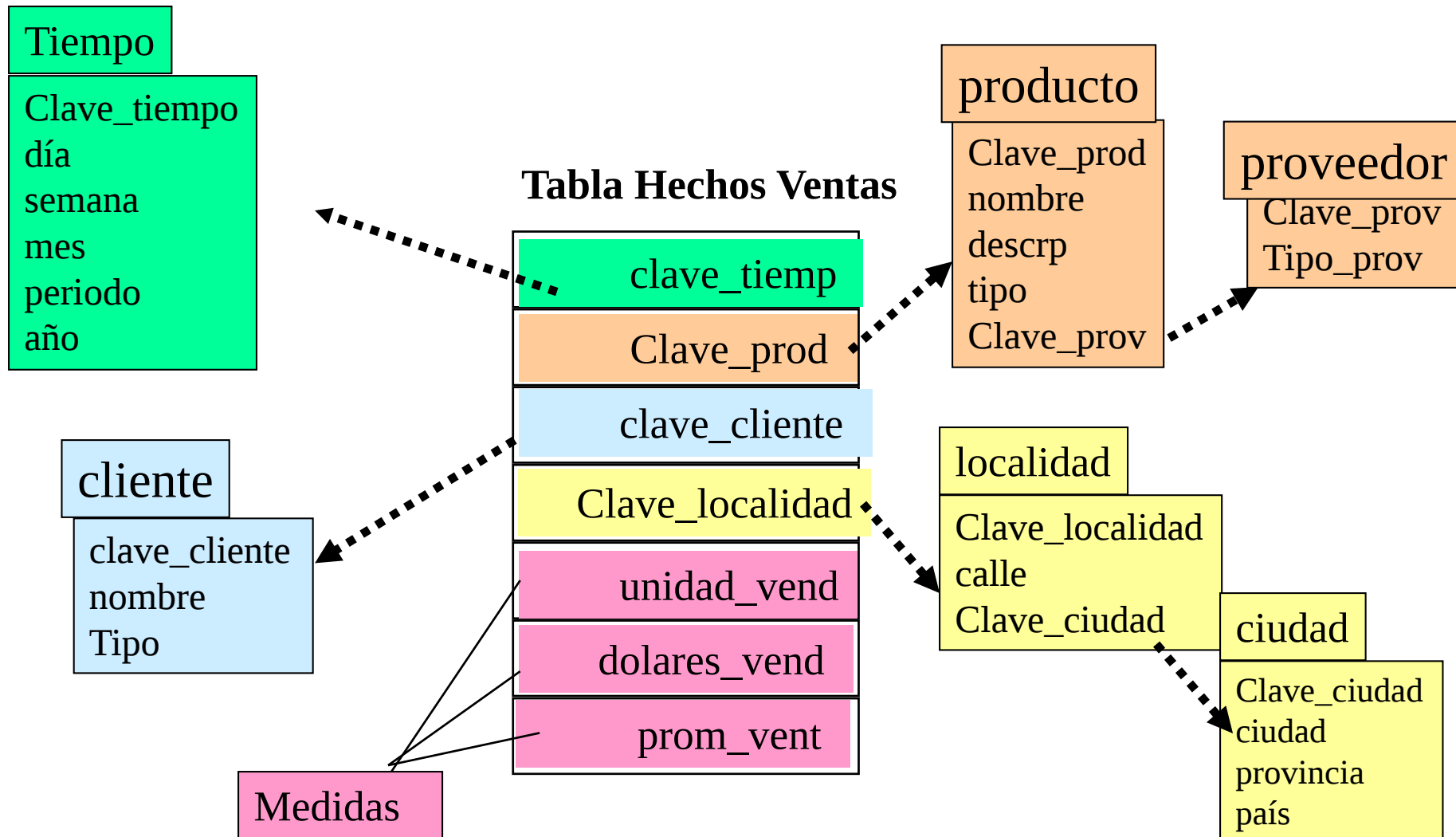
ID Distrito
Desc_distrito
ID_Región

ID Región
Descrip Región
Gerente Región..

Tabla hecho almacén

Clave Almacen
Clave producto
Clave Periodo
Monto
Unidades
Utilidad

Esquema Copo de nieve



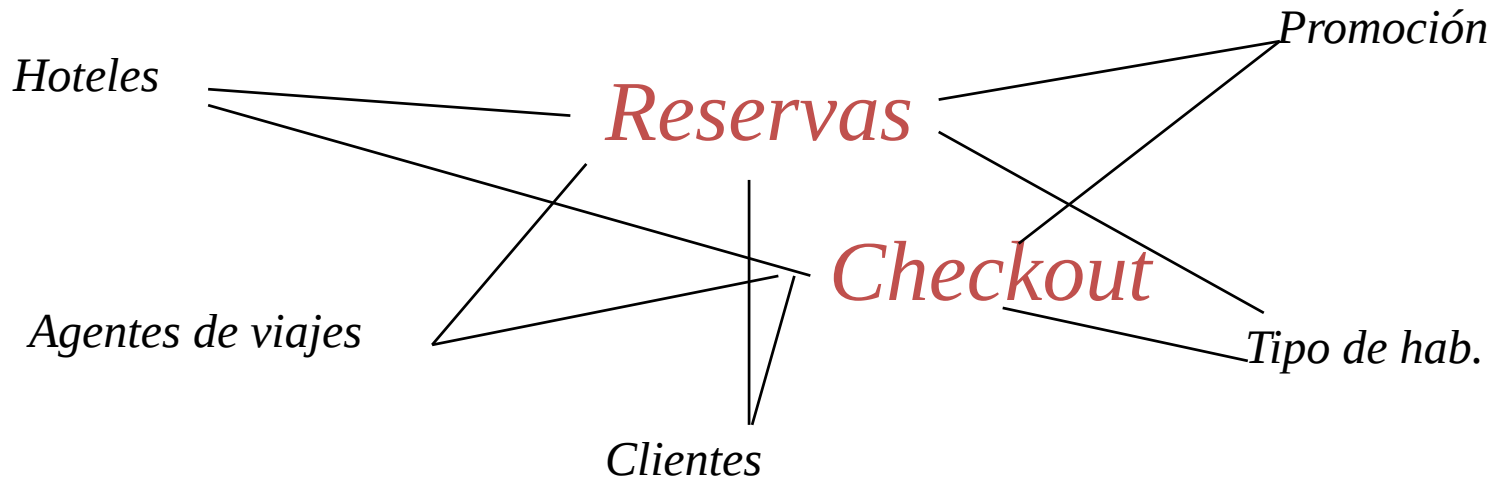
Esquema de Constelación de hechos

Varias tablas de hechos **comparten** tablas de dimensiones

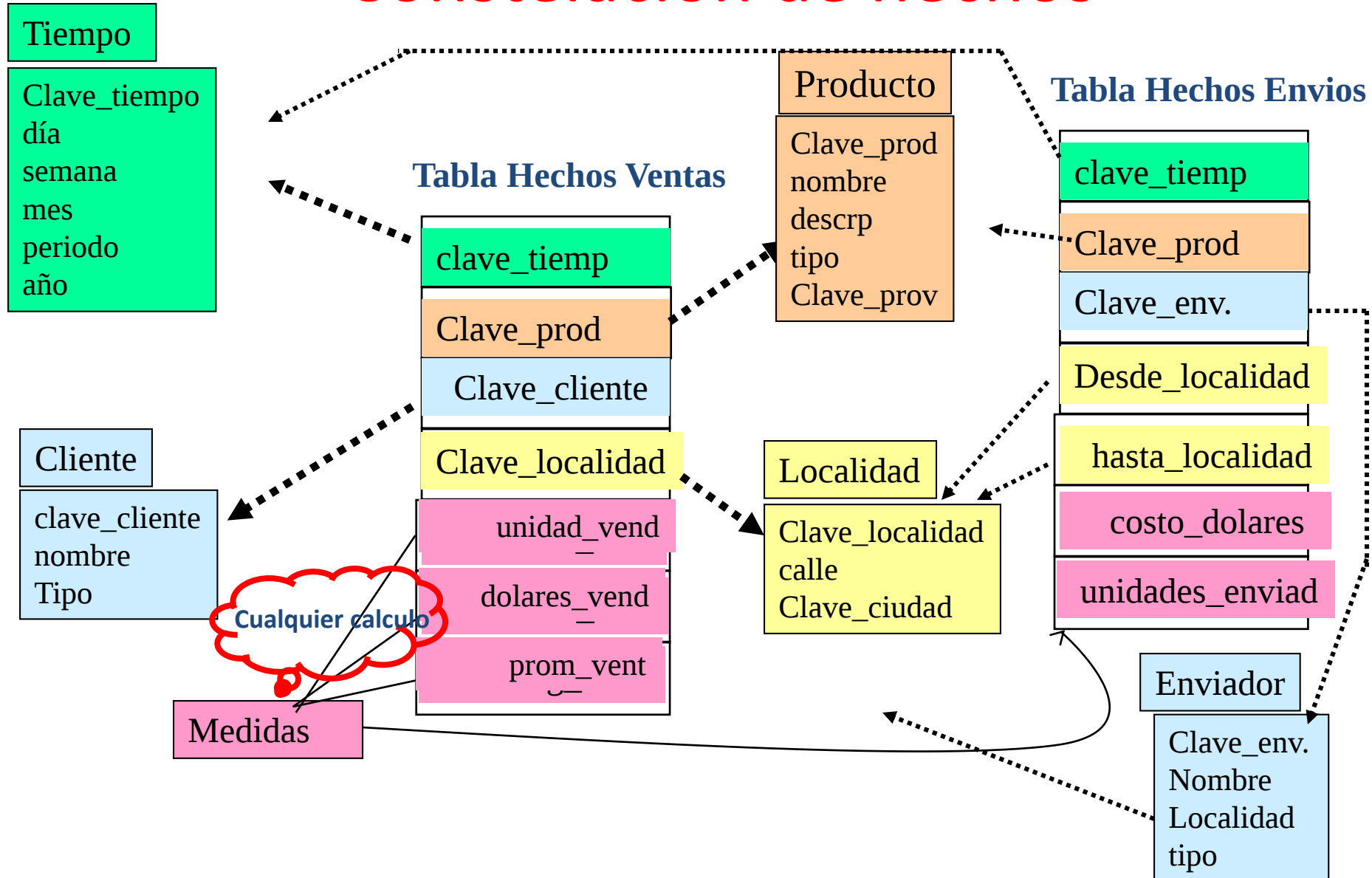
- vistos como una **colección de estrellas**, por lo tanto, llamados **esquema de galaxia o constelación de hecho**

Esquema de Constelación de hechos

Reservas y Checkout pueden compartir tablas de dimensiones en la industria hotelera



Esquema de Constelación de hechos



De las Tablas a los cubos de datos

- Un data warehouse se basa en un **modelo de datos multidimensional**
- Todo los datos se pueden ver en la forma de un **cubo de datos**
- Un cubo de datos permite ver múltiples dimensiones

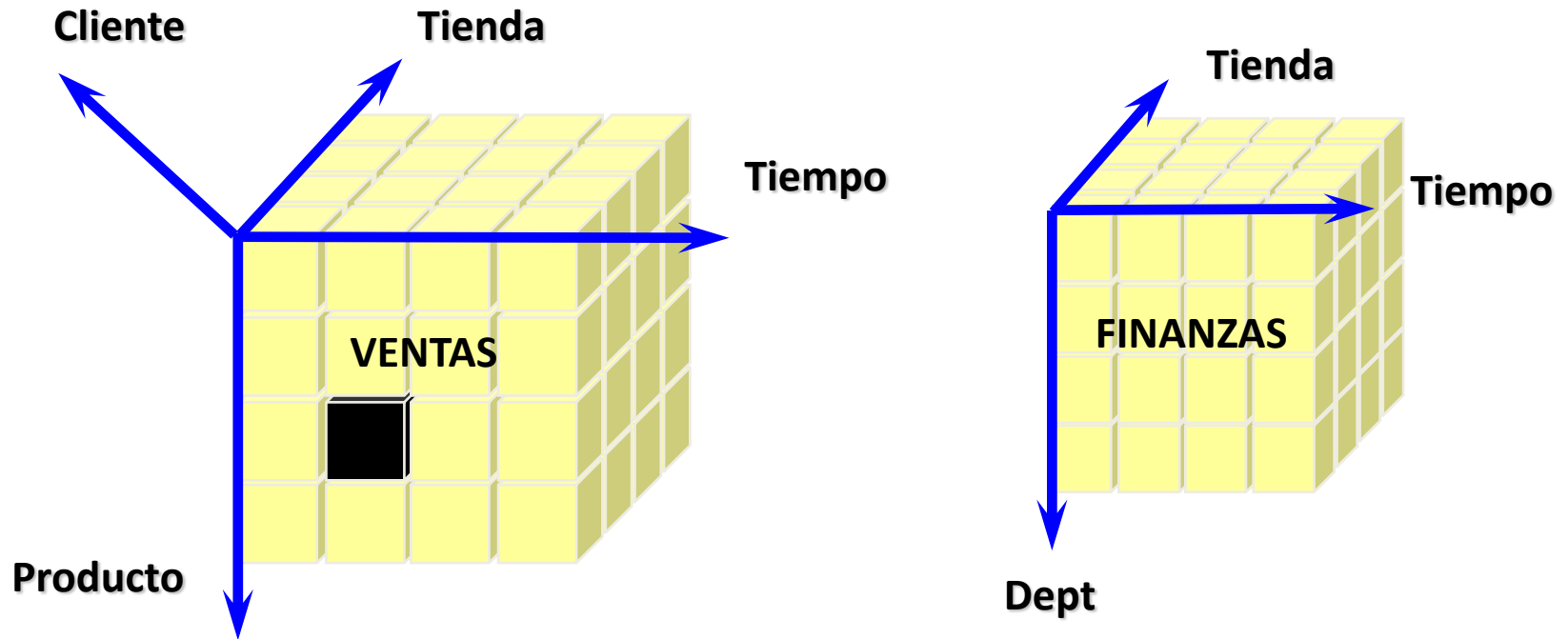
Un cubo n-D se llama un paralelepípedo.

Base de datos relacional

	Atributo 1 Nombre	Atributo 2 edad	Atributo 3 sexo	Atributo 4 No. Emp
Fila 1	Anderson	31	F	1001
Fila 2	Green	42	M	1007
Fila 3	Lee	22	M	1010
Fila 4	Ramos	32	F	1020

Tabla de empleados

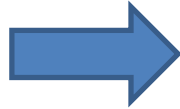
Modelo BD multidimensional



Los datos se encuentra en la intersección de las dimensiones

Dos dimensiones

City	Time	Total Revenue
Glasgow	Q1	29726
Glasgow	Q2	30443
Glasgow	Q3	30582
Glasgow	Q4	31390
London	Q1	43555
London	Q2	48244
London	Q3	56222
London	Q4	45632
Aberdeen	Q1	53210
Aberdeen	Q2	34567
Aberdeen	Q3	45677
Aberdeen	Q4	50056
.....		
.....		

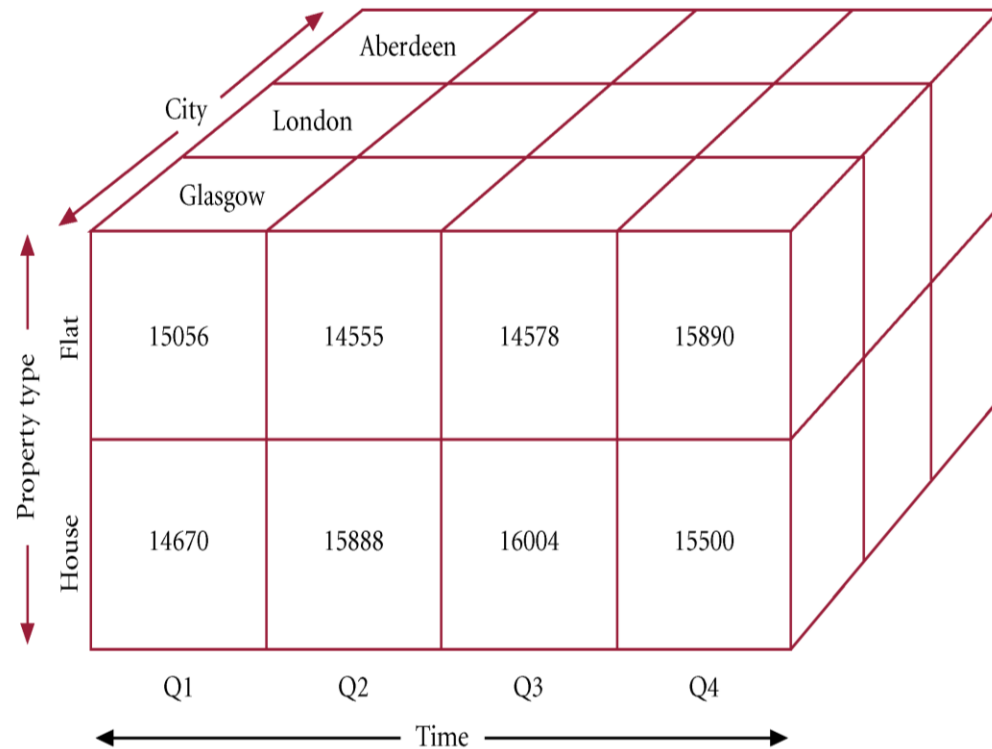


City				
City Quarter	Glasgow	London	Aberdeen	
Q1	29726	43555	53210	
Q2	30443	48244	34567	
Q3	30582	56222	45677	
Q4	31390	45632	50056	

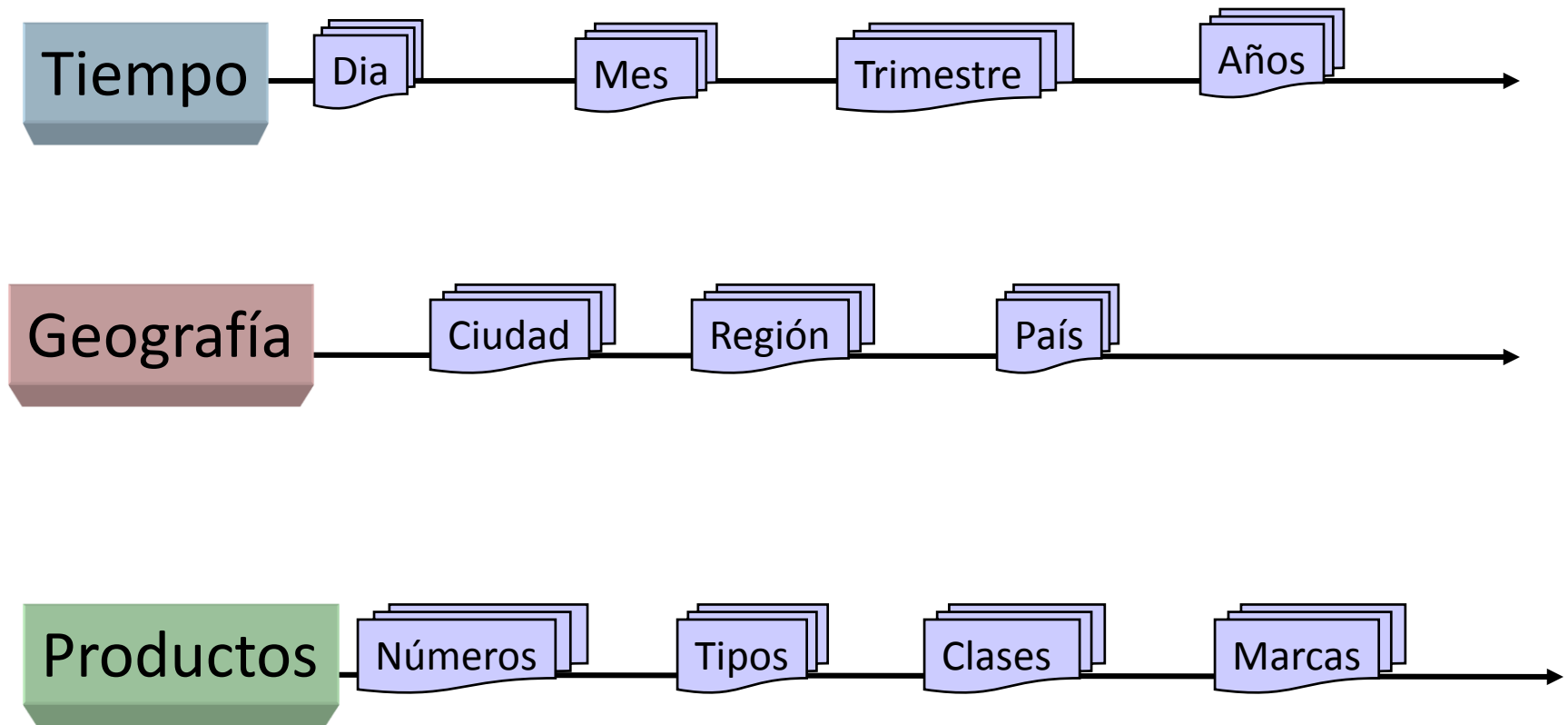
Tres dimensiones



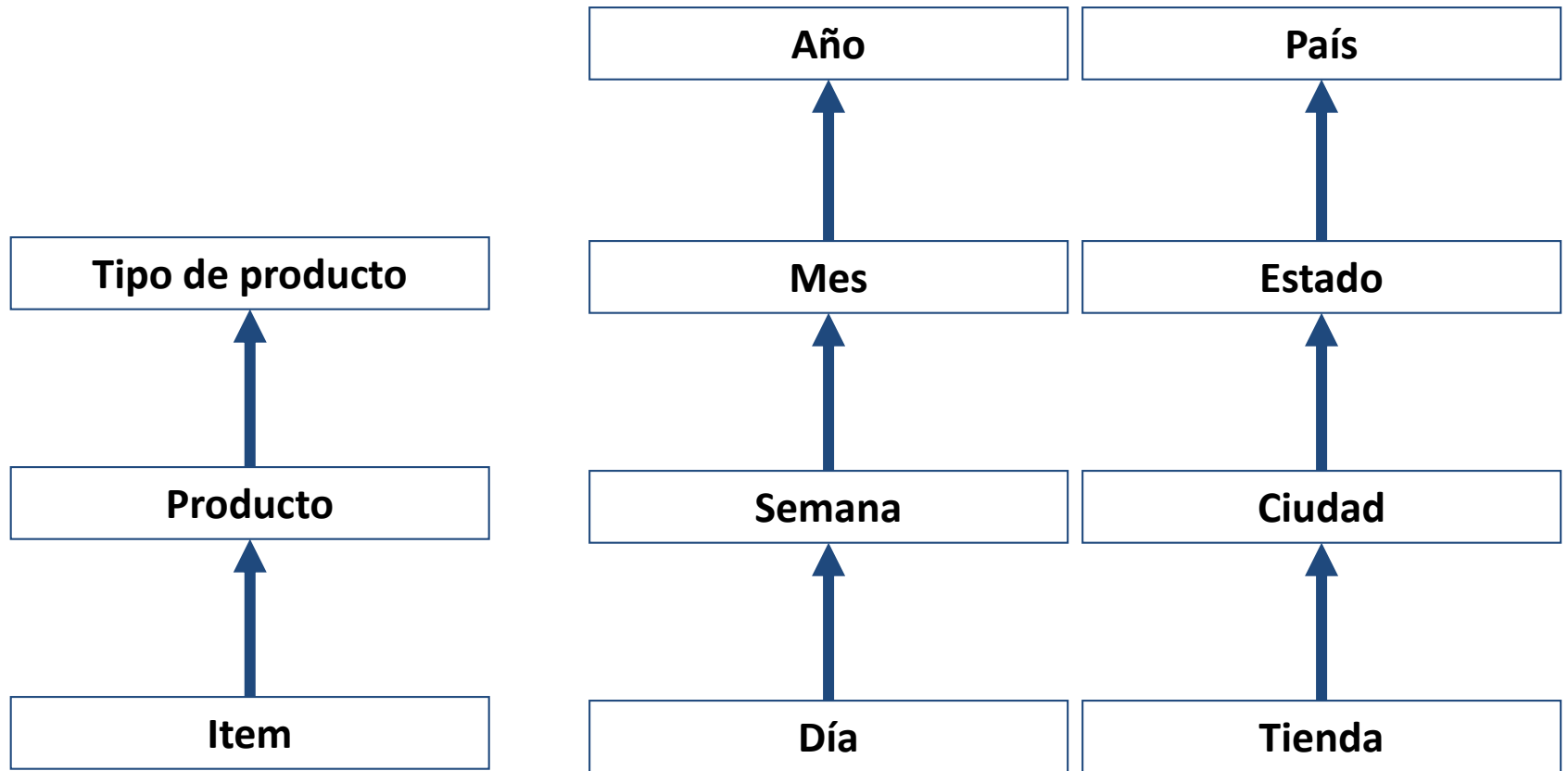
Property Type	City	Time	Total Revenue
Flat	Glasgow	Q1	15056
House	Glasgow	Q1	14670
Flat	Glasgow	Q2	14555
House	Glasgow	Q2	15888
Flat	Glasgow	Q3	14578
House	Glasgow	Q3	16004
Flat	Glasgow	Q4	15890
House	Glasgow	Q4	15500
Flat	London	Q1	19678
House	London	Q1	23877
Flat	London	Q2	19567
House	London	Q2	28677
.....			
.....			



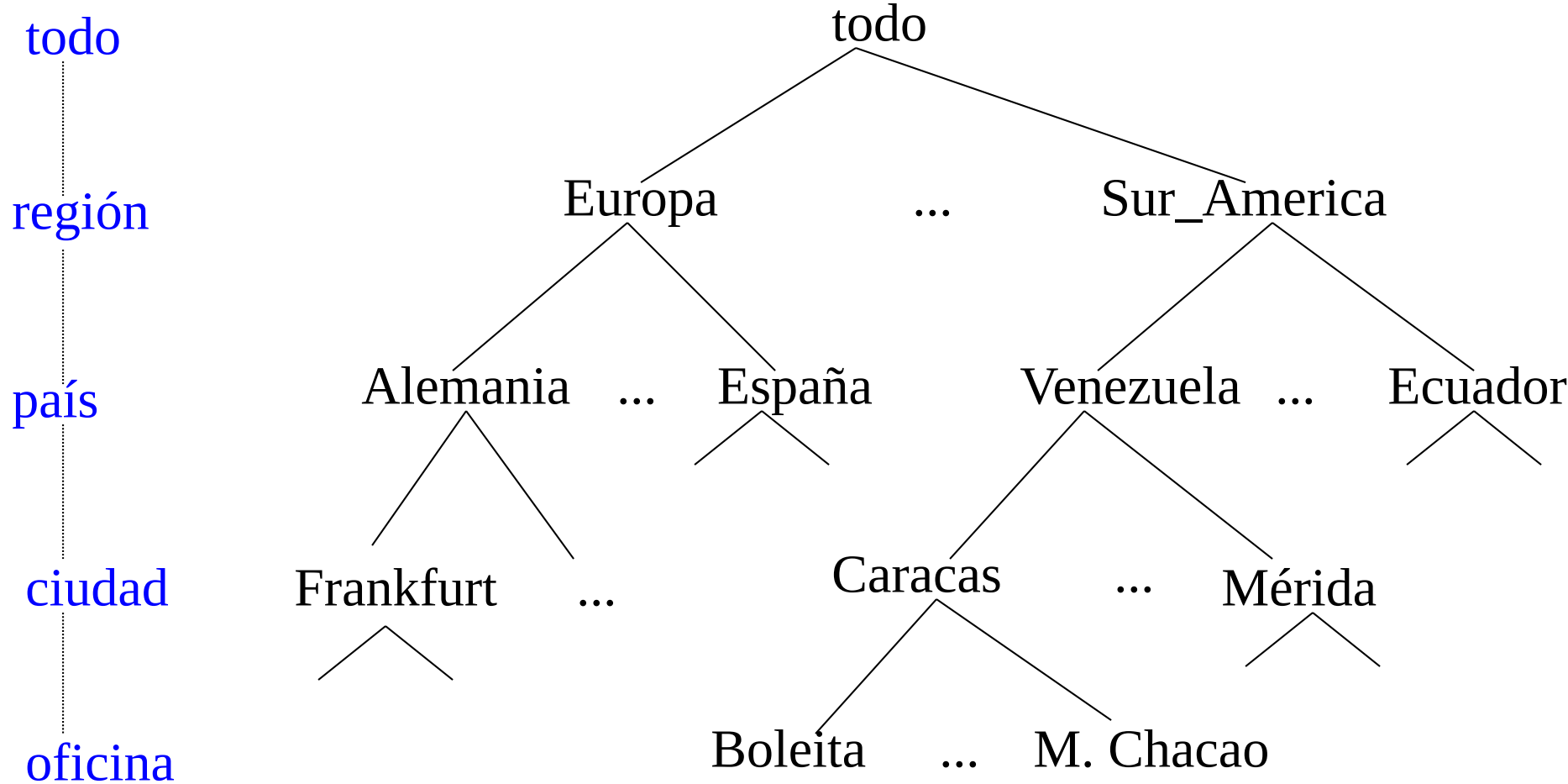
La granularidad de las dimensiones



Jerarquía Dimensional



Jerarquía Dimensional (localidad)



Jerarquía Dimensional

- jerarquía de esquema

día < mes < cuatrimestre < año

- Se pueden agrupar las jerarquías

{día 1 al 10}

{días} < 30

Las multidimensiones

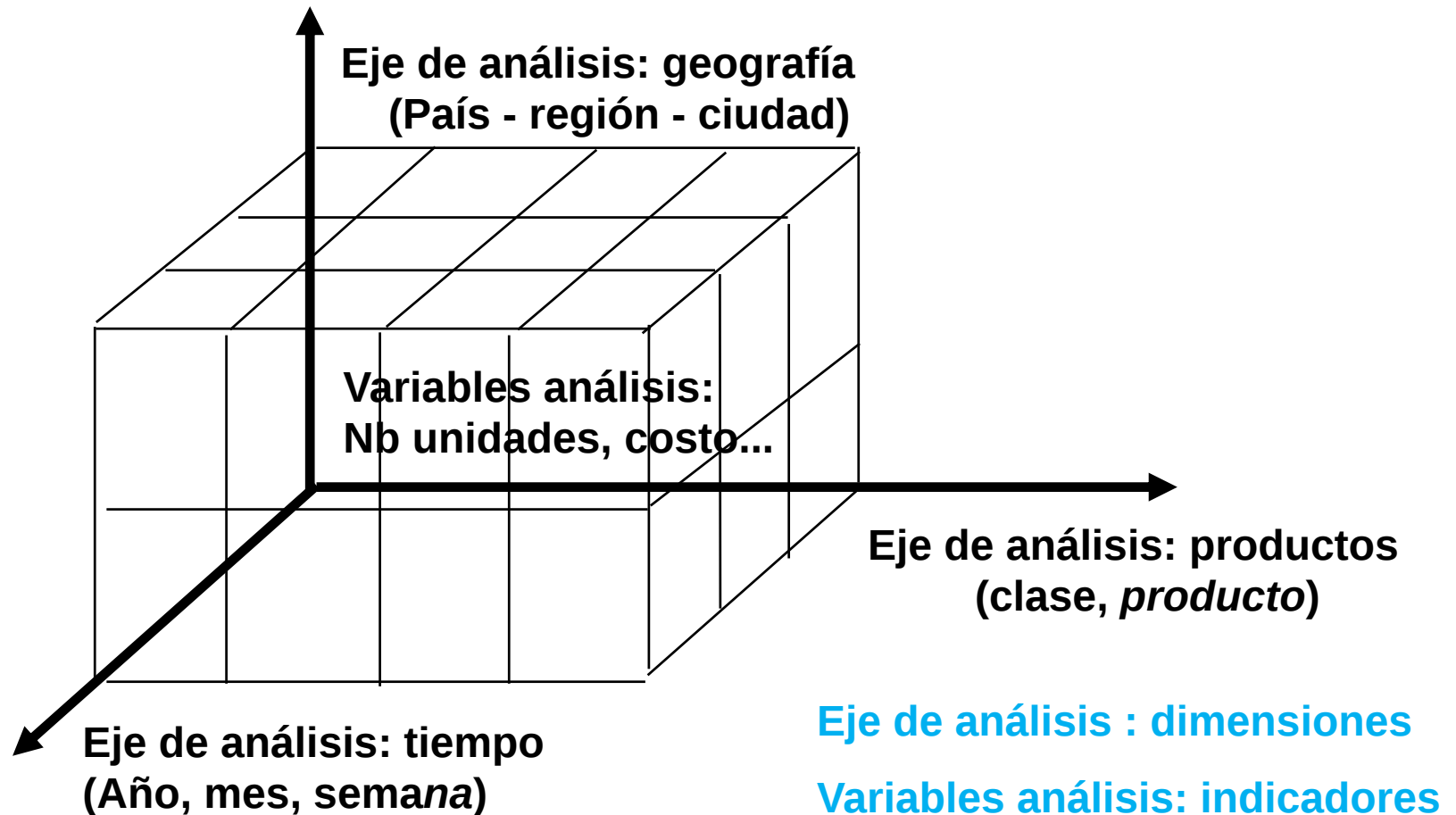
- Dimensiones:

- Tiempo
- Geografía
- Productos
- Clientes
- Canales de ventas.....

- Indicadores:

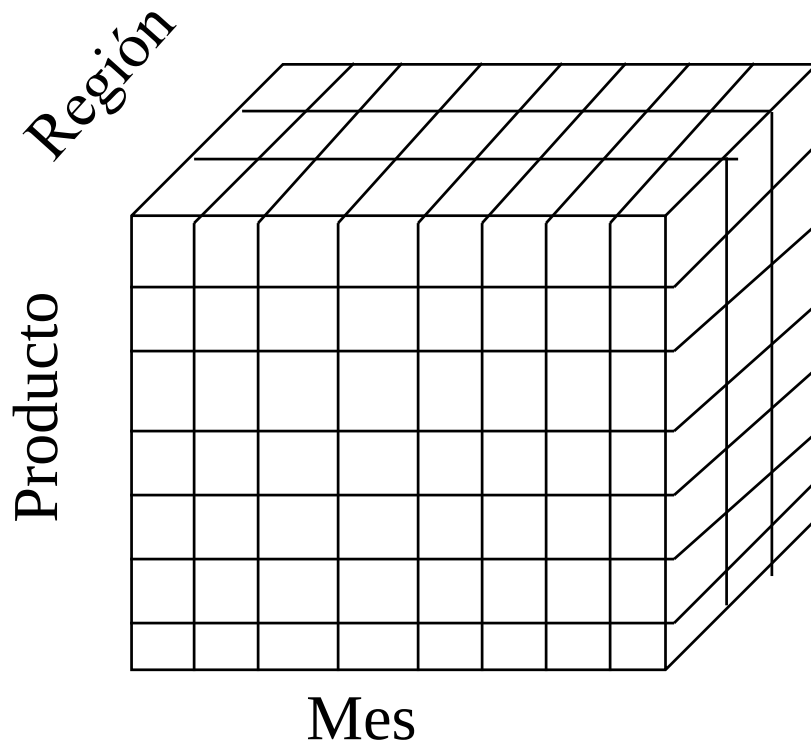
- Número de unidades vendidas
- Costo

Cubo de dato y las dimensiones



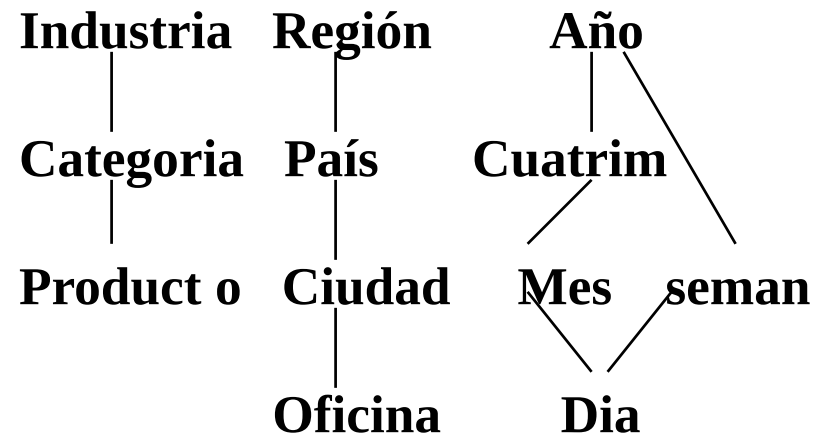
Datos Multidimensionales

El volumen de ventas en función del producto,
el mes, y el área



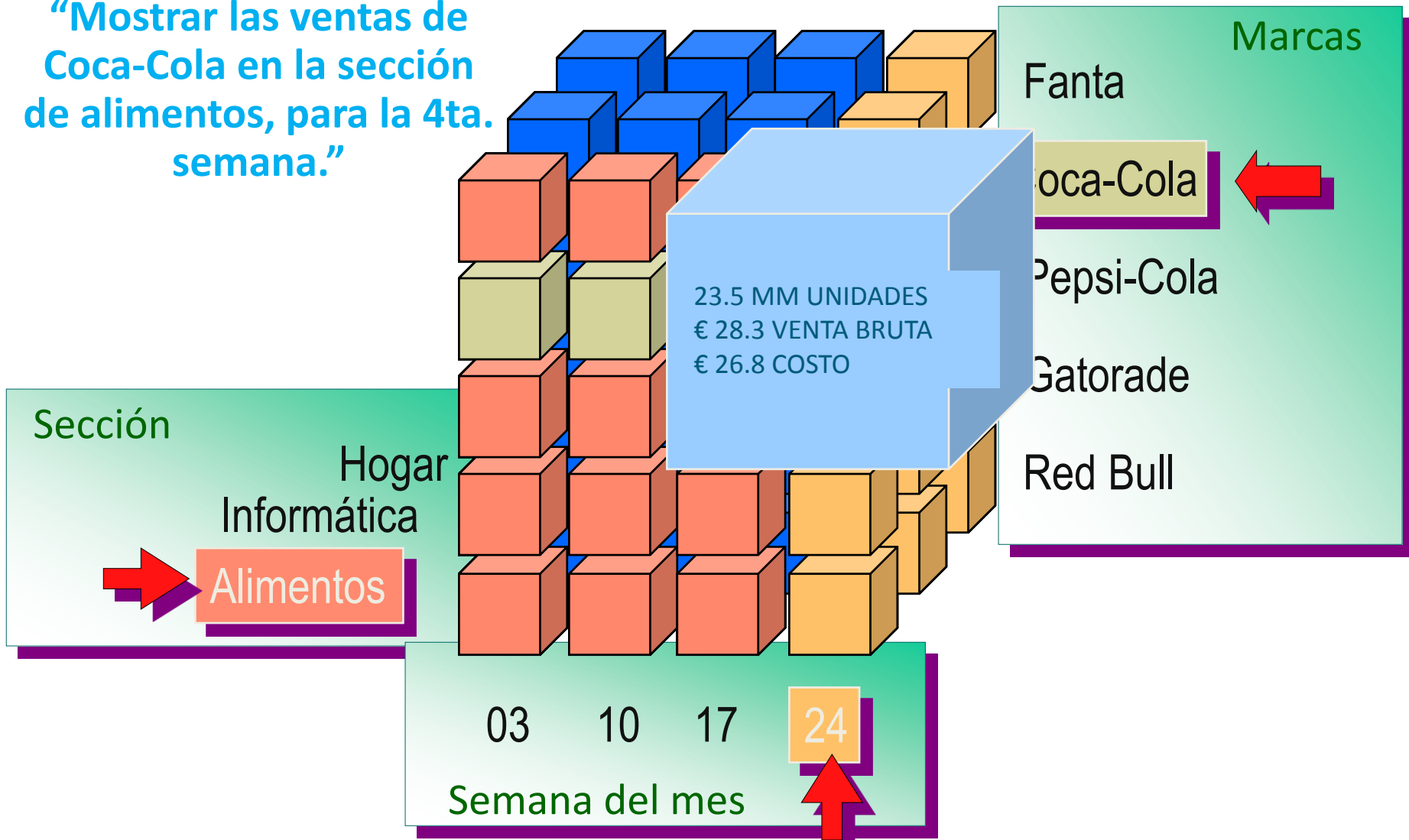
Dimensiones: Producto, Localidad, Tiempo

Caminos jerárquicos

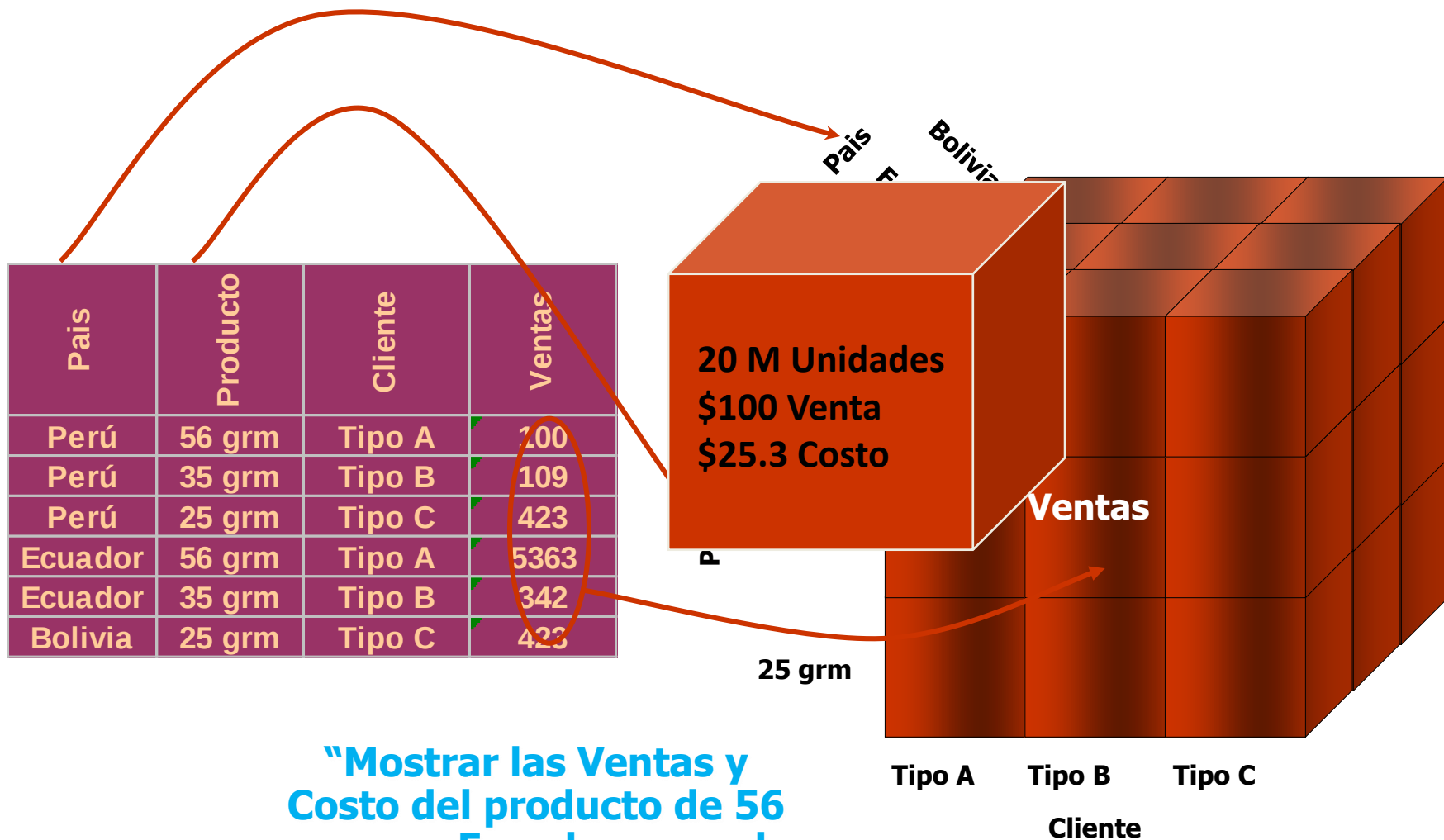


Cubo Multidimensional

“Mostrar las ventas de Coca-Cola en la sección de alimentos, para la 4ta. semana.”

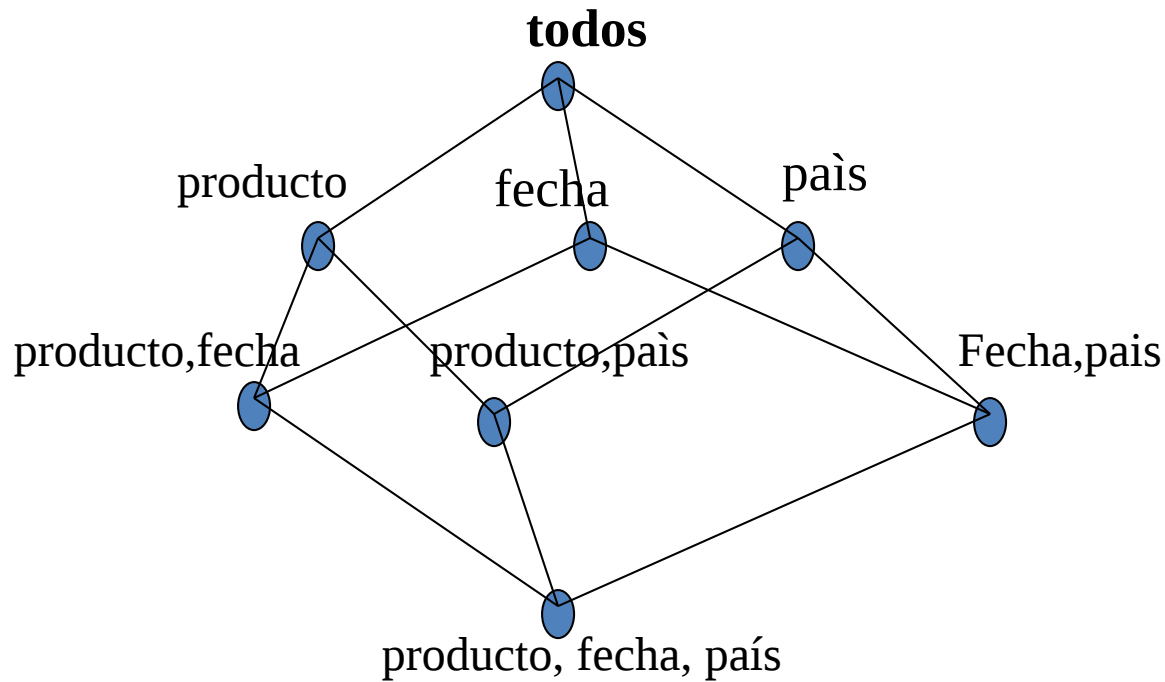


Cubo Multidimensional



**“Mostrar las Ventas y
Costo del producto de 56
grm en Ecuador para el
Tipo de Cliente A”**

Cuboides correspondientes al Cubo



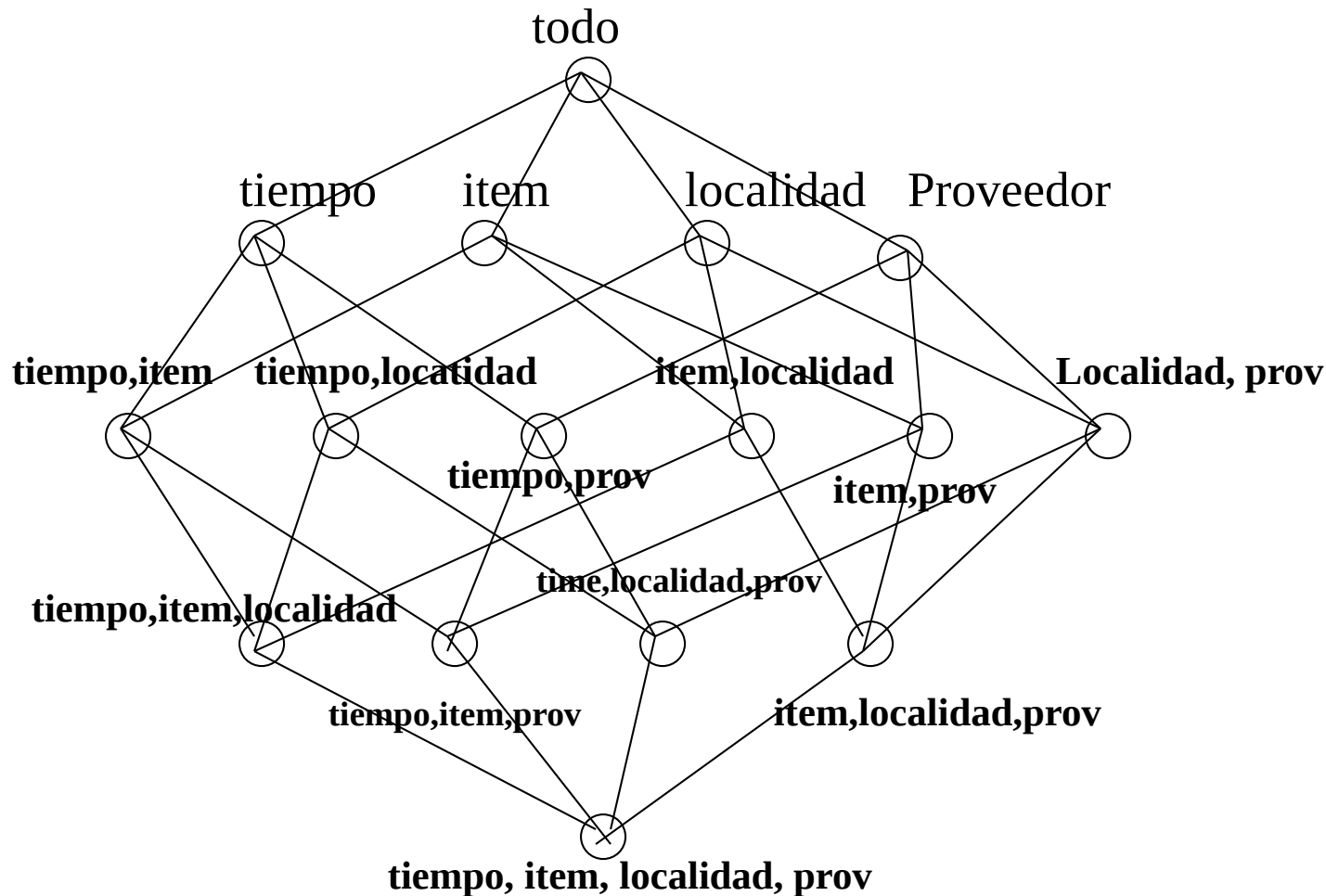
0-D cuboide

1-D cuboides

2-D cuboides

3-D cuboides

Cuboides correspondientes al Cubo



0-D cuboide

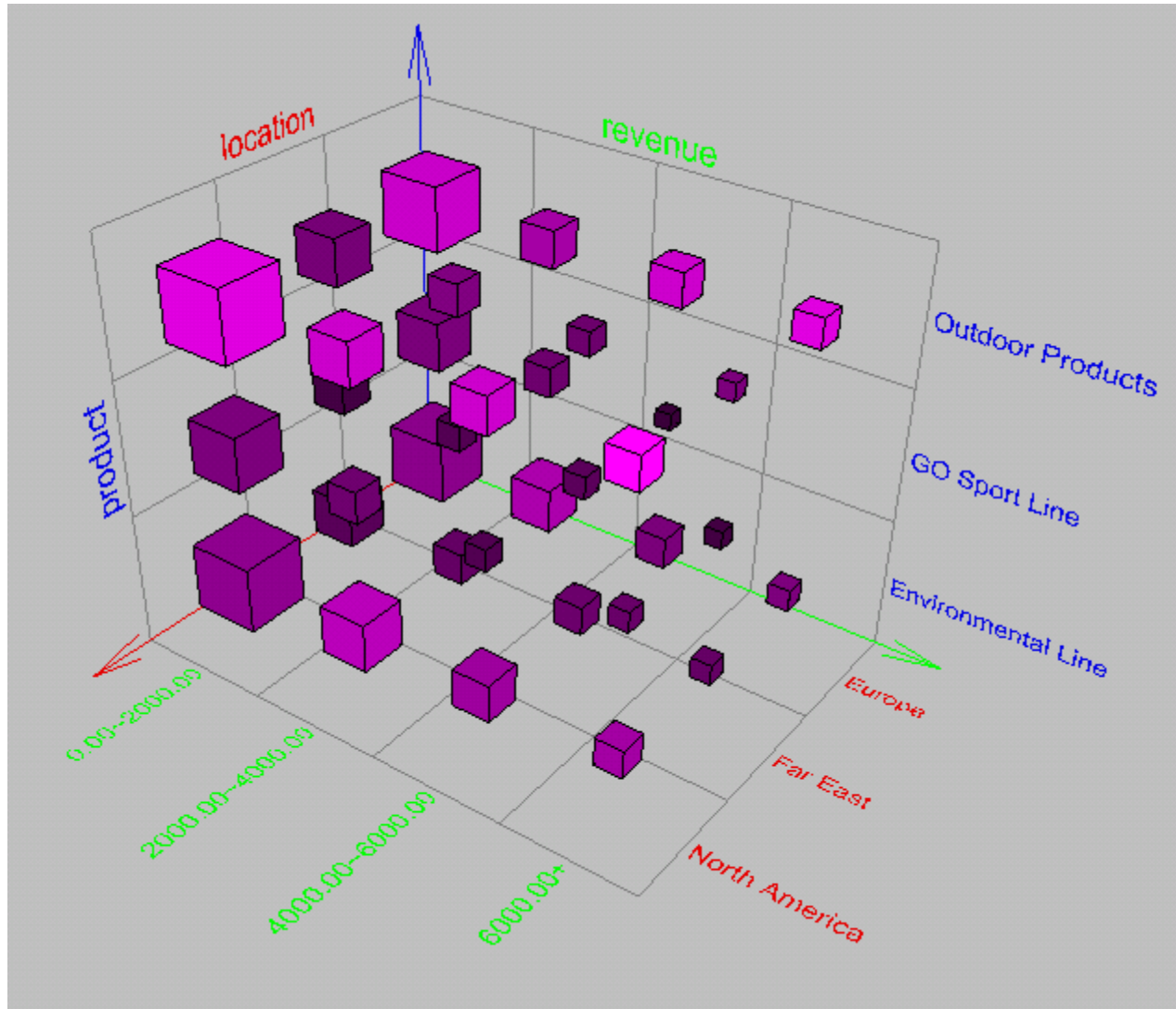
1-D cuboides

2-D cuboides

3-D cuboides

4-D cuboides

Navegar por un cubo de datos



- Visualización
- OLAP

Navegar por un cubo de datos

2 dimensiones

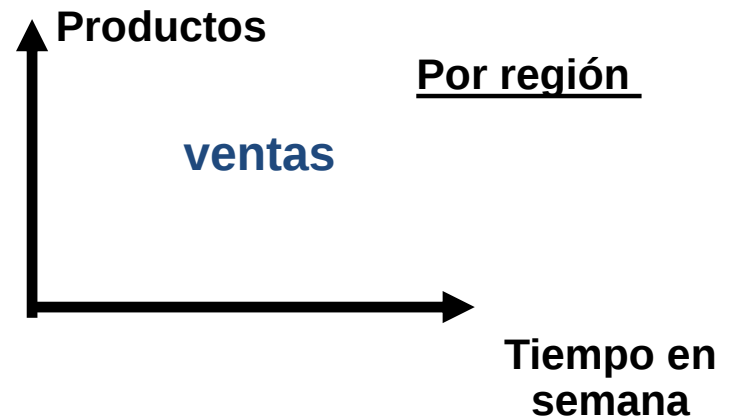


1 dimensión

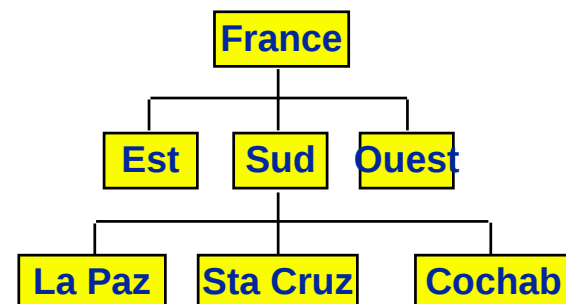


3 dimensiones

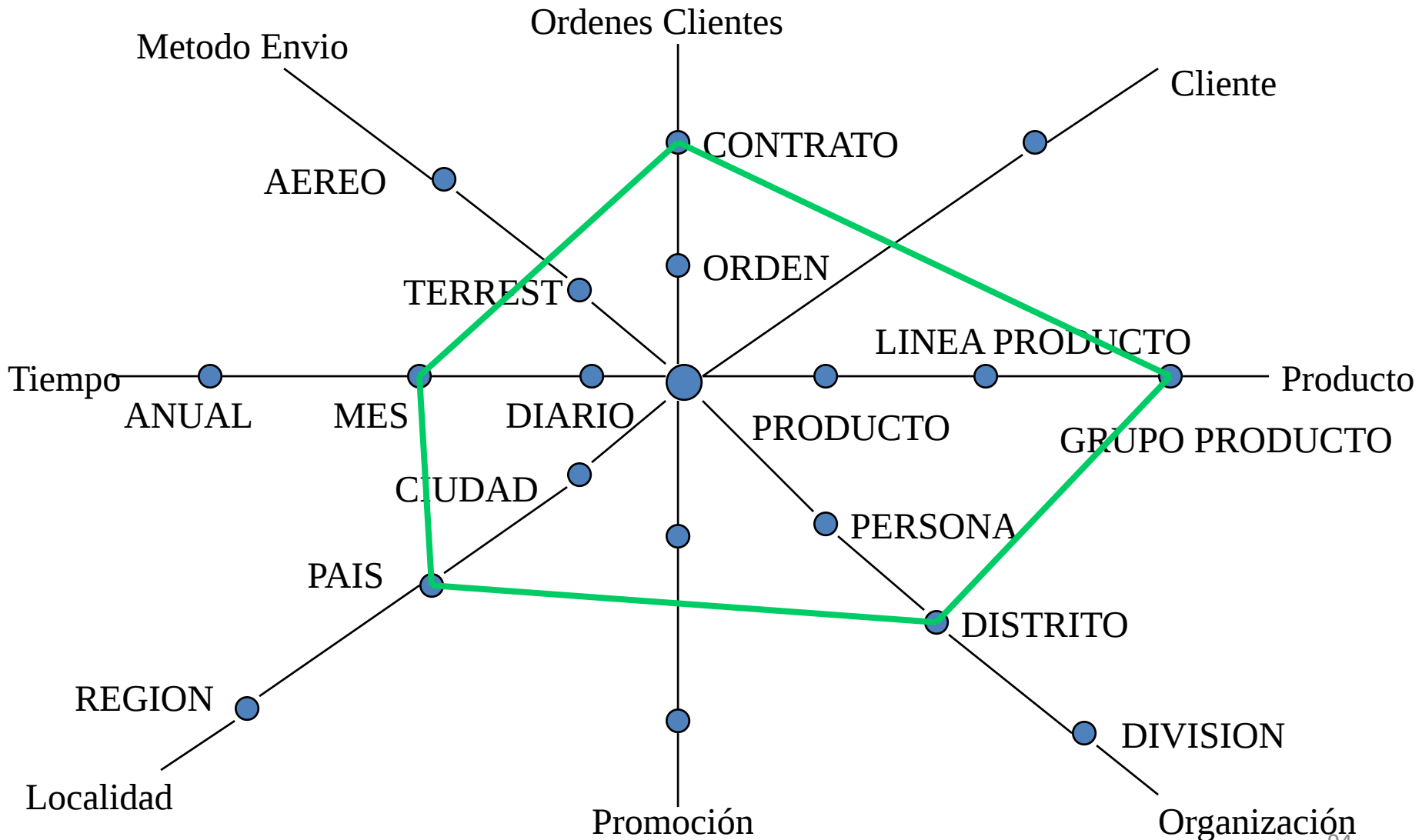
Copo de nieve



Zoom a una dimensión



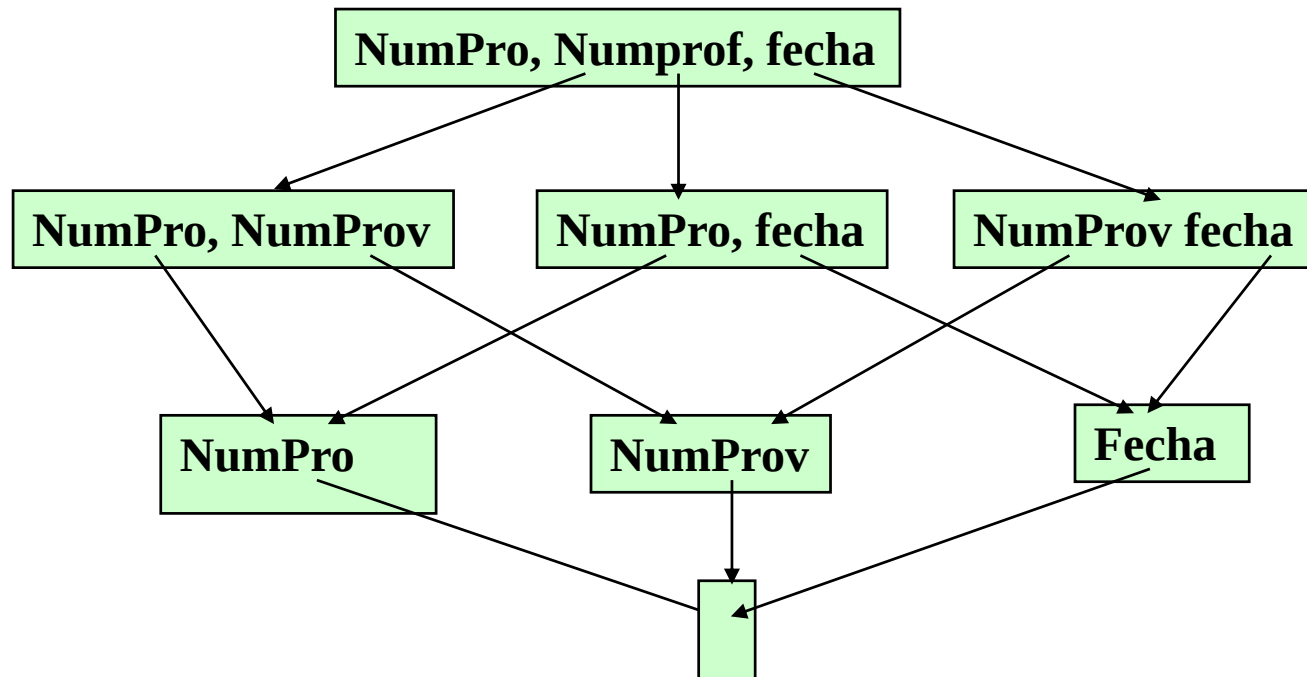
Modelo de consulta



Navegar por un cubo de datos

- Un cubo se puede rotar, agrupar, etc.

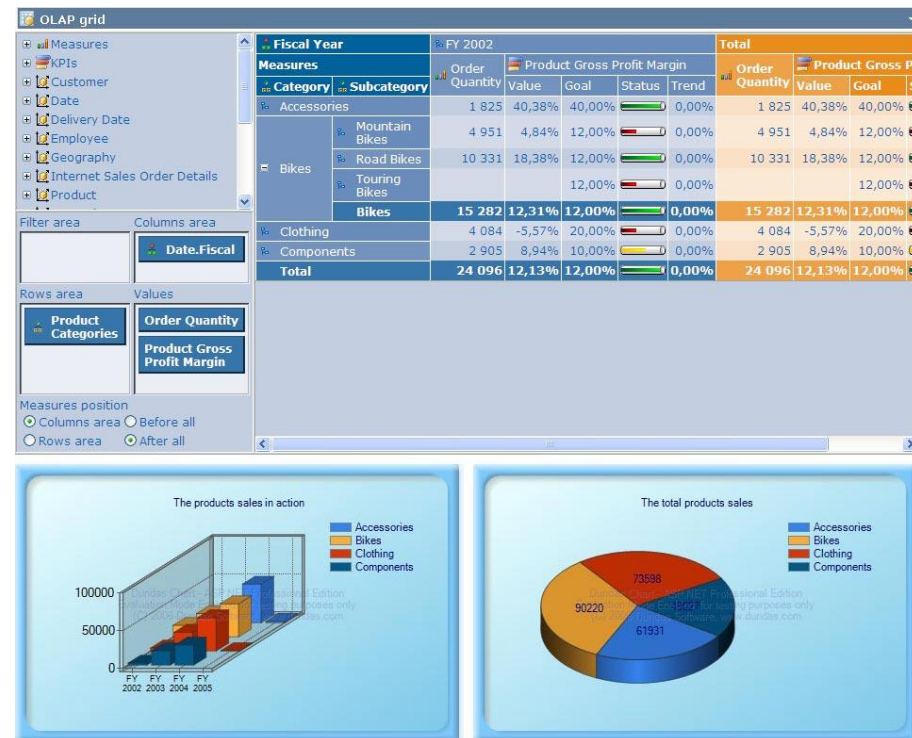
Se obtienen retículas de puntos de vista



Herramientas para explotación del Datawarehouse

Análisis multidimensional (OLAP online analytical processing)

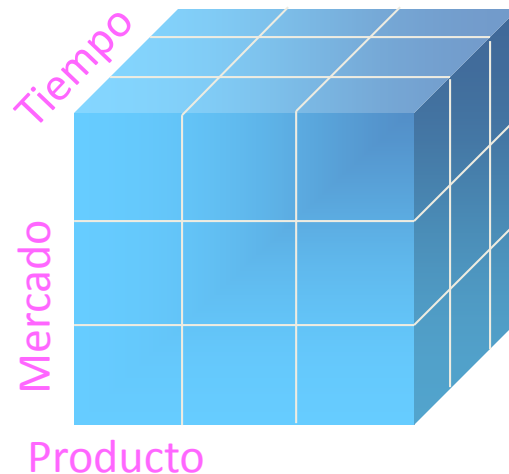
Facilitan el análisis de datos a través de dimensiones y jerarquías, utilizando consultas rápidas predefinidas



On-Line Analytical Processing (OLAP)

Idea básica: los usuarios deben poder manipular los modelos de datos organizacionales a través de muchas dimensiones para comprender que se está ocurriendo.

- Los datos utilizados en OLAP deberían estar en la forma de un cubo multidimensional.



Herramienta Multidimensional Especializada

- **Beneficios:**
 - Acceso rápido a grandes volúmenes de datos
 - Bibliotecas extensas de funciones complejas de análisis
 - Capacidades de modelado y predicción
 - Puede acceder a las estructuras de bases de datos multidimensionales y relacionales

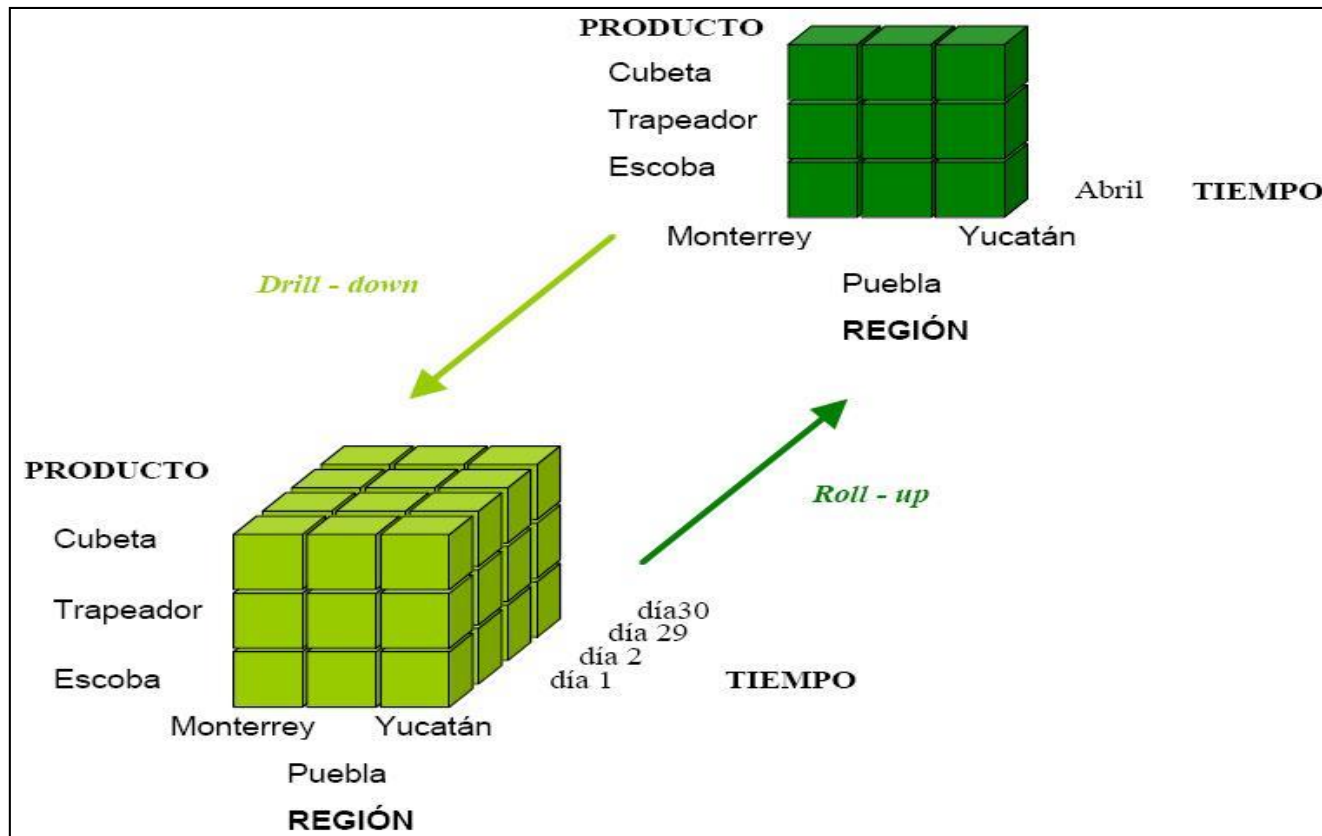
Arquitecturas OLAP

- **OLAP Relacional (ROLAP)**
 - Usa un esquema relacional para manejar la navegación y administrar los datos consolidados
 - Incluye agregación
 - Gran escalabilidad
- **OLAP Multidimensional (MOLAP)**
 - Almacenamiento con técnicas multidimensionales
 - Acceso rápido a datos pre-calculados previamente
- **OLAP Híbrido (HOLAP)**
 - Bajo nivel MOLAP, Alto nivel ROLAP
- **Motores de BD especializados**
 - Manejan consultas especializadas (como las de SQL) con esquemas estrella o copo de nieve

Operaciones clásicas OLAP

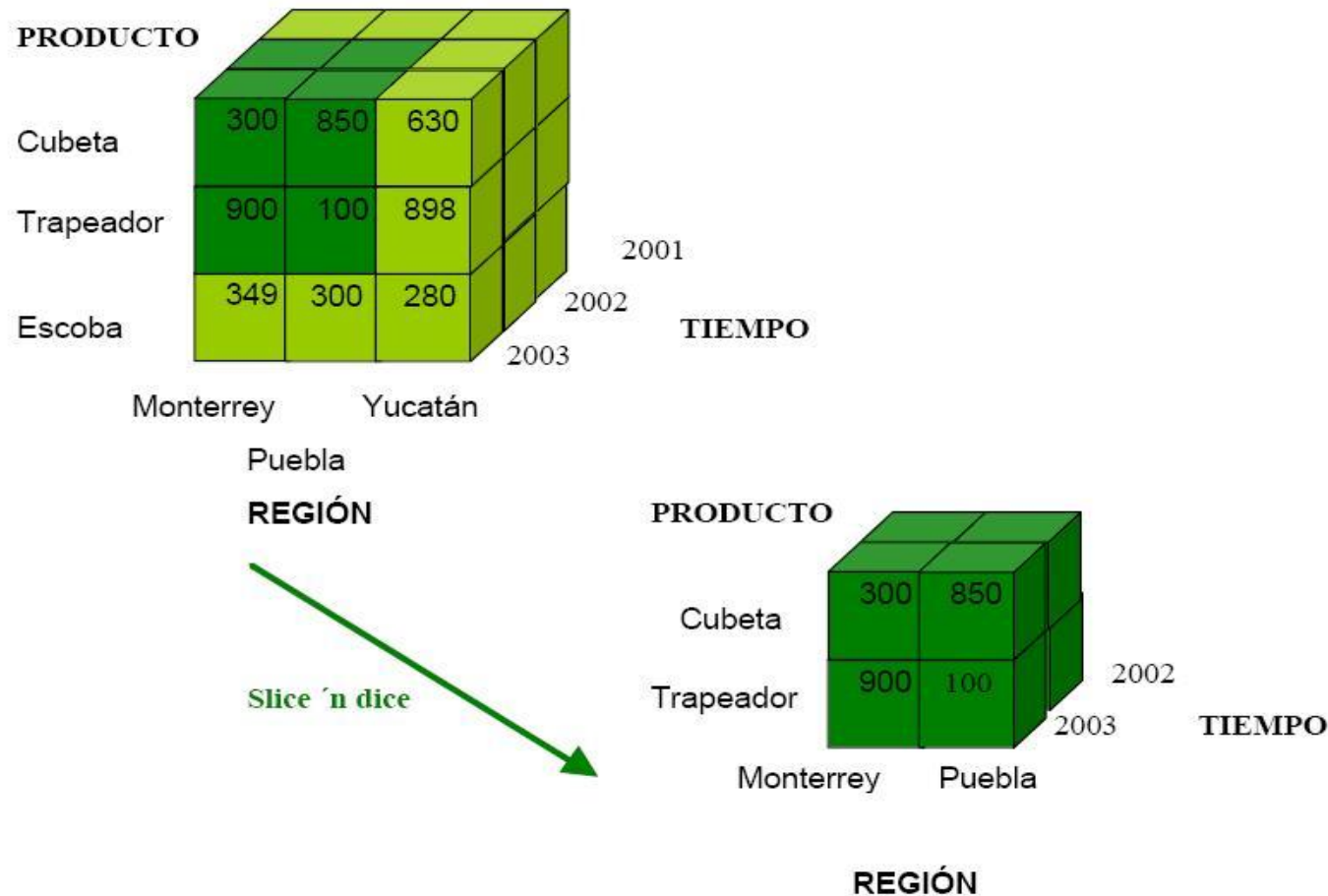
Roll up (drill-up): agrega medidas que van de un nivel N_i a un nivel mas general N_j de una dimensión.

Drill down (roll down): es la operación inversa. A partir de un nivel superior este operador permitir bajar de nivel.



Operaciones clásicas OLAP

Slice and dice: permite restringir los valores asociados a una o varias dimensiones del cubo, es decir, toma un subconjunto de dimensiones y de niveles seleccionados del DW.



Otras operaciones

drill across

navegar a través de más de una tabla de hechos

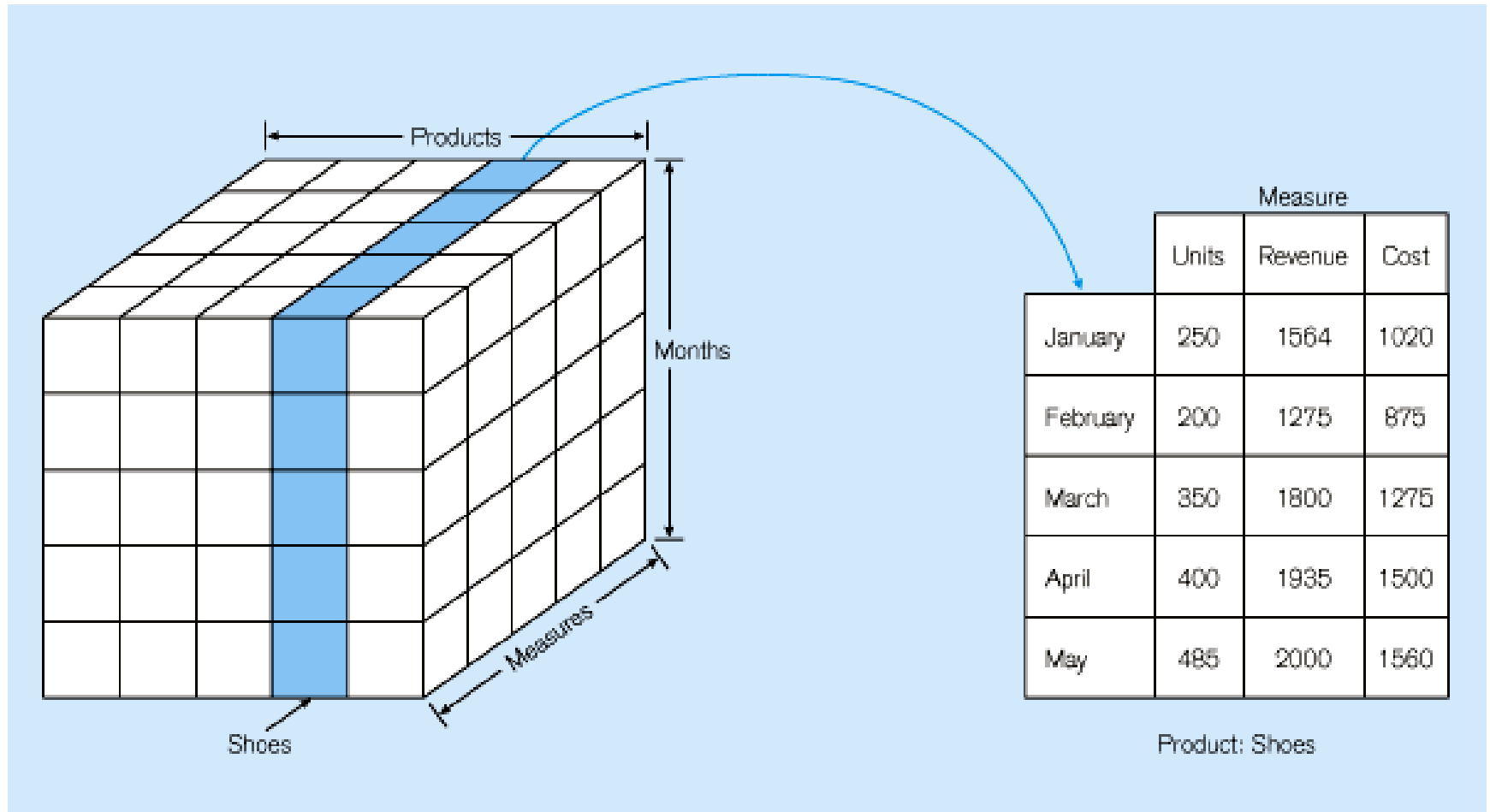
drill through

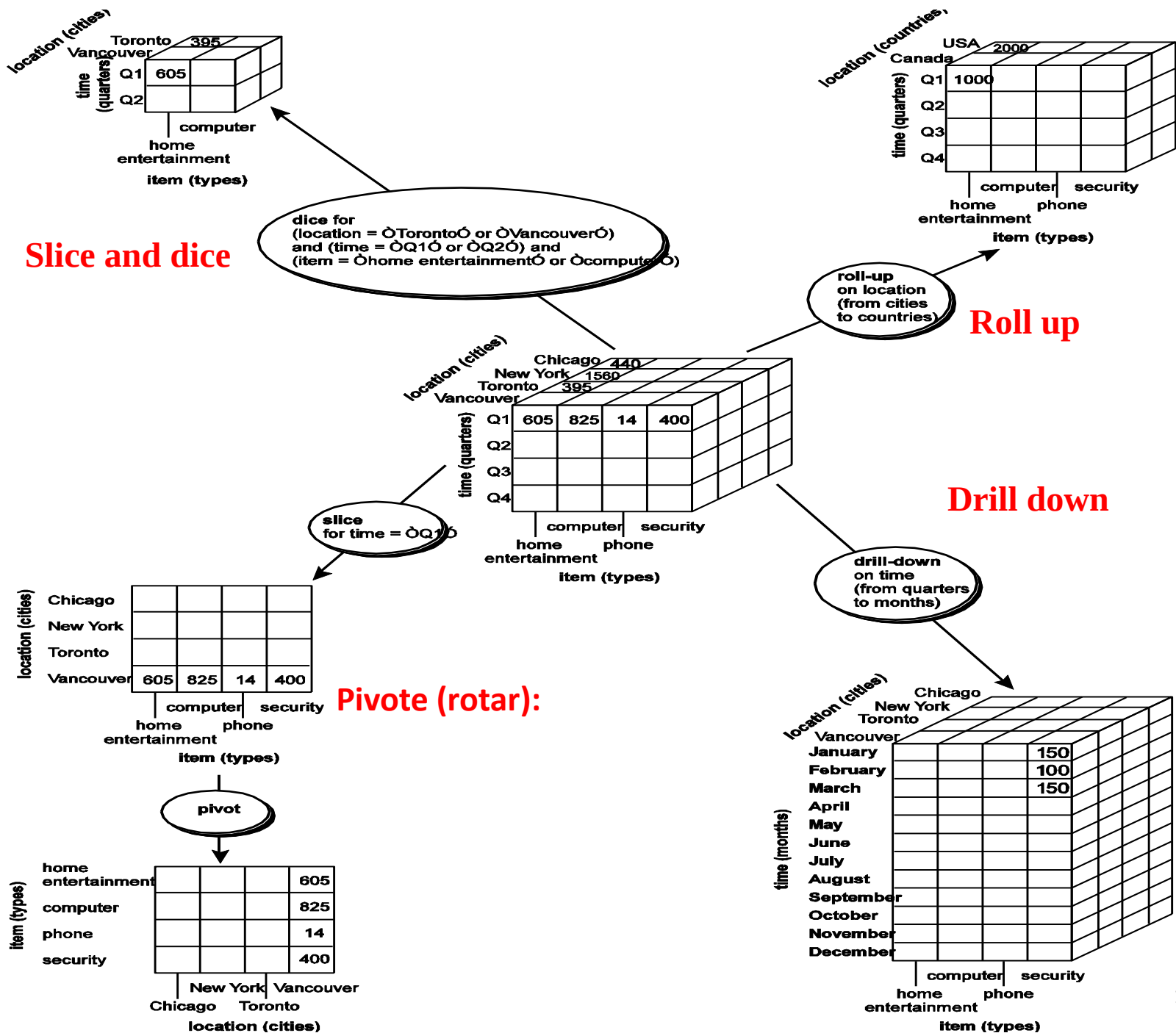
navegar a través del nivel inferior del cubo a tablas relacionales

Pivote (rotar)

Rotar el cubo

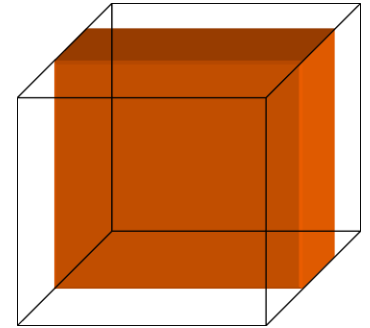
Cortando/rebanando un cubo de datos



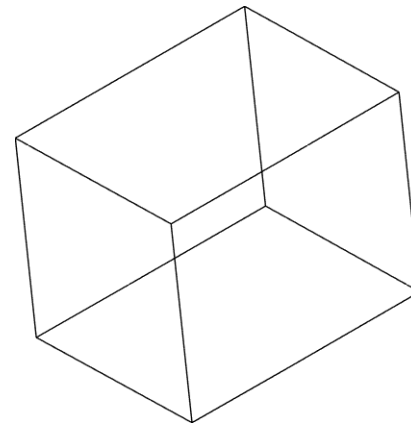


Resumen Operaciones clásicas OLAP

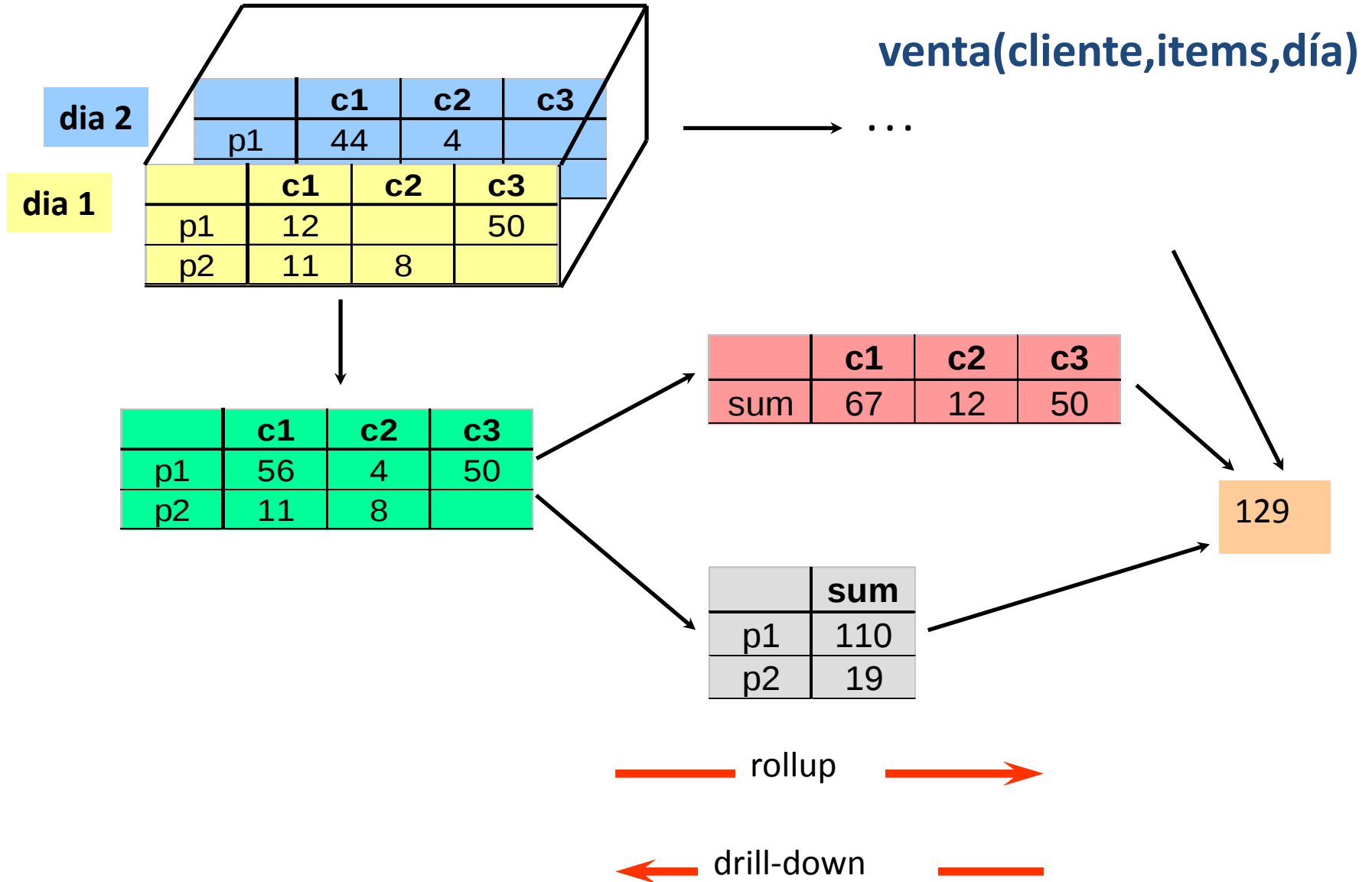
- *Rollup*: decrease nivel de detalle
- *Drill-down*: aumenta nivel de detalle
- *Slice-and-dice*: selección y proyección



- *Pivot*: re-orienta vista multidimensional

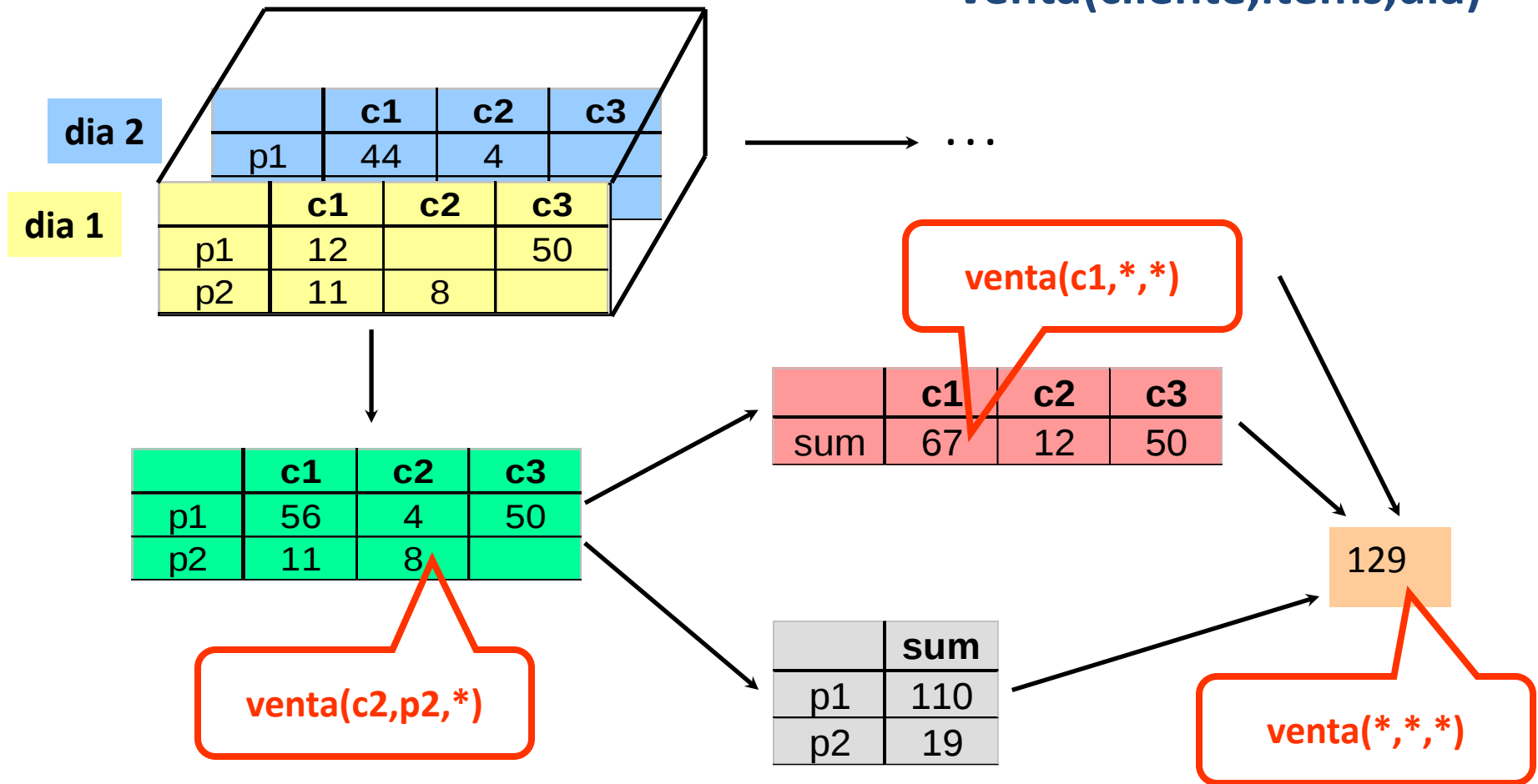


Agregación en Cubos



Agregación en Cubos

venta(cliente,item,día)



Cara de Cubos

venta(cliente,item,día)

The diagram illustrates three stacked tables representing sales data for different days. The tables are color-coded: blue for the top table, light blue for the middle table, and yellow for the bottom table. A red callout points to the value 19 in the middle table, labeled **venta(*,p2,*)**.

dia 2		dia 2				
	*		c1	c2	c3	*
		p1	56	4	50	110
		p2	11	8		19
dia 2			c1	c2	c3	*
		p1	44	4		48
dia 1			c1	c2	c3	*
		p1	12		50	62
		p2	11	8		19
		*	23	8	50	81

venta(*,p2,*)

Agregación usando jerarquía

venta(cliente,items,día)

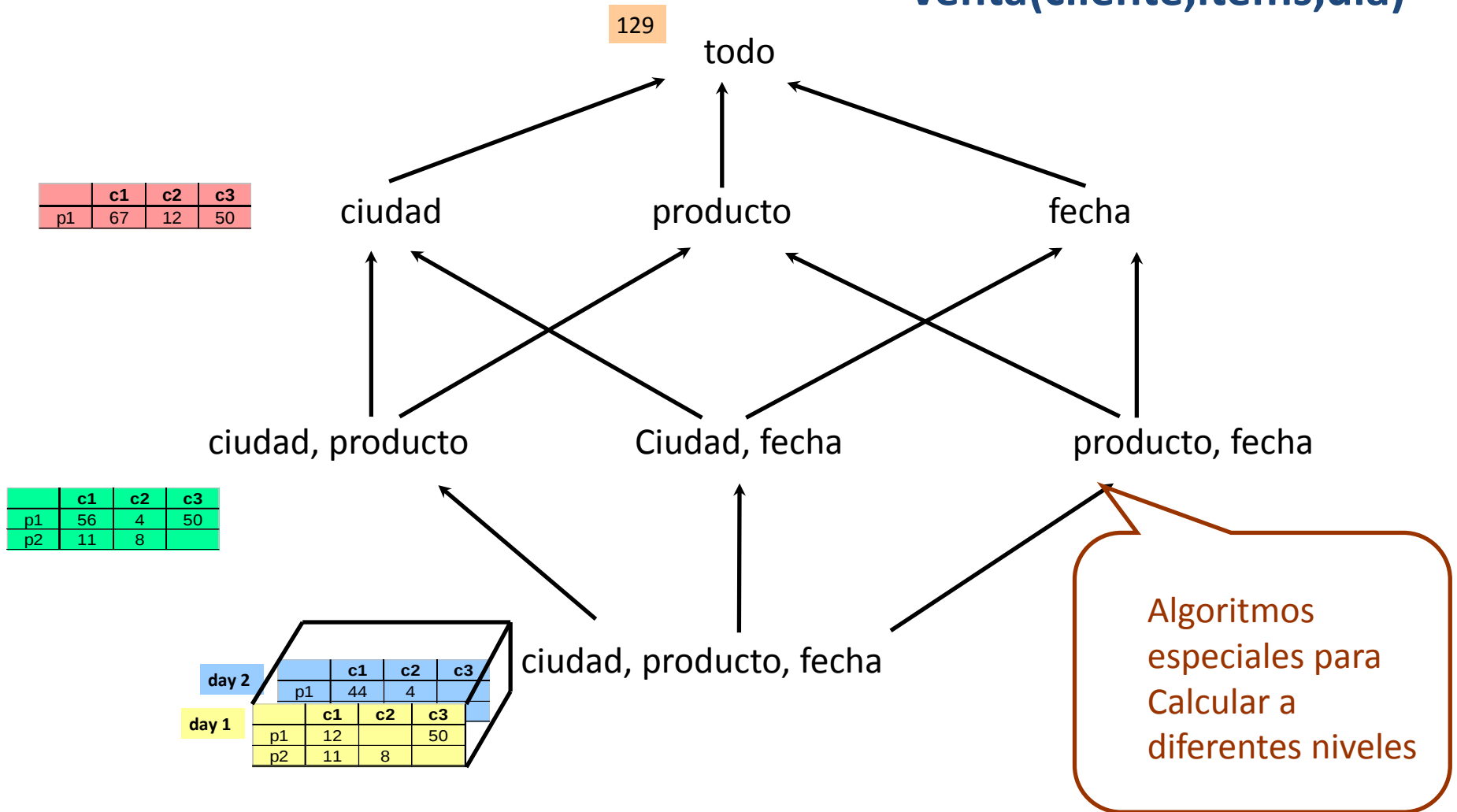
	region A	region B
p1	56	54
p2	11	8

cliente
|
region
|
país

(cliente c1 en Region A;
cliente c2, c3 en Region B)

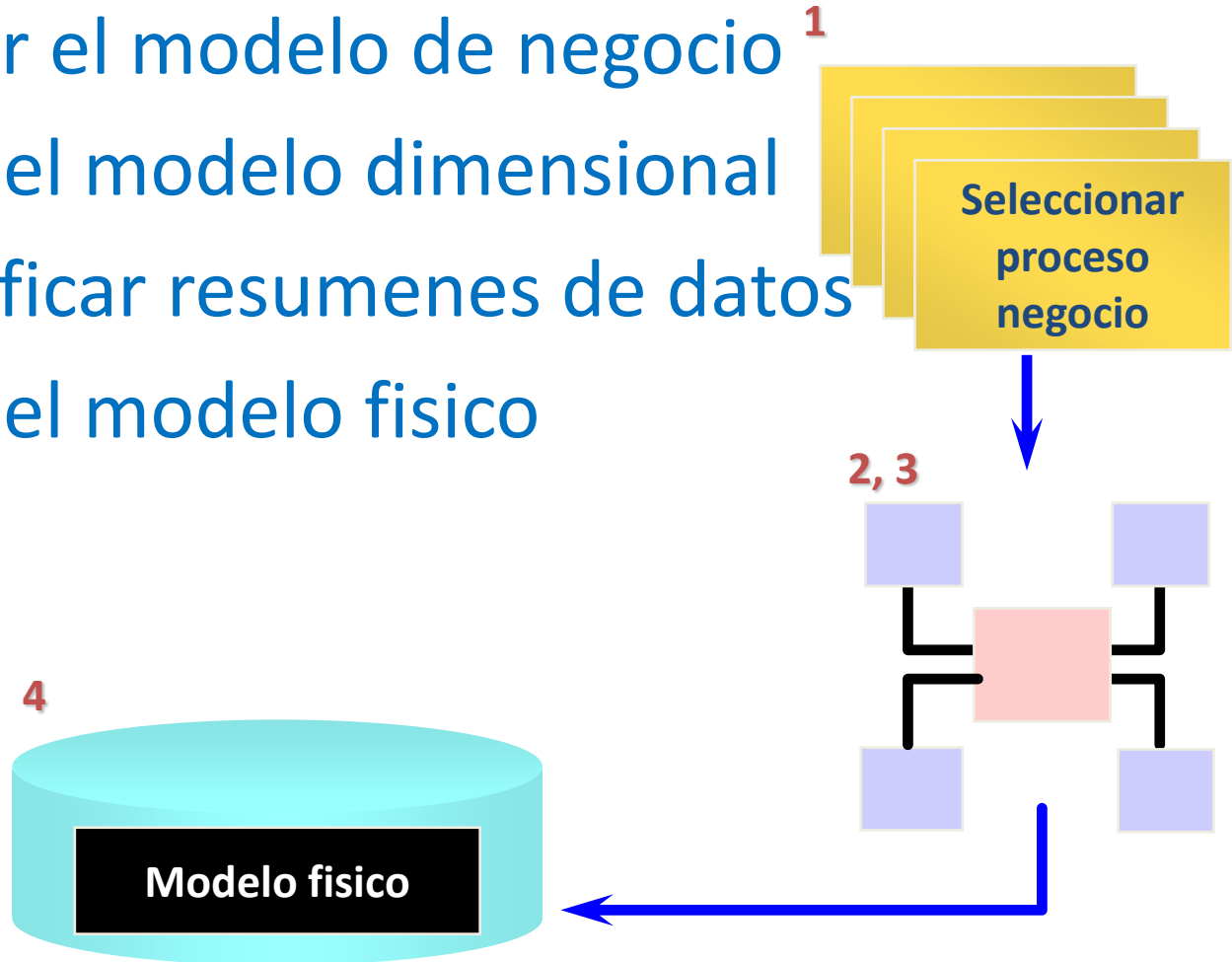
Agregación en Cubos

venta(cliente,item,día)



Modelado en Data Warehouse

1. Definir el modelo de negocio
2. Crear el modelo dimensional
3. Identificar resúmenes de datos
4. Crear el modelo físico



Crear Modelo Dimensional

- Seleccionar una entidad para comenzar a armar tabla de hechos
- Determinar granularidad
- Identificar claves operacionales para tabla de hechos
- Buscar jerarquías
- Añadir dimensiones
- Caracterizar los atributos de las dimensiones

Granularidad (unidad de análisis)

Determina lo que representa cada registro de la tabla de hechos: el nivel de detalles.

- Ejemplos
 - Puntos en el tiempo
 - Lineas en un documento
- Depende del proyecto de IN

Crear Modelo Dimensional

- Identificar tablas de hechos
 - Traducir medidas pregunta madre en tablas de hechos
 - Analizar las fuentes de datos para las medidas
 - Identificar tablas de dimensiones
- Enlazar tabla de hechos con las tablas de dimensiones
- Crear vistas para los usuarios (operaciones OLAP)

Identificar resúmenes de datos

- Proporciona un acceso rápido a datos pre-calculados
- Reduce el uso de E/S, CPU y memoria
- Se calcula desde las fuentes de datos y otros resúmenes pre-calculados
- Por lo general, se guardan en las tablas de hechos

3. Identificar resúmenes de datos

- Promedio
- Máximo
- Total
- Porcentaje

	Unidad	Venta	Tienda
Producto A			
Total			
Producto B			
Total			
Producto C			
Total			



Inconvenientes

- El almacén de datos no suele ser estático.
Los costos de mantenimiento son elevados.
- Ante una petición de información estos pueden devolver una información sub-óptima, que también supone una pérdida para la organización.
- Se pueden quedar obsoletos relativamente pronto