



UNIVERSIDAD
DE LOS ANDES
MÉRIDA VENEZUELA

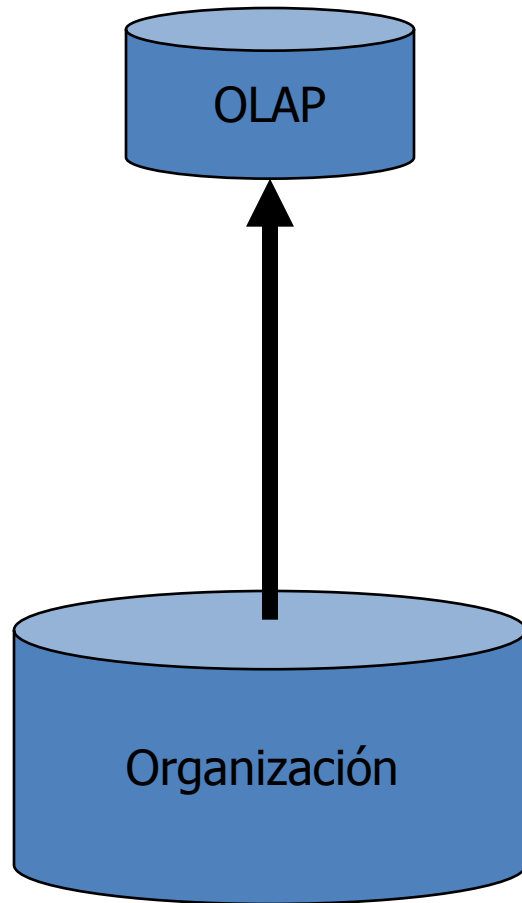


Modelado de Datos

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

Modelado de Datos y de Información

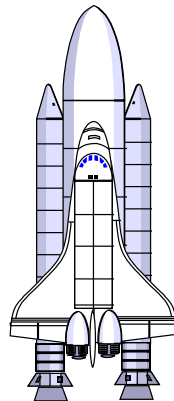
Usuarios tienen diferentes vistas de los datos



Turistas: Navegar por la información recolectada



Agricultores: información de los caminos a los datos, etc.



Exploradores: Busca cosas desconocidos que se esconden en los datos detallados

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:

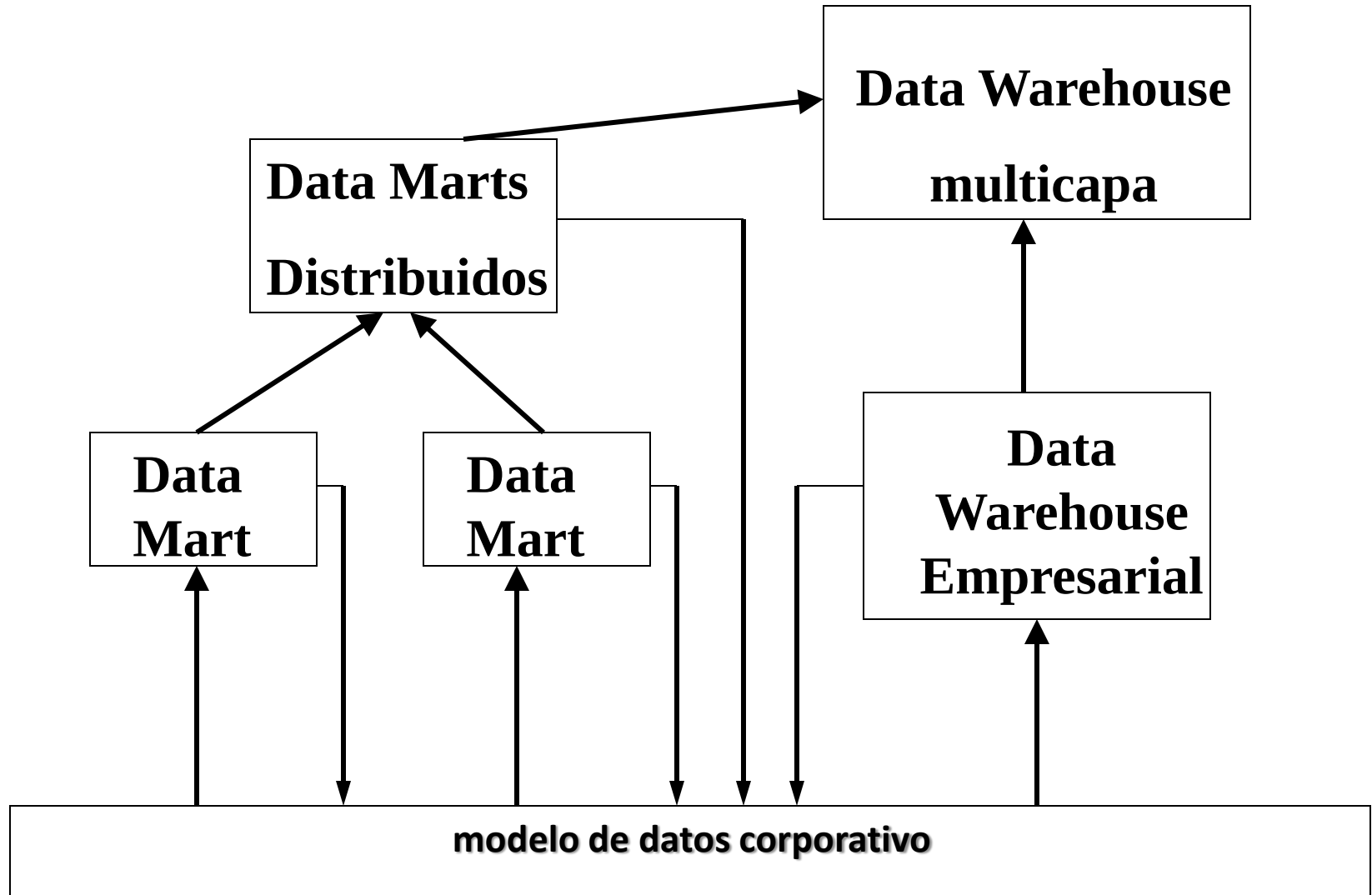
- **Situación actual:** Cualquier solución propuesta de data warehouse debe estar muy orientada por las necesidades del negocio, debe ser compatible con la arquitectura técnica existente y planeada de la compañía.
- **Tipo y características del negocio:** Tener el conocimiento exacto sobre el tipo de 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.

Un enfoque de desarrollo de Data Warehouse



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**

- Recopilación de datos útiles sobre un tema

Ejemplo: ventas

- Sintetizar una visión única de los temas a analizar

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

- 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

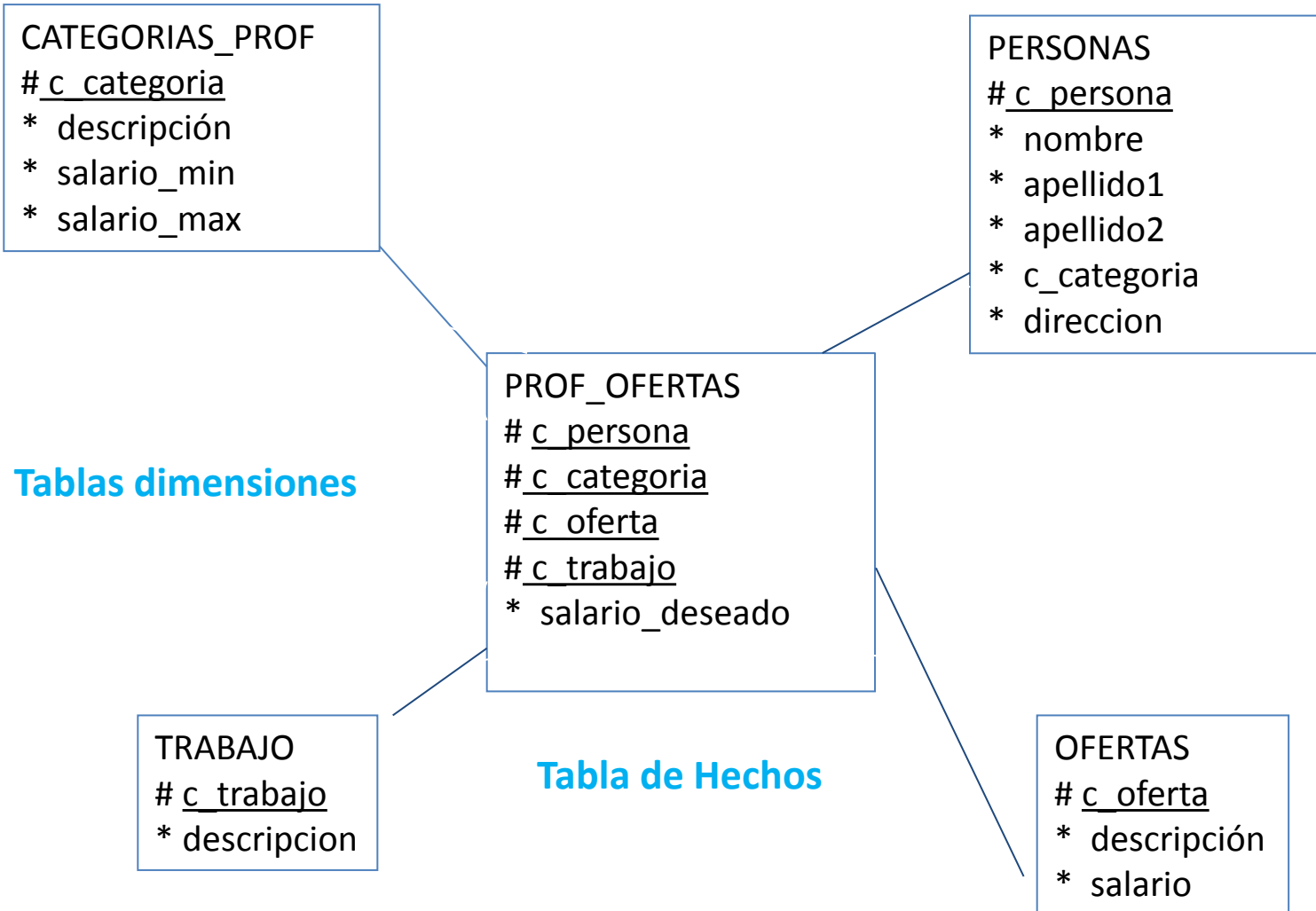
- Dividir en varias tablas unidas por referencias

4. Integrar todo

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



Esquema en estrella



Esquema en estrella

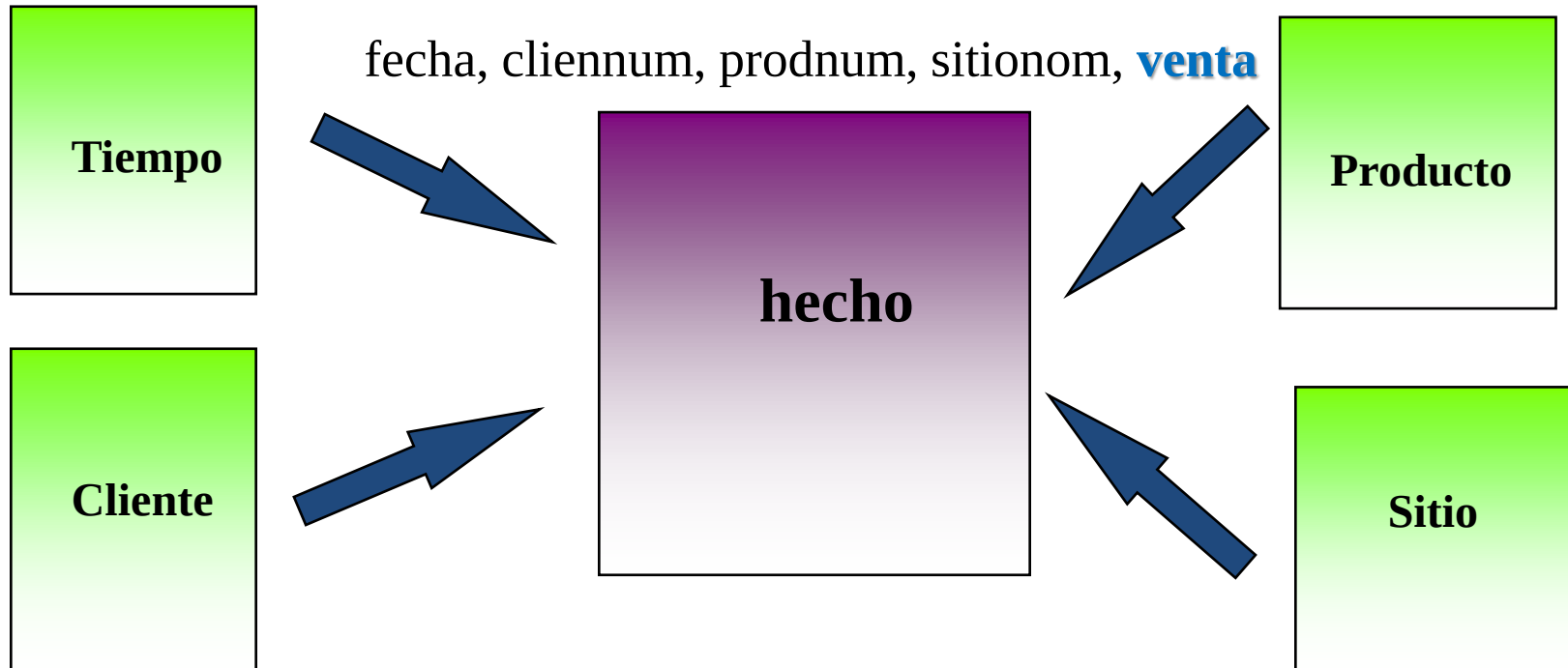
- Modelado relacional actual no satisface las necesidades actuales
- Representaciones de datos multidimensionales
- Optimizar las operaciones de consulta de datos en lugar de las operaciones de actualización de datos
- Los datos no representan una transacción del negocio en particular.
- Los datos pueden 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 (4-15= dimensiones
 - Las dimensiones se de-normalizan
- Una tabla de hechos en el medio conectado a un conjunto de tablas de dimensiones

Esquema en estrella

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



Esquema en estrella: Componentes

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

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
- Recibe enlace dimensiones
- 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

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

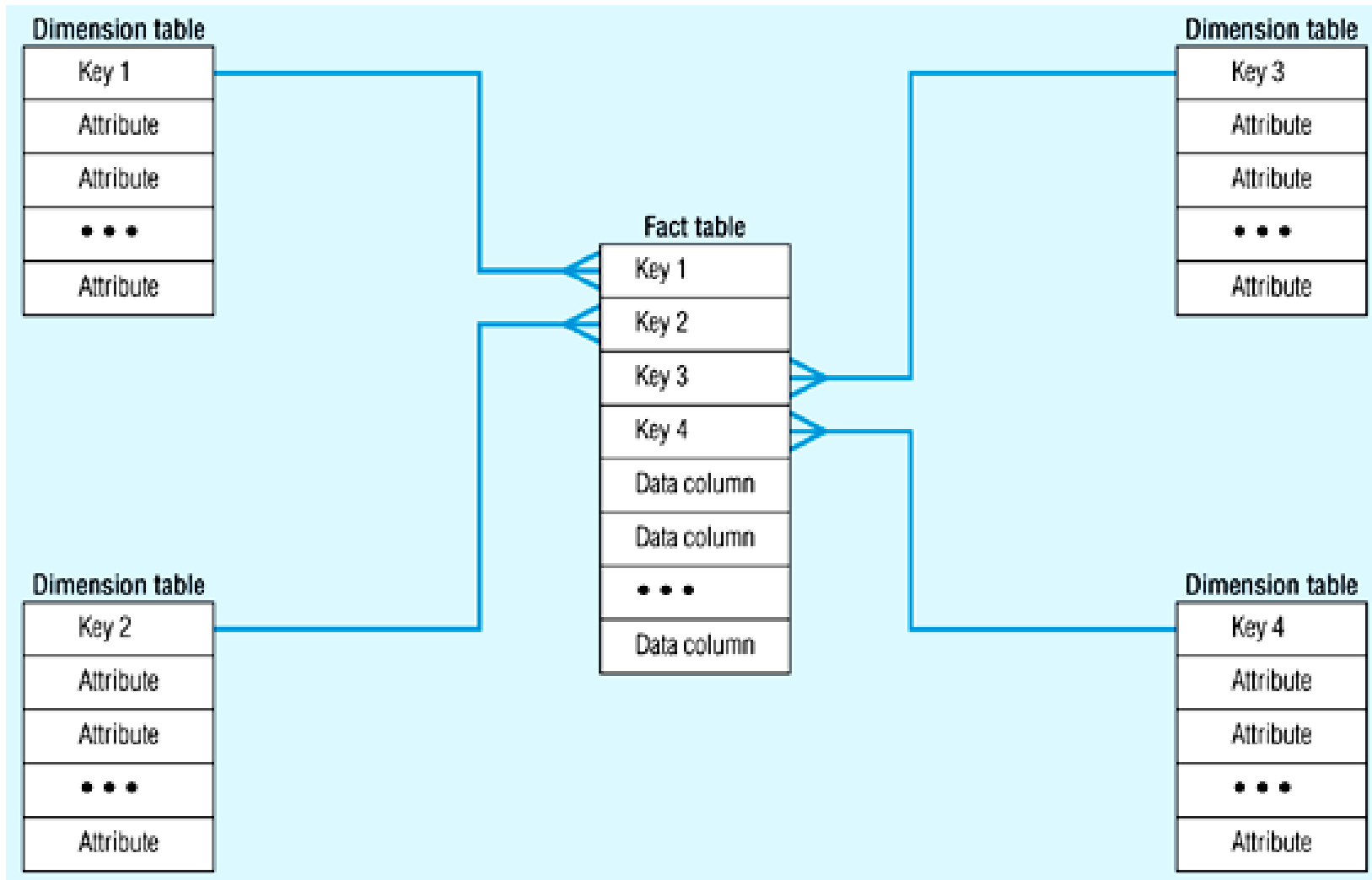
Esquema en estrella: Tabla de Dimensiones

- 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

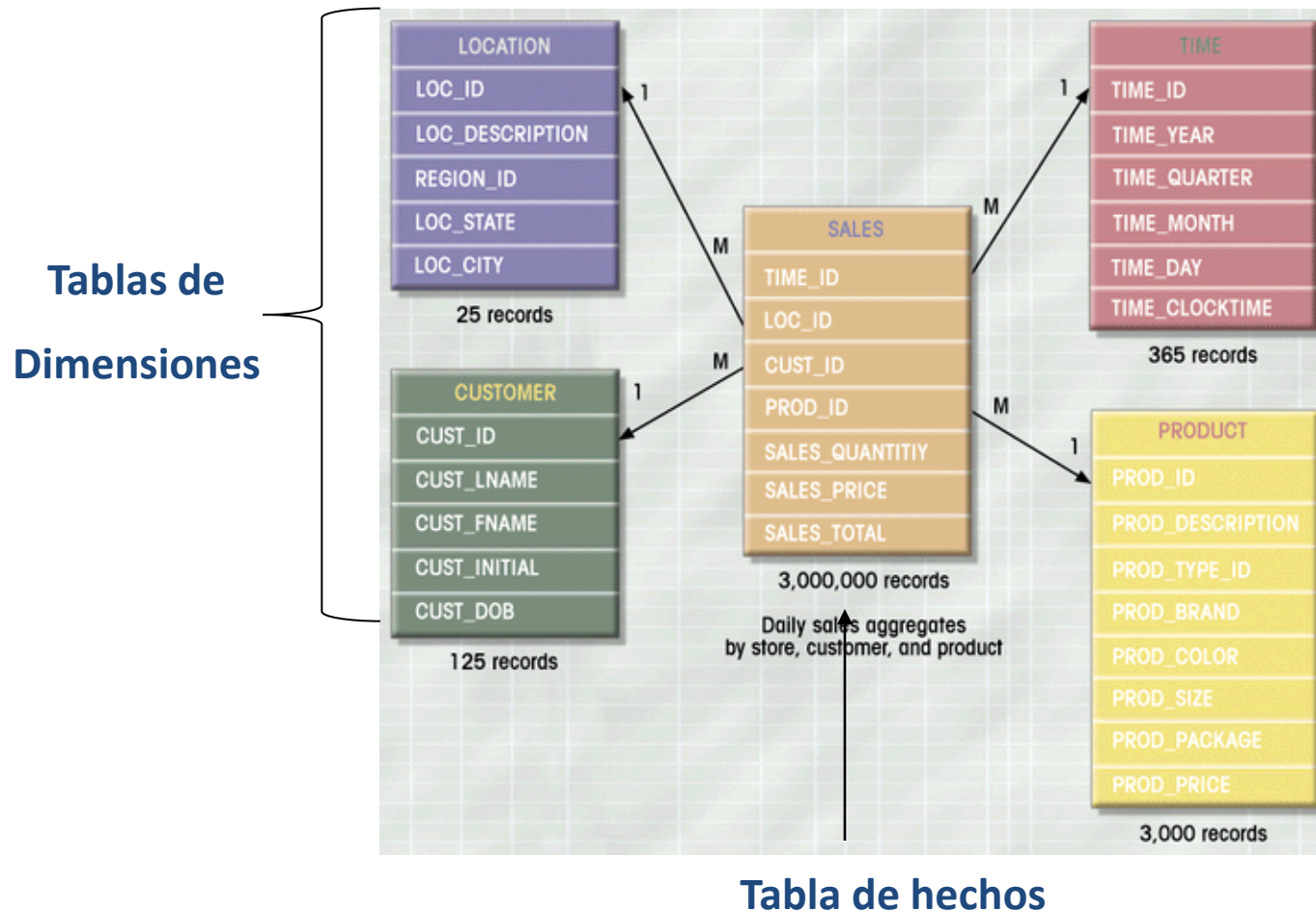
Esquema en estrella: Atributos

- 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)

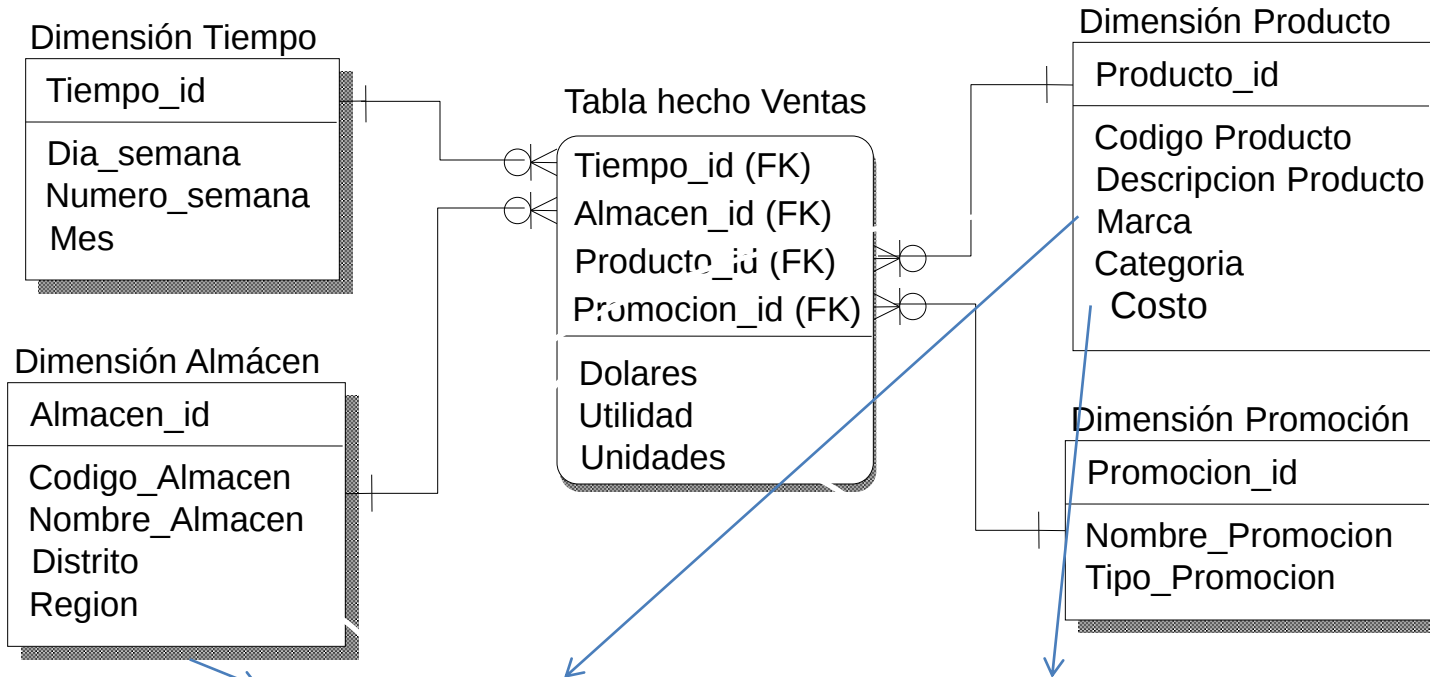
Esquema en estrella: Atributos



Ejemplo de esquema en estrella para ventas



Ejemplo de esquema en estrella para Ventas

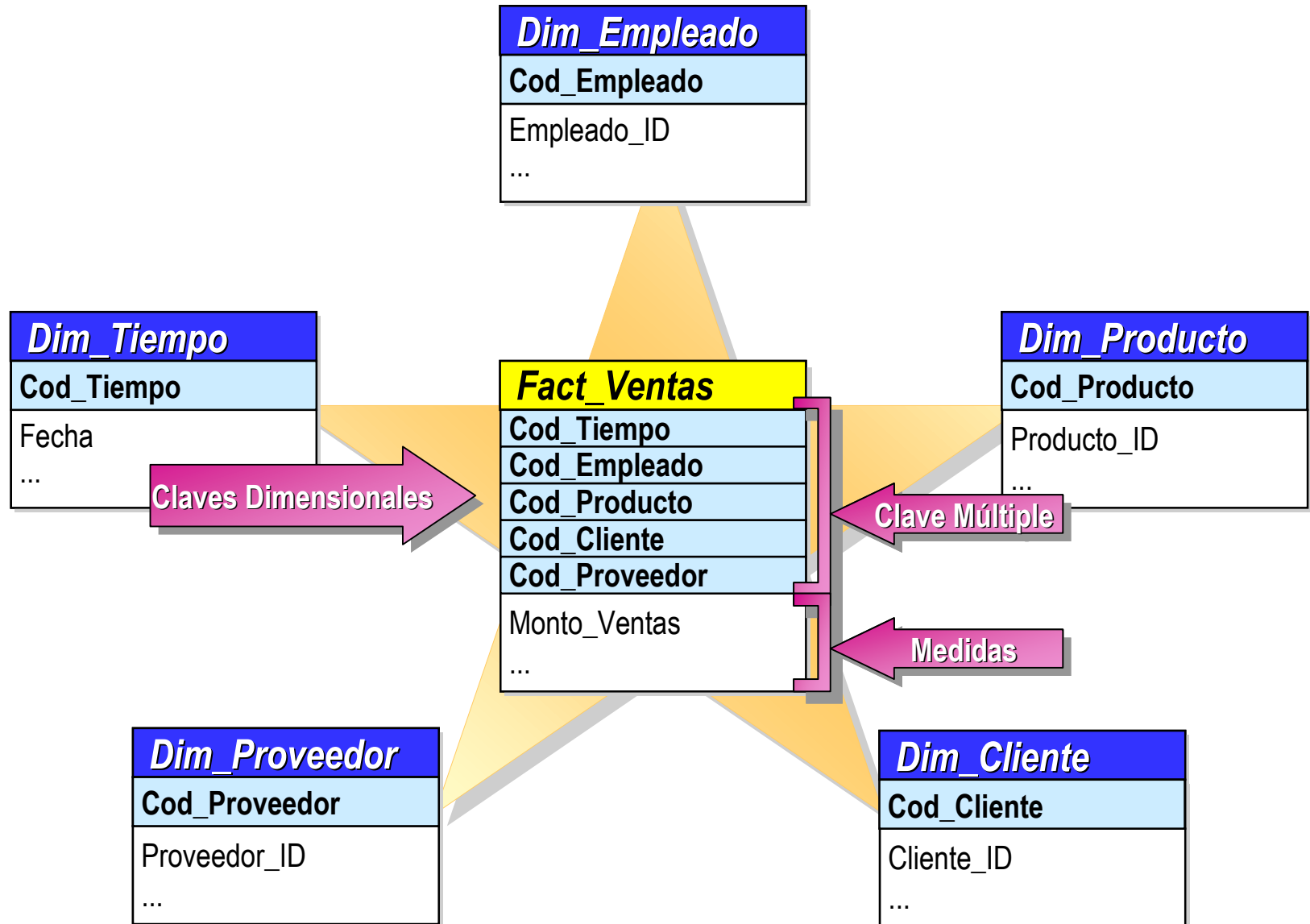


Distrito	Marca	Total Dolares	Total Costo	Utilidad	Unidades
Atherton	Clean Fast	\$ 1,233	\$ 1,058	\$ 175	10
Atherton	More Power	\$ 2,239	\$ 2,200	\$ 39	2
Atherton	Zippy	\$ 848	\$ 650	\$ 198	4
Belmont	Clean Fast	\$ 2,097	\$ 1,848	\$ 249	6
Belmont	More Power	\$ 2,428	\$ 2,350	\$ 78	3
Belmont	Zippy	\$ 633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

Ejemplo de esquema en estrella para ventas



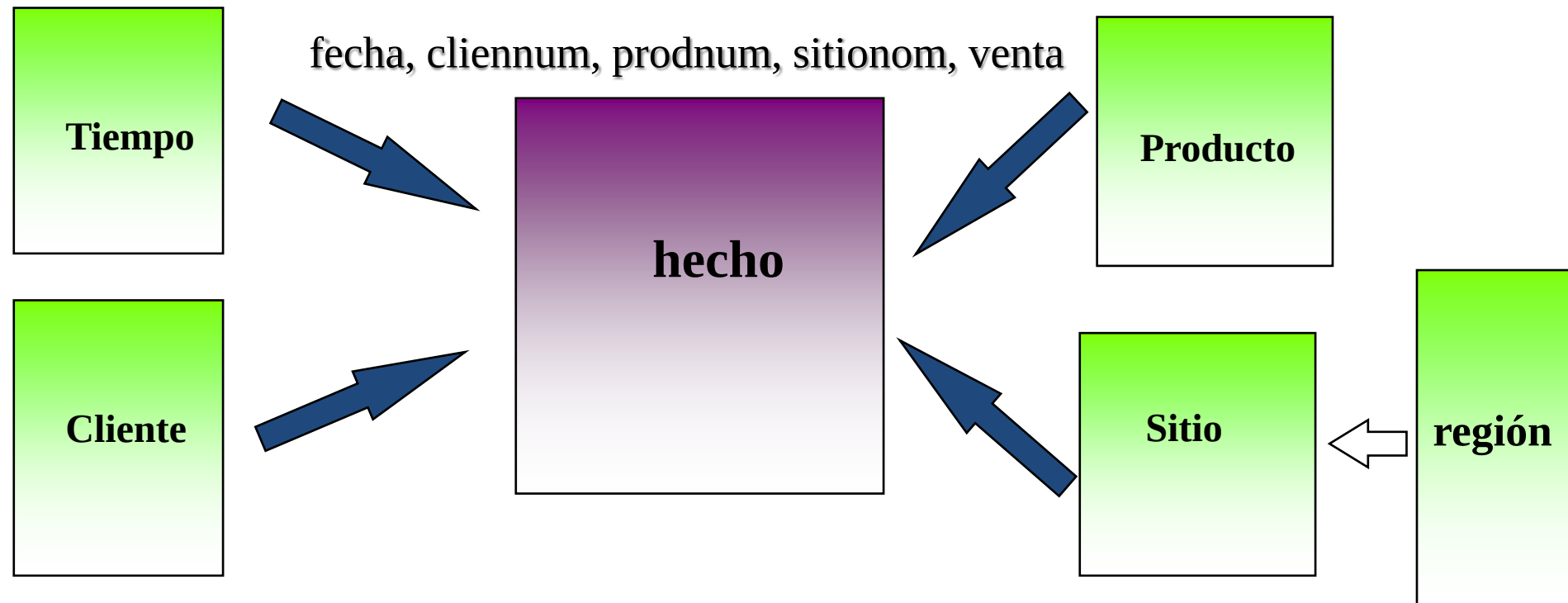
Esquema Copo de nieve

- Un refinamiento del esquema en estrella donde alguna jerarquía dimensional se despliega en un conjunto de tablas de dimensiones más pequeñas,
- Forma similar a la del copo de nieve
 - 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

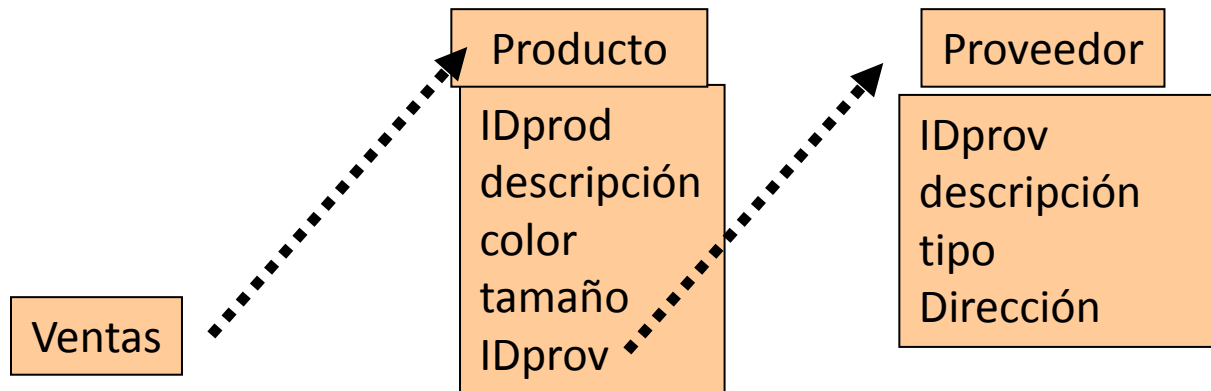
Representa jerarquía dimensional



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

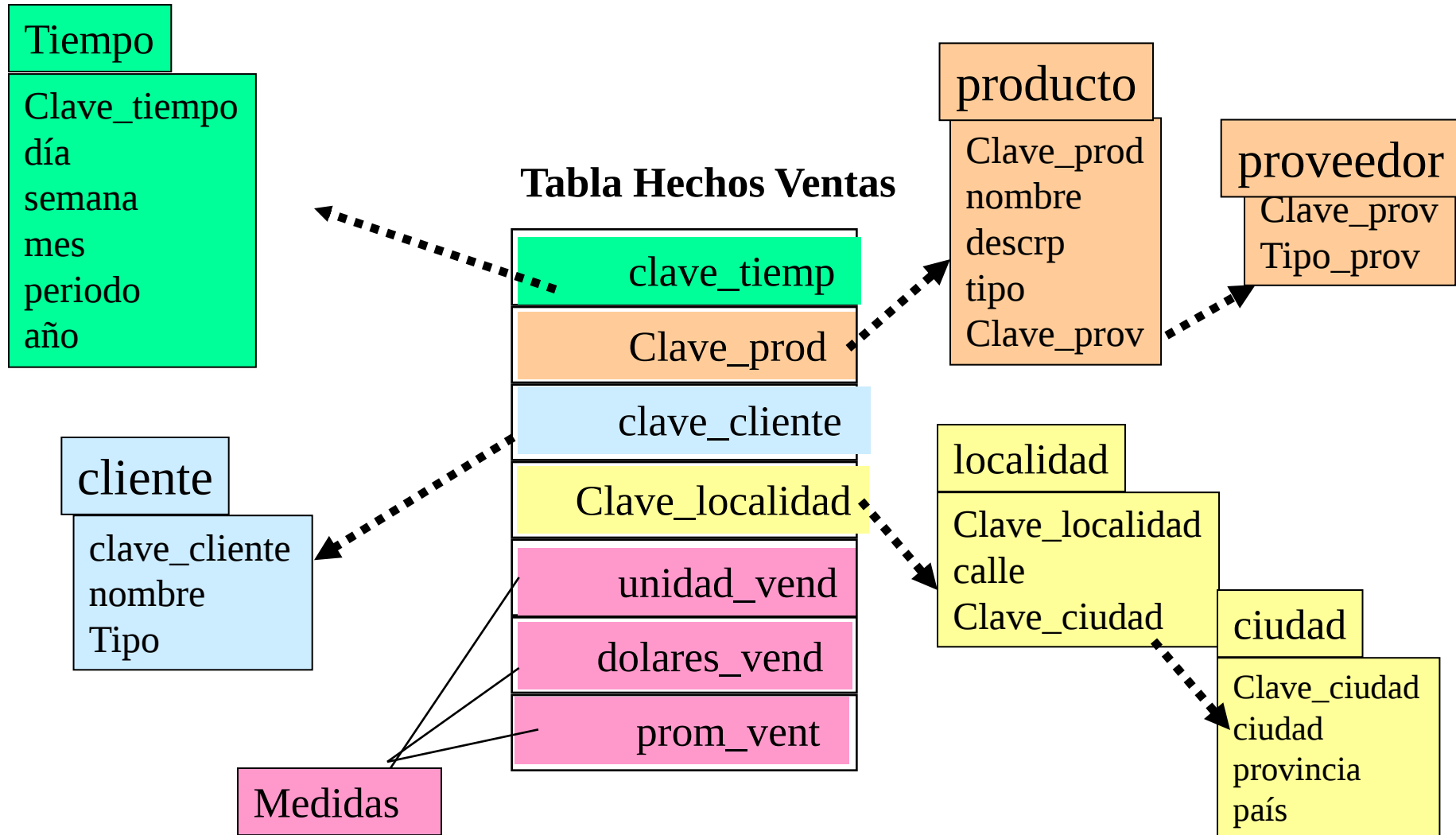
Descripción Región

Gerente Región..

Tabla hecho almacén

Clave Almacen
Clave producto
Clave Periodo
Monto
Unidades
Utilidad

Esquema Copo de nieve

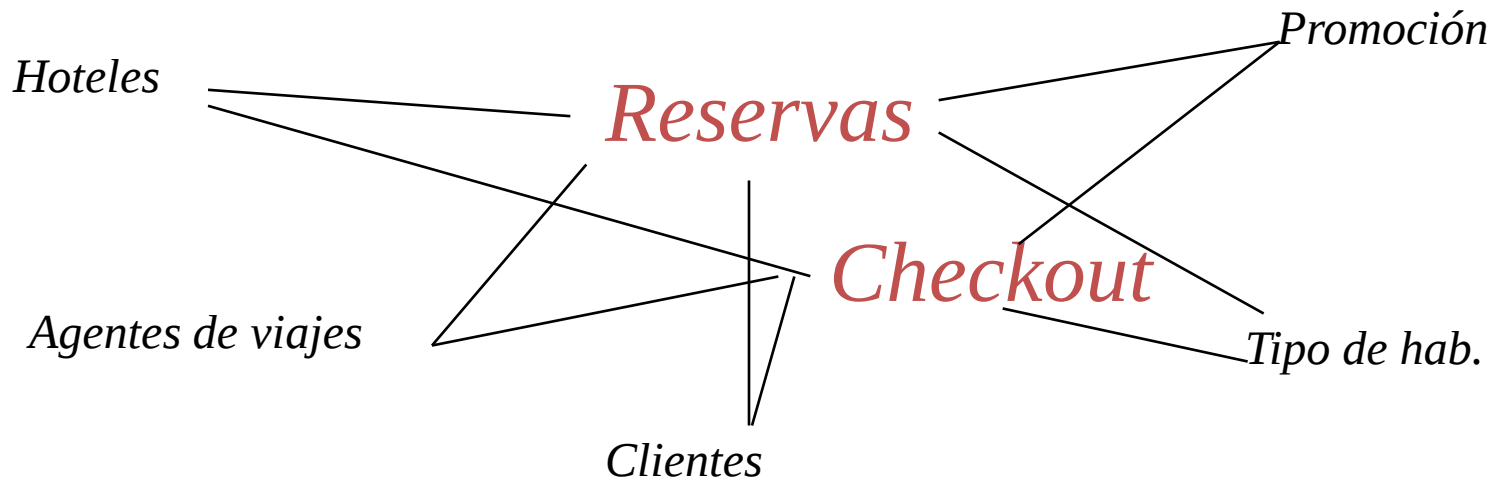


Esquema de Constelación de hechos

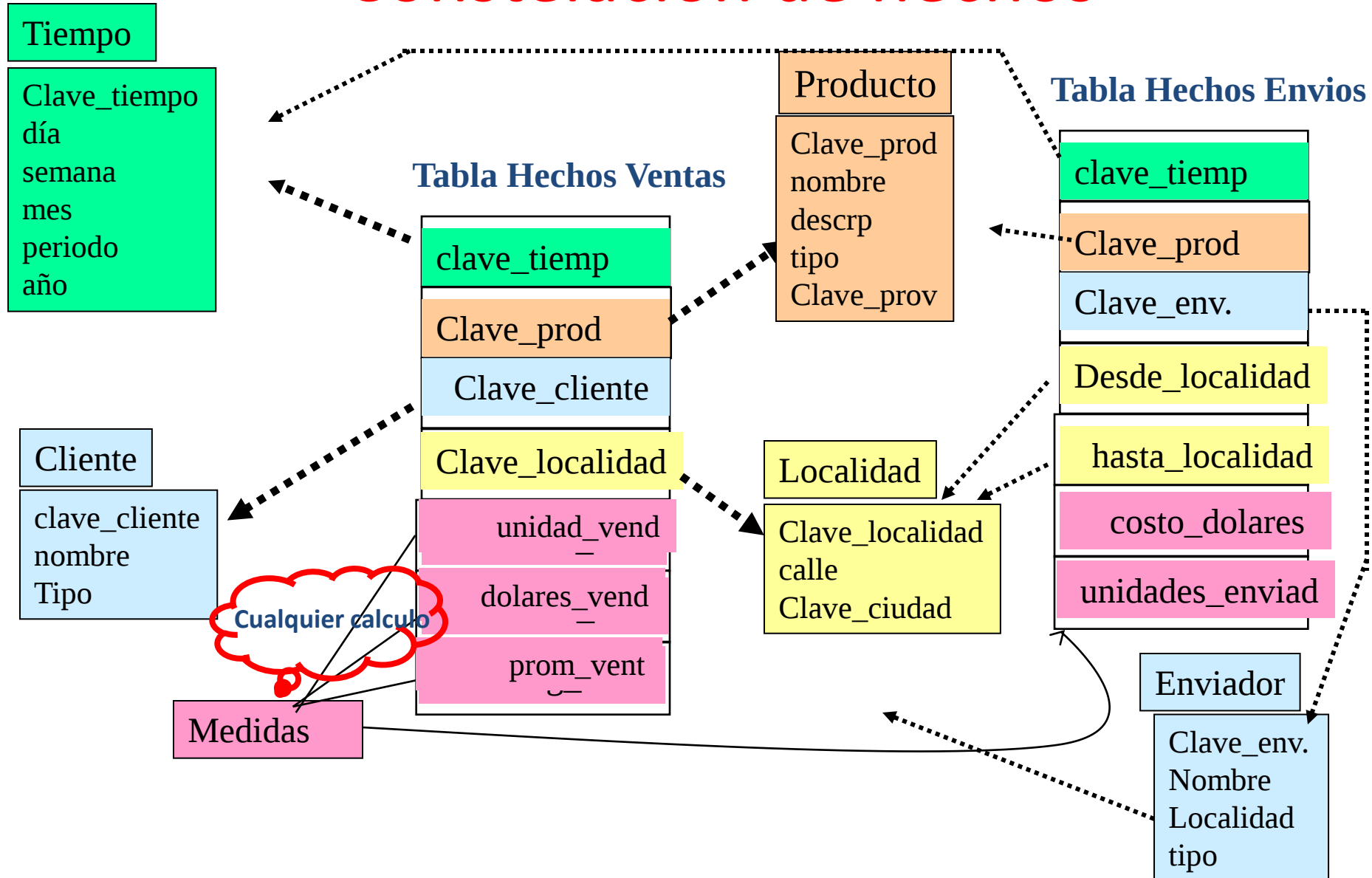
- Varias tablas de hechos **comparten** tablas de dimensiones,
- vistos como una **colección de estrellas**, por lo tanto, llamados **esquema de galaxia o constelación de hecho**

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

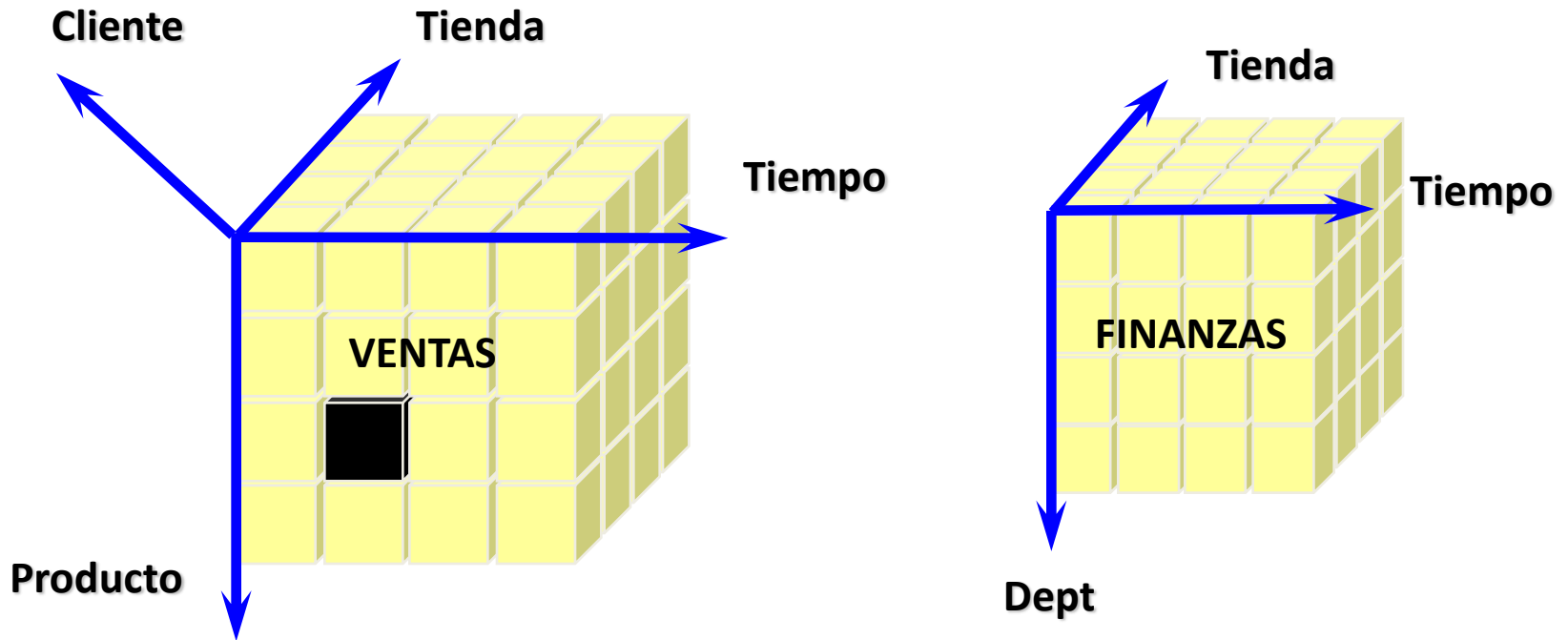
Un cubo n-D se llama un paralelepípedo.

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

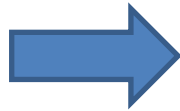
Modelo BD multidimensional



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
.....		
.....		

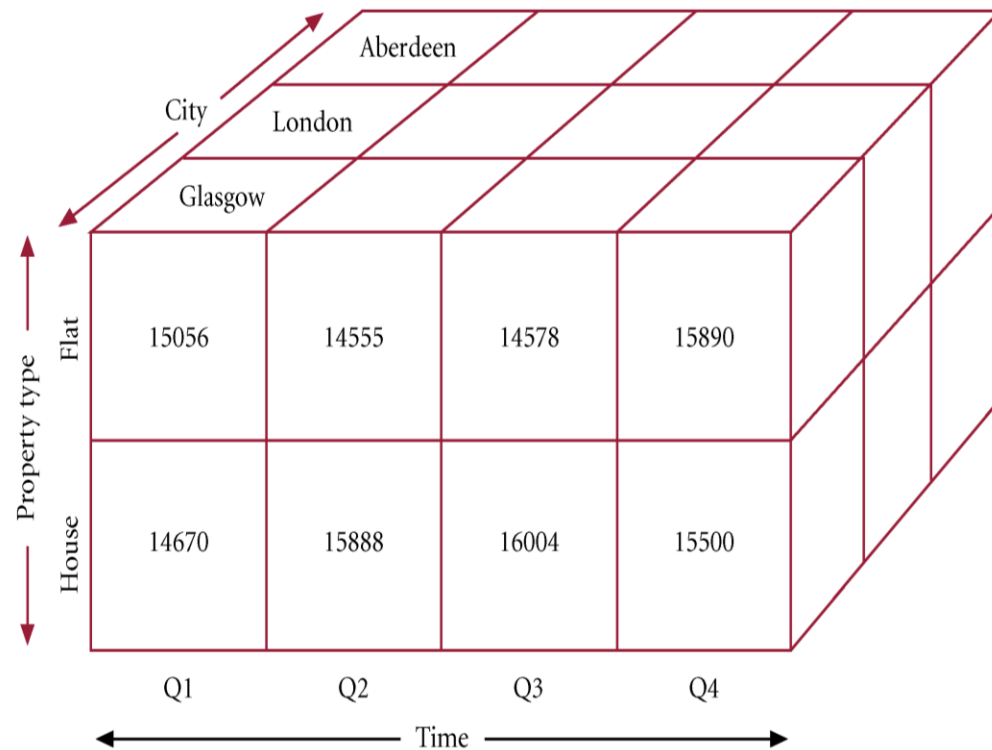


City				
City Quarter	Glasgow	London	Aberdeen	
Q1	29726	43555	53210	
Q2	30443	48244	34567	
Q3	30582	56222	45677	
Q4	31390	45632	50056	

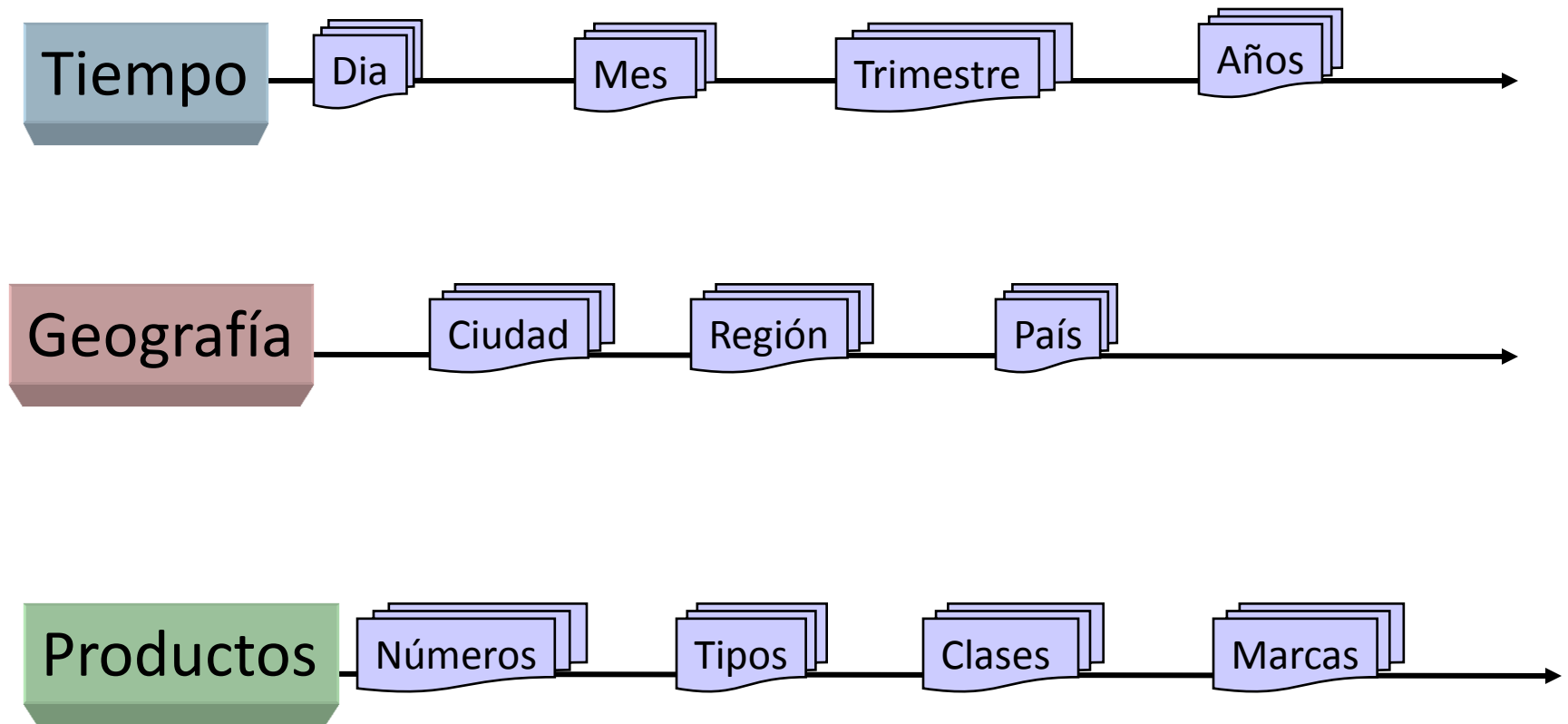
Tres dimensiones



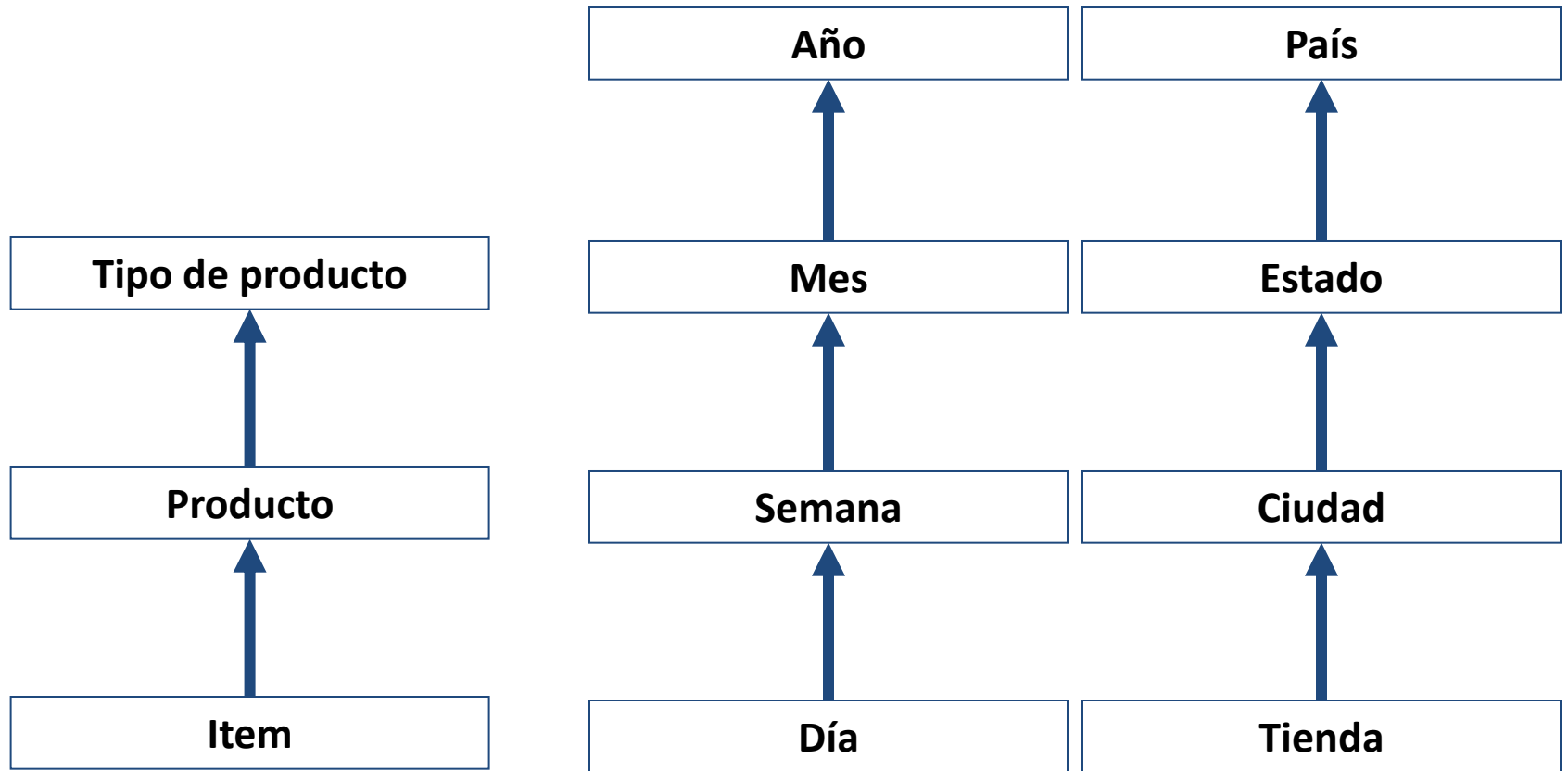
Property Type	City	Time	Total Revenue
Flat	Glasgow	Q1	15056
House	Glasgow	Q1	14670
Flat	Glasgow	Q2	14555
House	Glasgow	Q2	15888
Flat	Glasgow	Q3	14578
House	Glasgow	Q3	16004
Flat	Glasgow	Q4	15890
House	Glasgow	Q4	15500
Flat	London	Q1	19678
House	London	Q1	23877
Flat	London	Q2	19567
House	London	Q2	28677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



La granularidad de las dimensiones



Jerarquía Dimensional



Jerarquía Dimensional (localidad)

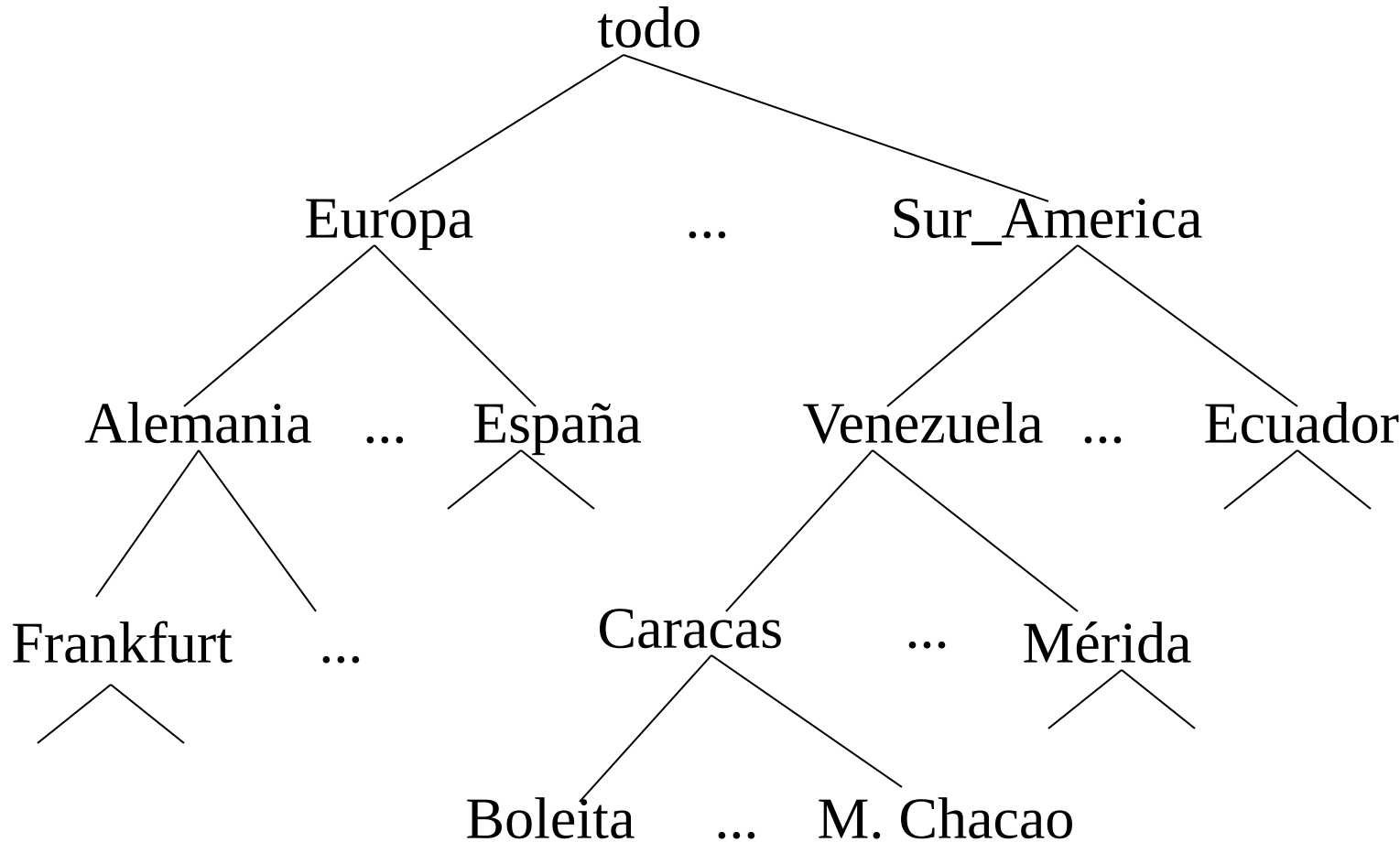
todo

región

país

ciudad

oficina



Jerarquía Dimensional

- jerarquía de esquema

día < mes < cuatrimestre < año

- Se pueden agrupar las jerarquías

{día 1 al 10}

{días} < 30

Las multidimensiones

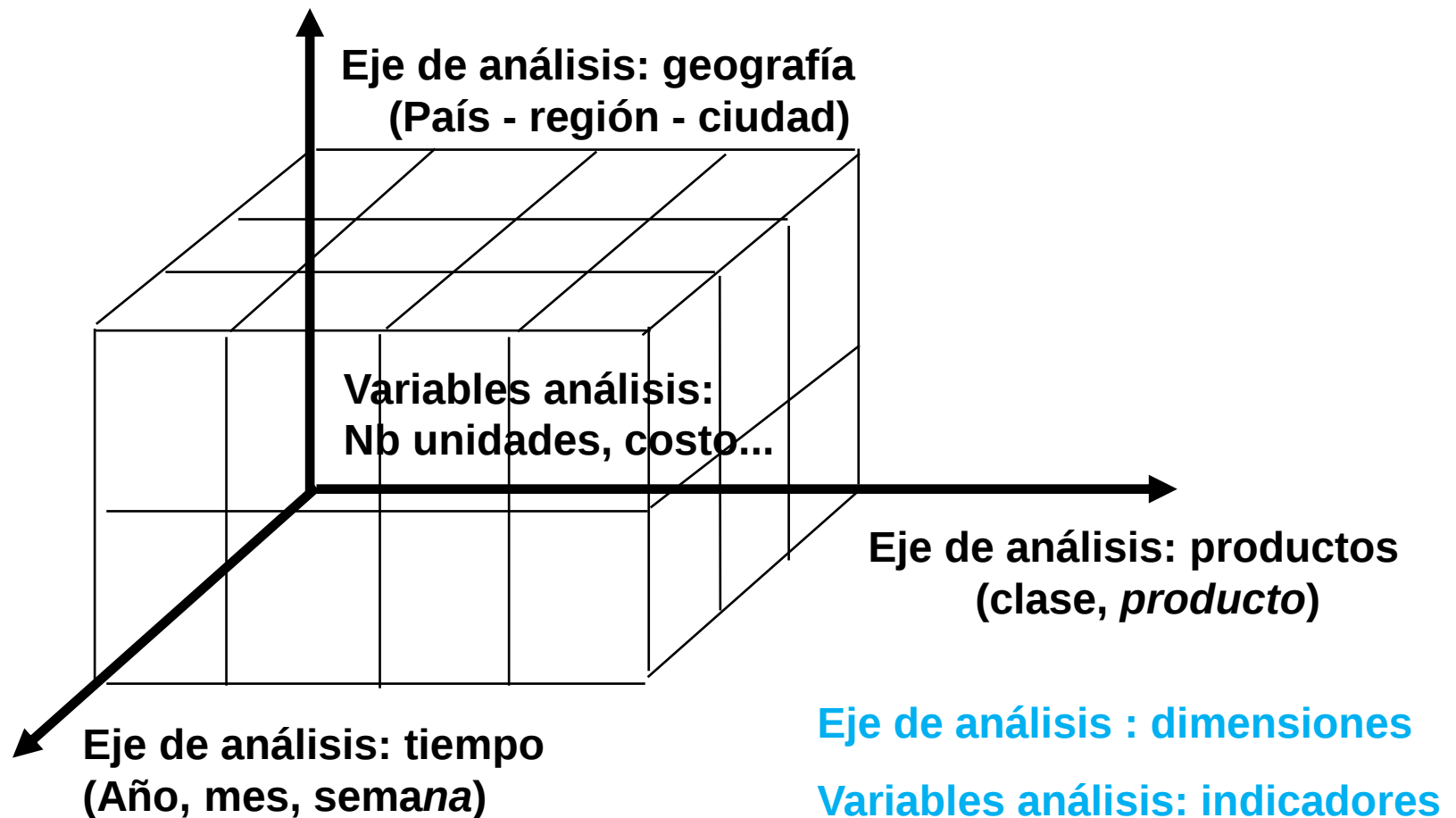
- Dimensiones:

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

- Indicadores:

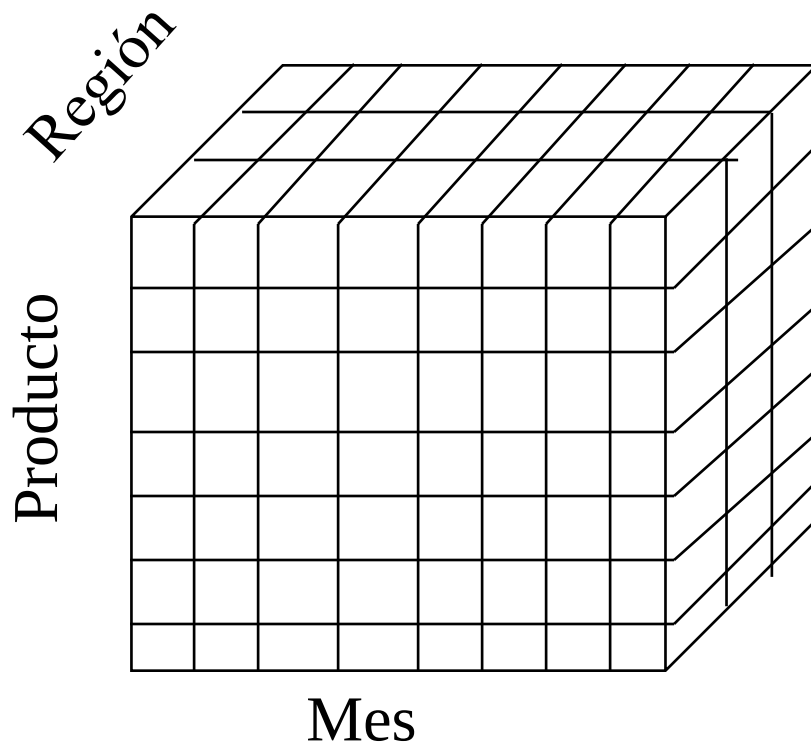
- Número de unidades vendidas
- Costo

Cubo de dato y las dimensiones



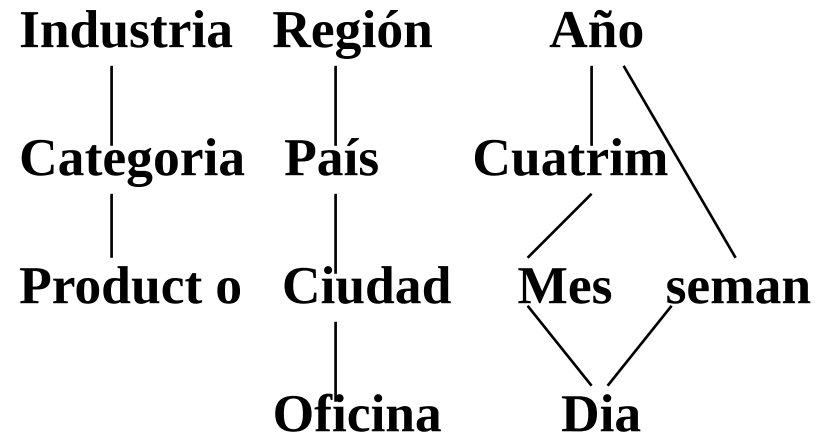
Datos Multidimensionales

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



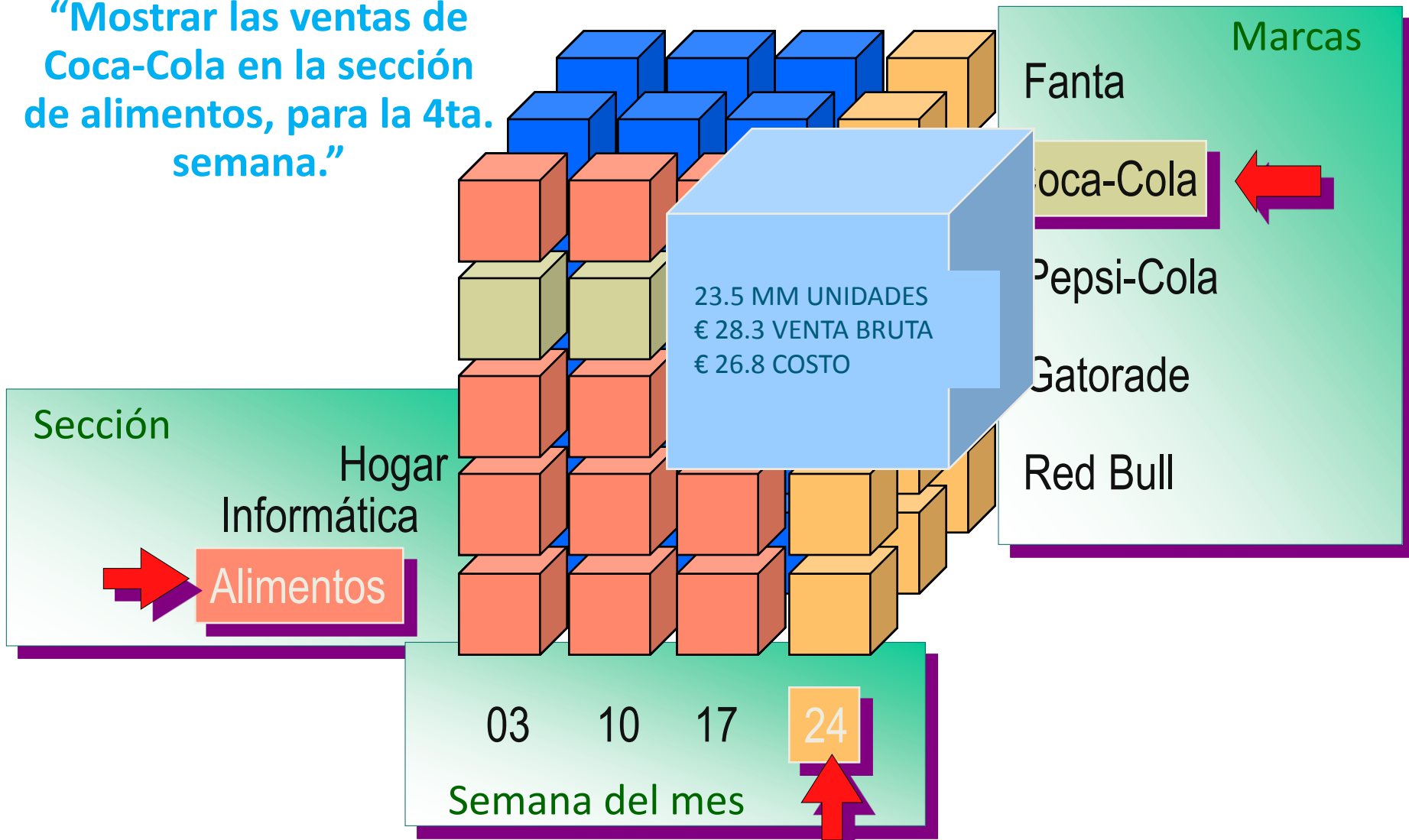
Dimensiones: Producto, Localidad, Tiempo

Caminos jerárquicos

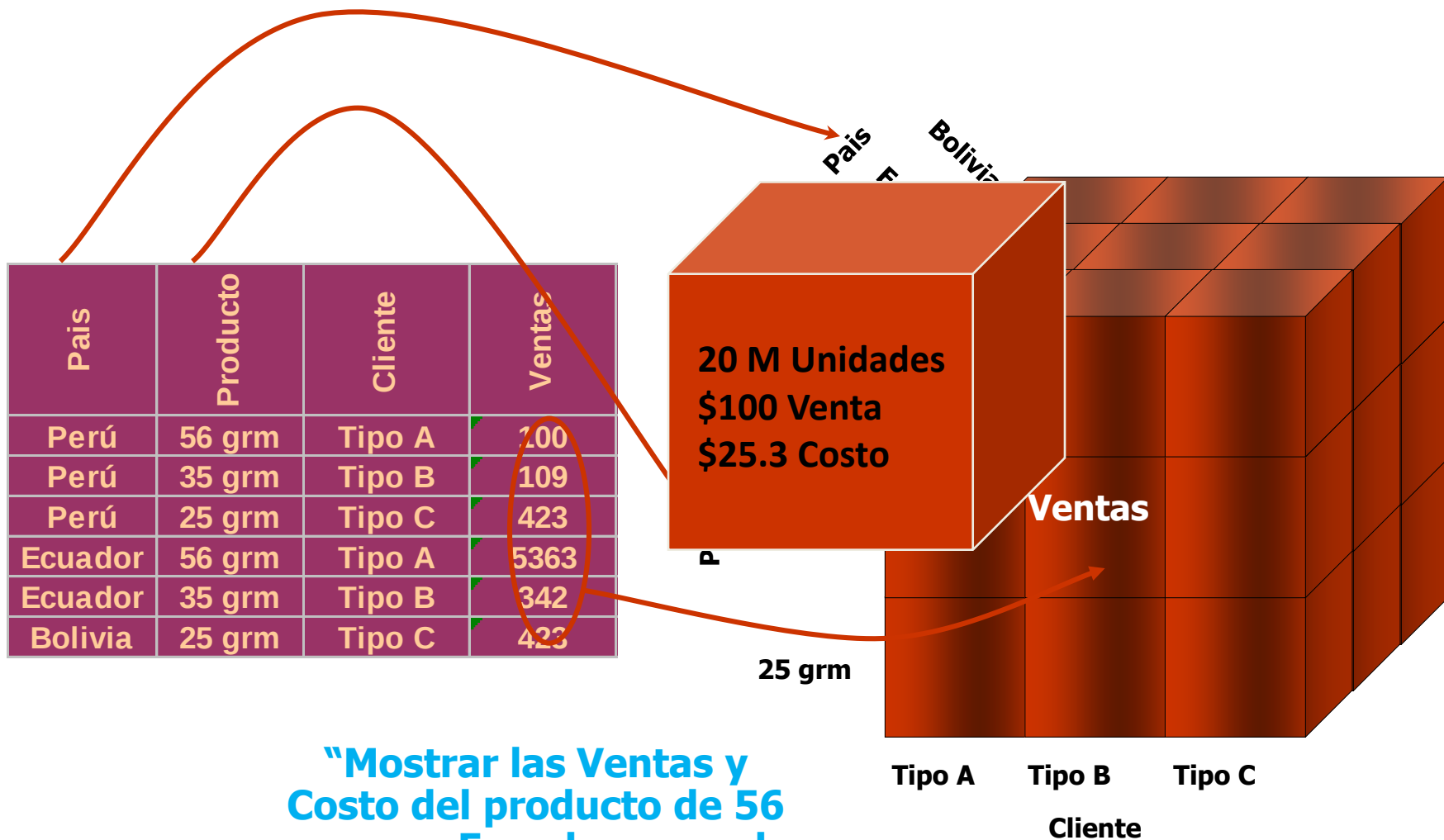


Cubo Multidimensional

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

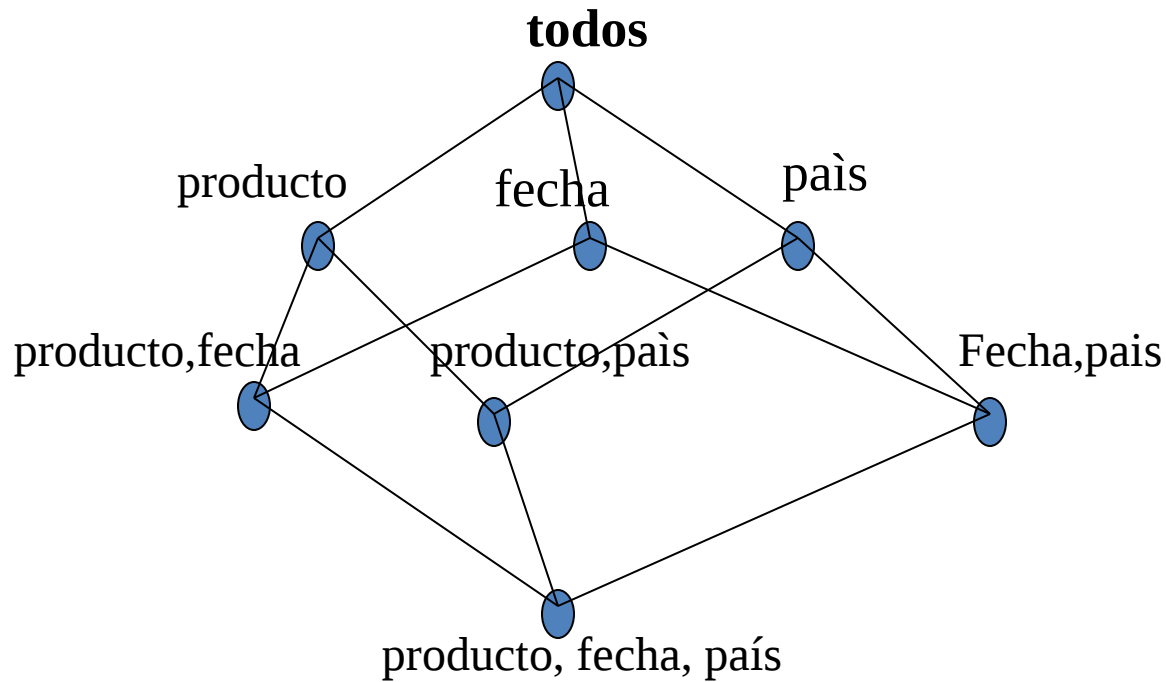


Cubo Multidimensional



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

Cuboides correspondientes al Cubo



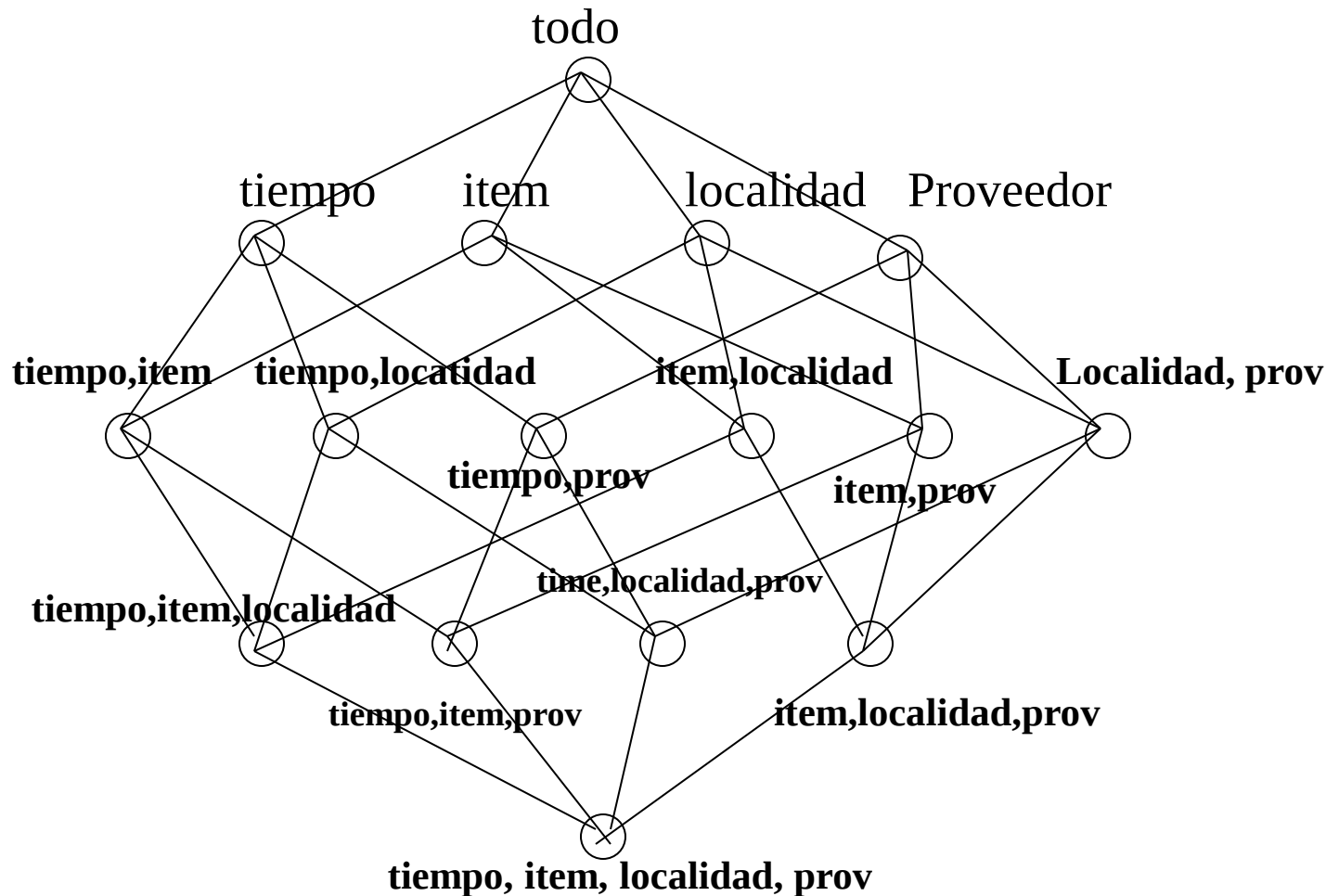
0-D cuboide

1-D cuboides

2-D cuboides

3-D cuboides

Cuboides correspondientes al Cubo



0-D cuboide

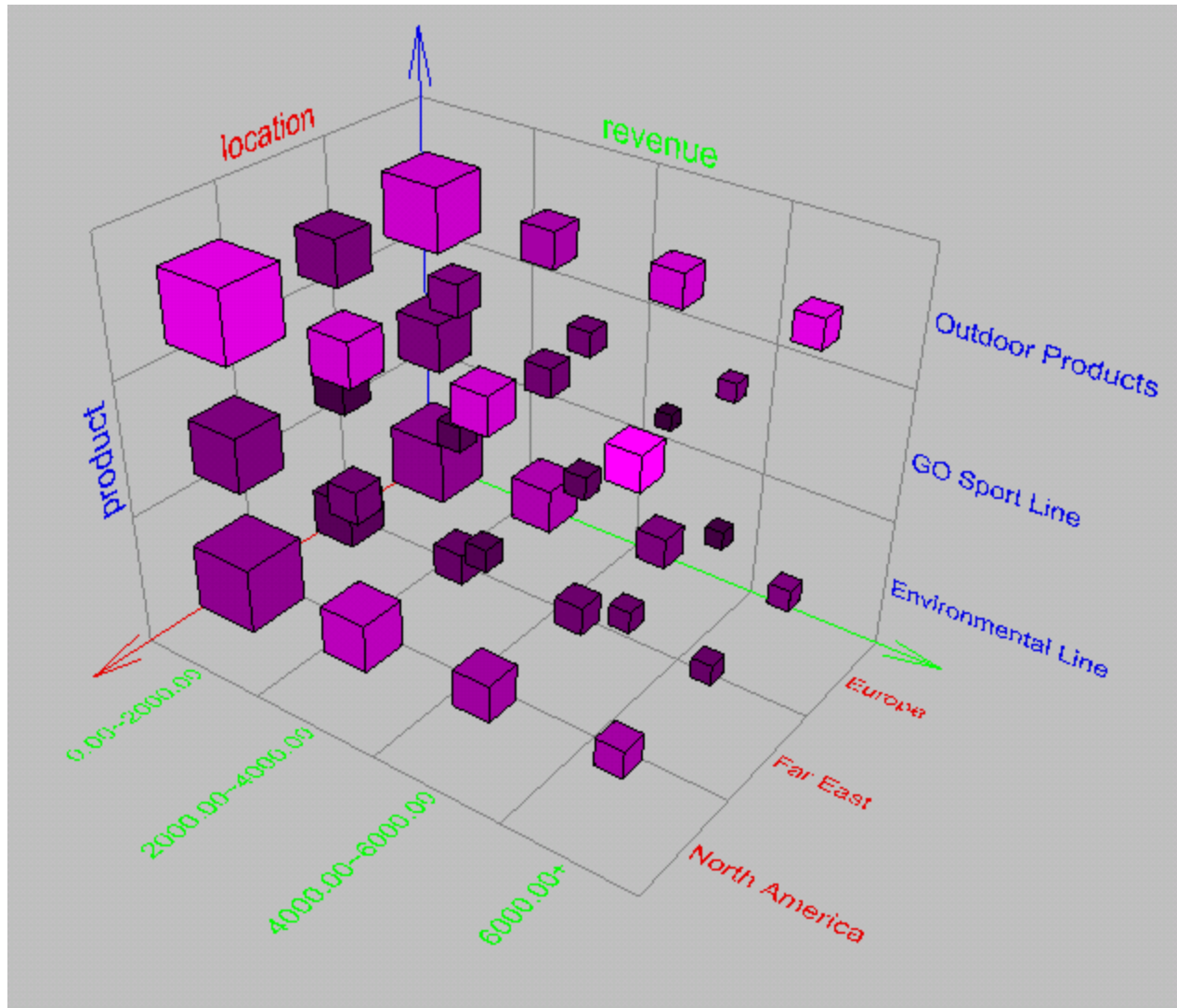
1-D cuboides

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

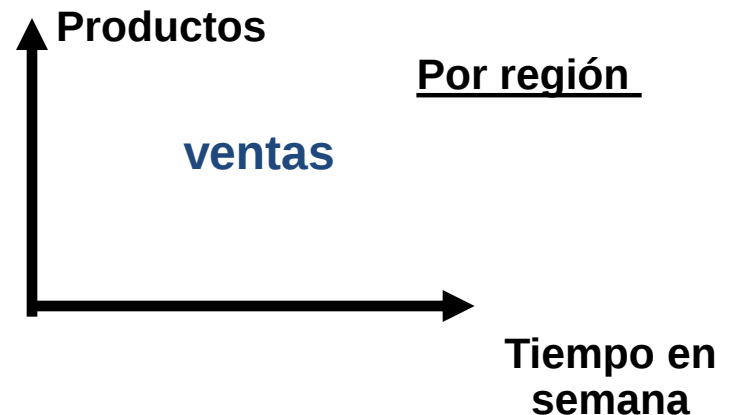


1 dimensión

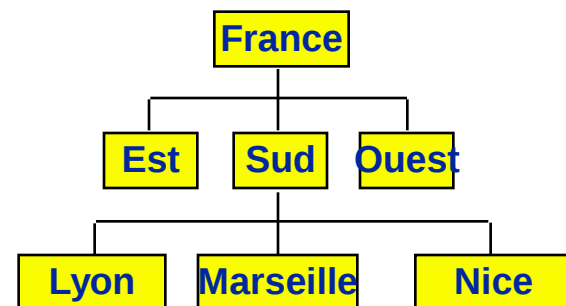


3 dimensiones

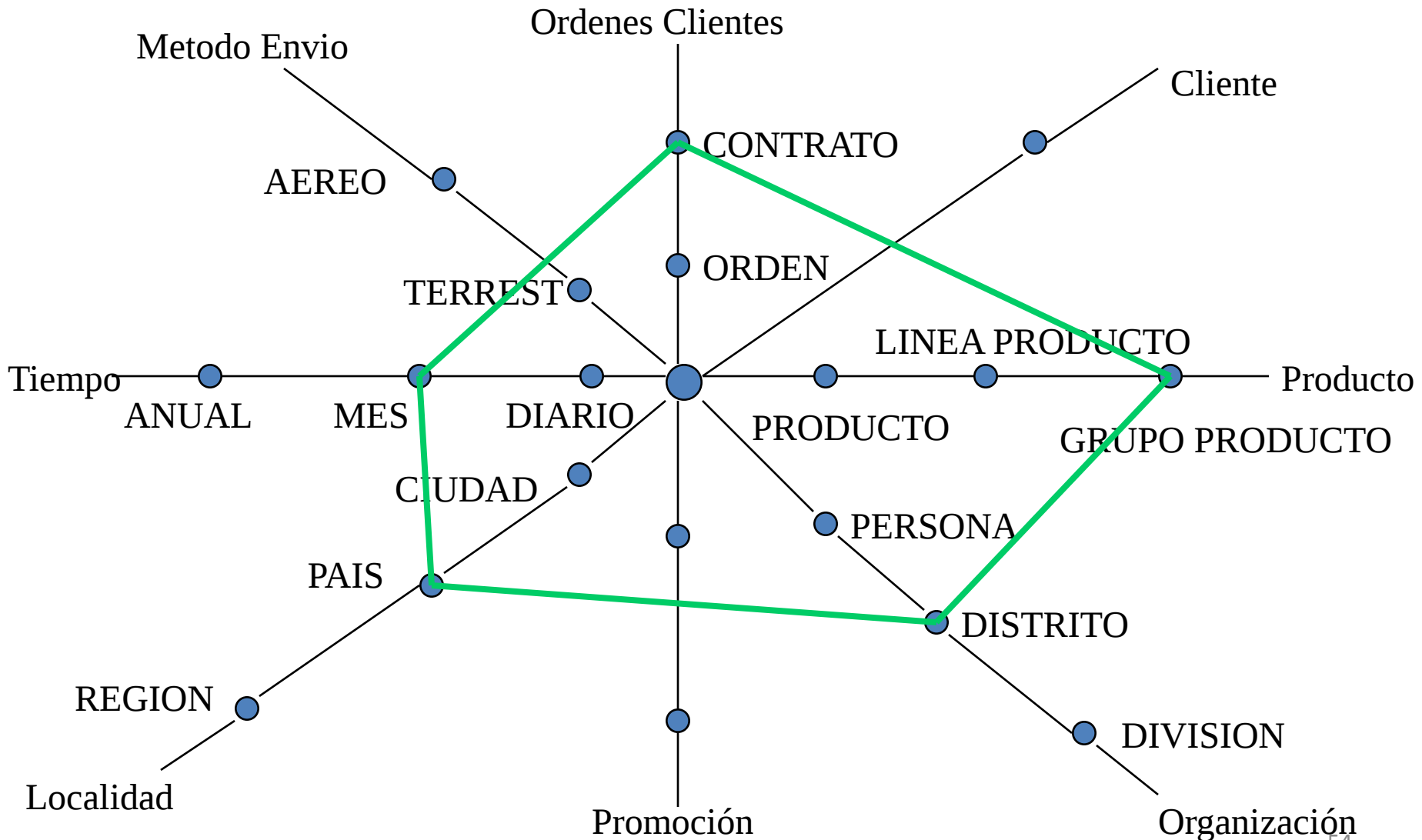
Copo de nieve



Zoom a una dimensión



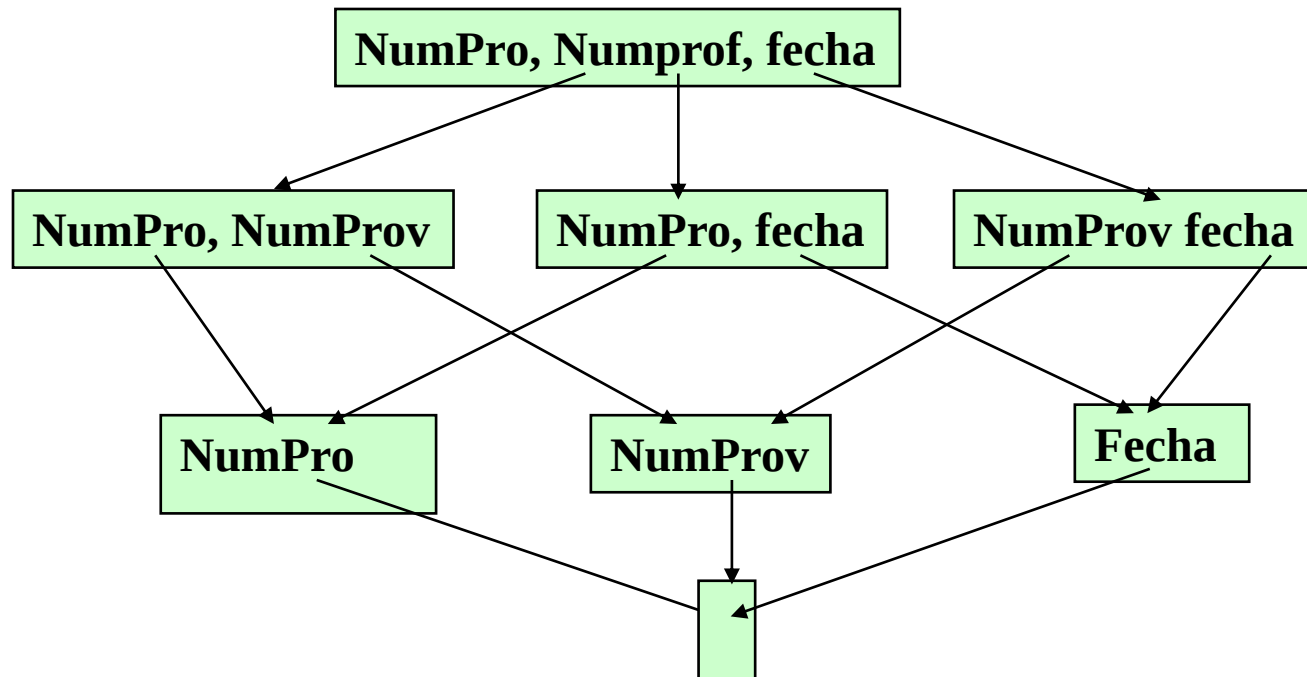
Modelo de consulta



Navegar por un cubo de datos

- Un cubo se puede rotar, agrupar, etc.

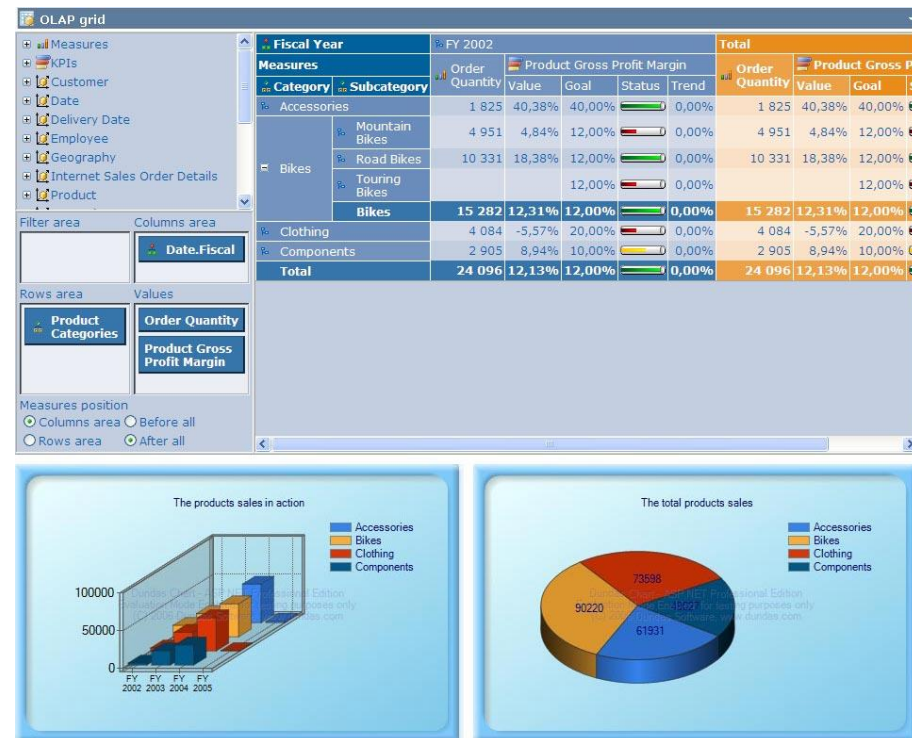
Se obtienen retículas de puntos de vista



Herramientas para explotación del Datawarehouse

Análisis multidimensional (OLAP online analytical processing)

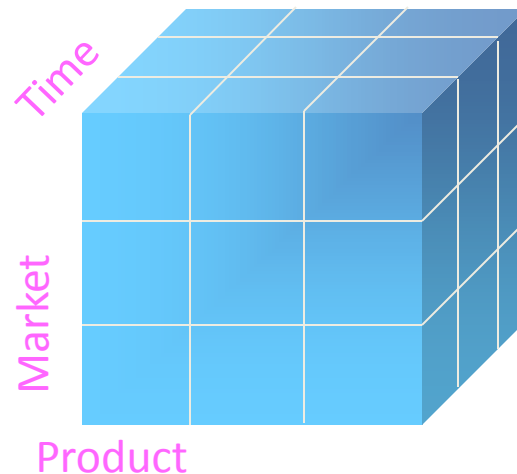
Facilitan el análisis de datos a través de dimensiones y jerarquías, utilizando consultas rápidas predefinidas



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

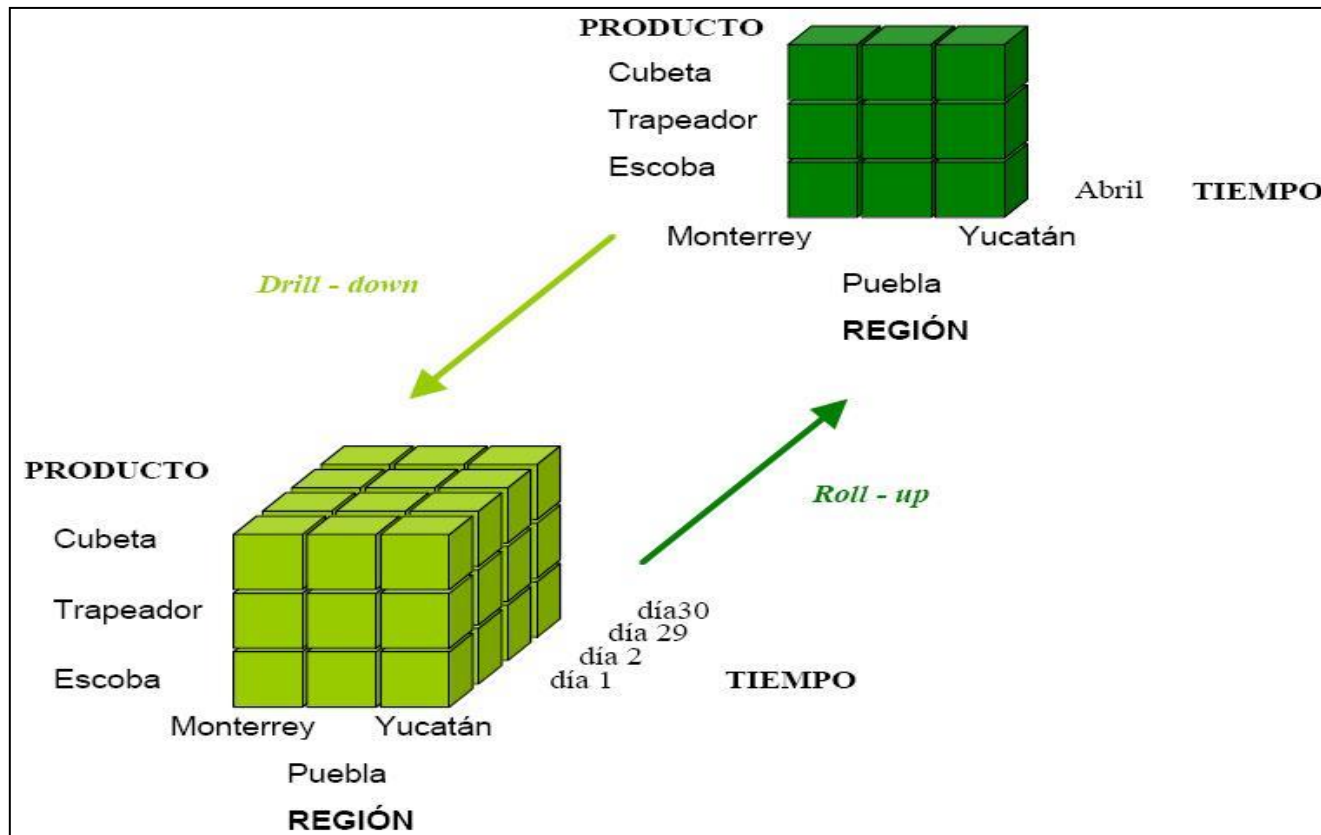
- **Motores de BD especializados**

- Manejan consultas especializadas (como las de SQL) con esquemas estrella o copo de nieve

Operaciones clásicas OLAP

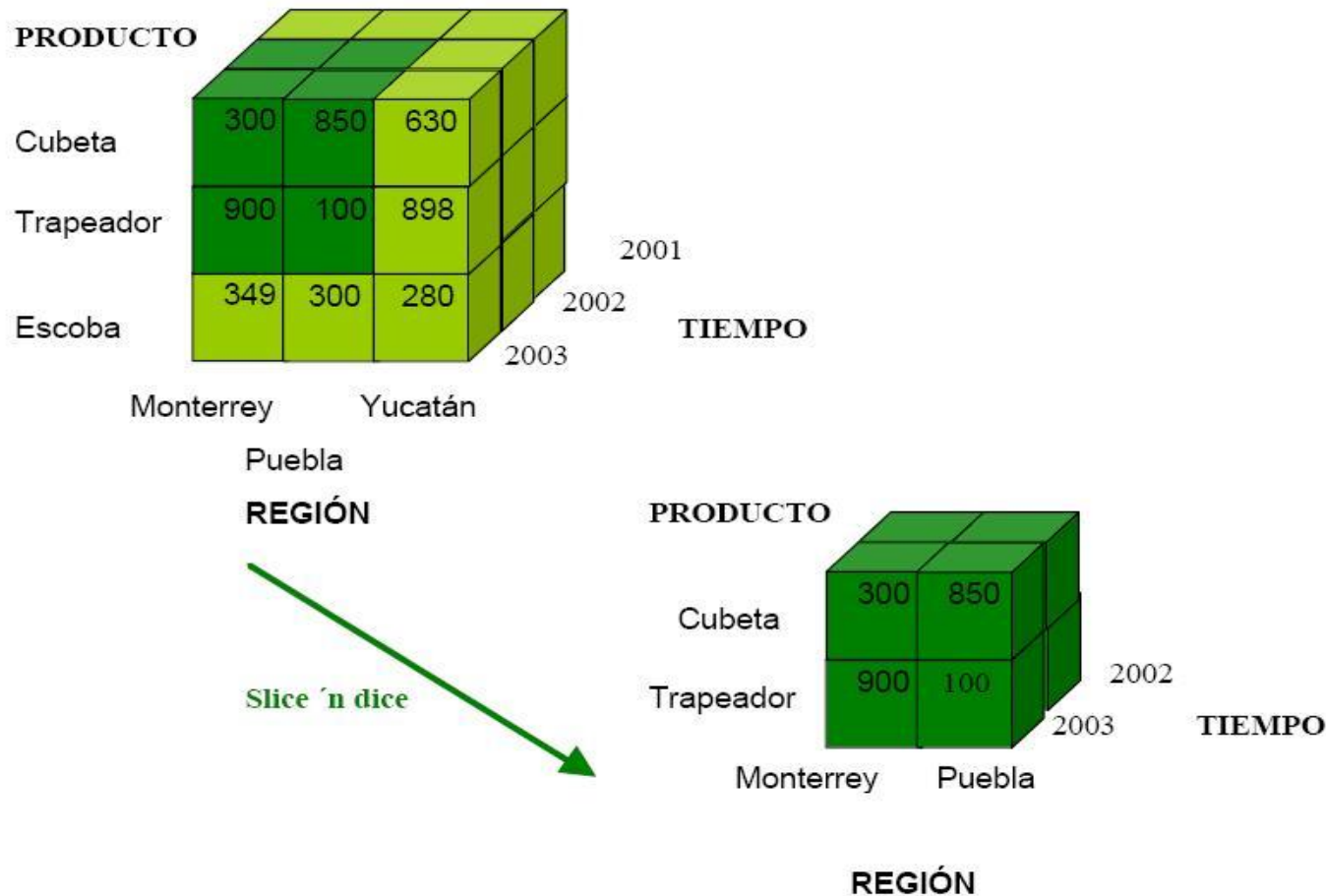
Roll up (drill-up): agrega medidas que van de un nivel N_i a un nivel mas general N_j de una dimensión.

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

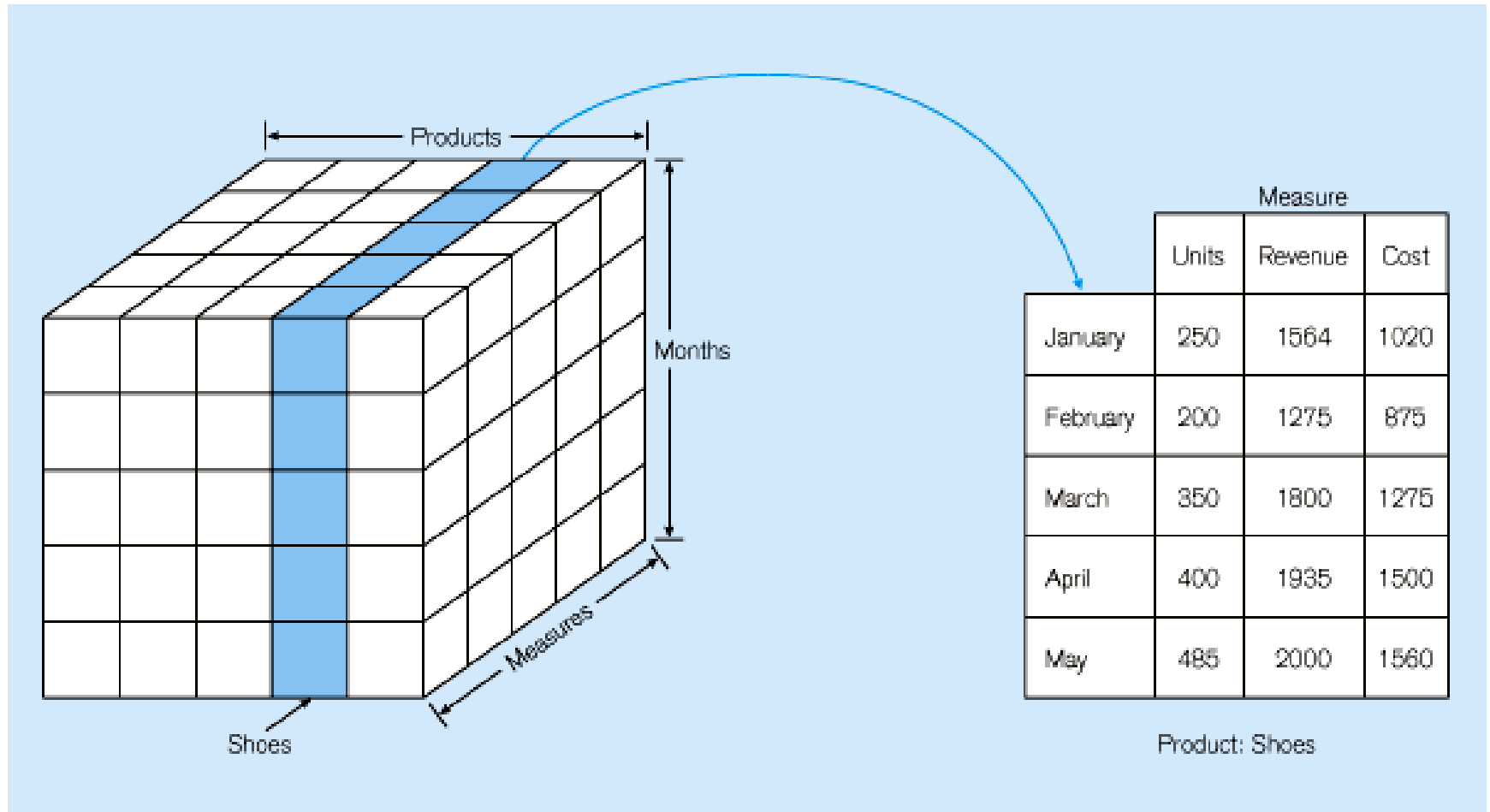
drill through

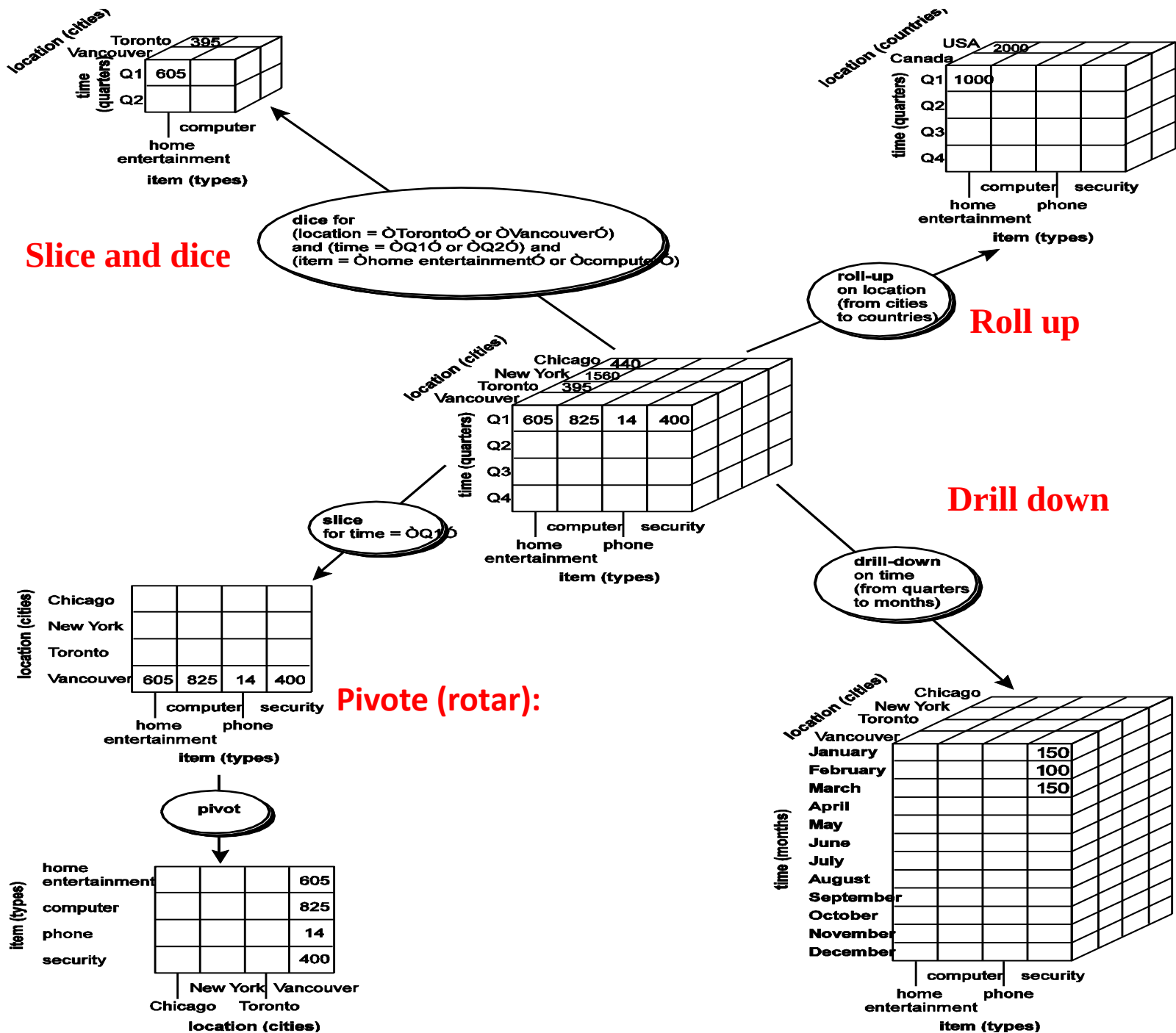
navegar a través del nivel inferior del cubo a tablas relacionales

Pivote (rotar)

Rotar el cubo

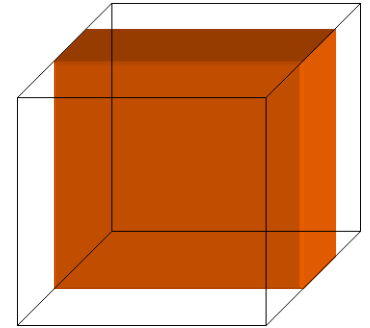
Cortando/rebanando un cubo de datos



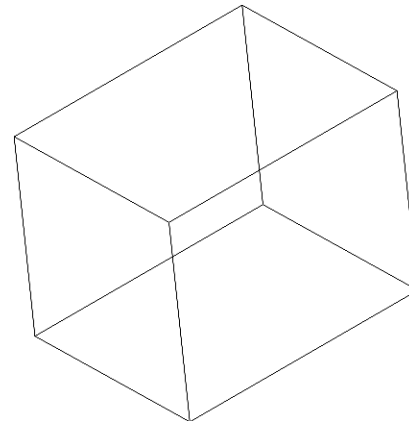


Resumen Operaciones clásicas OLAP

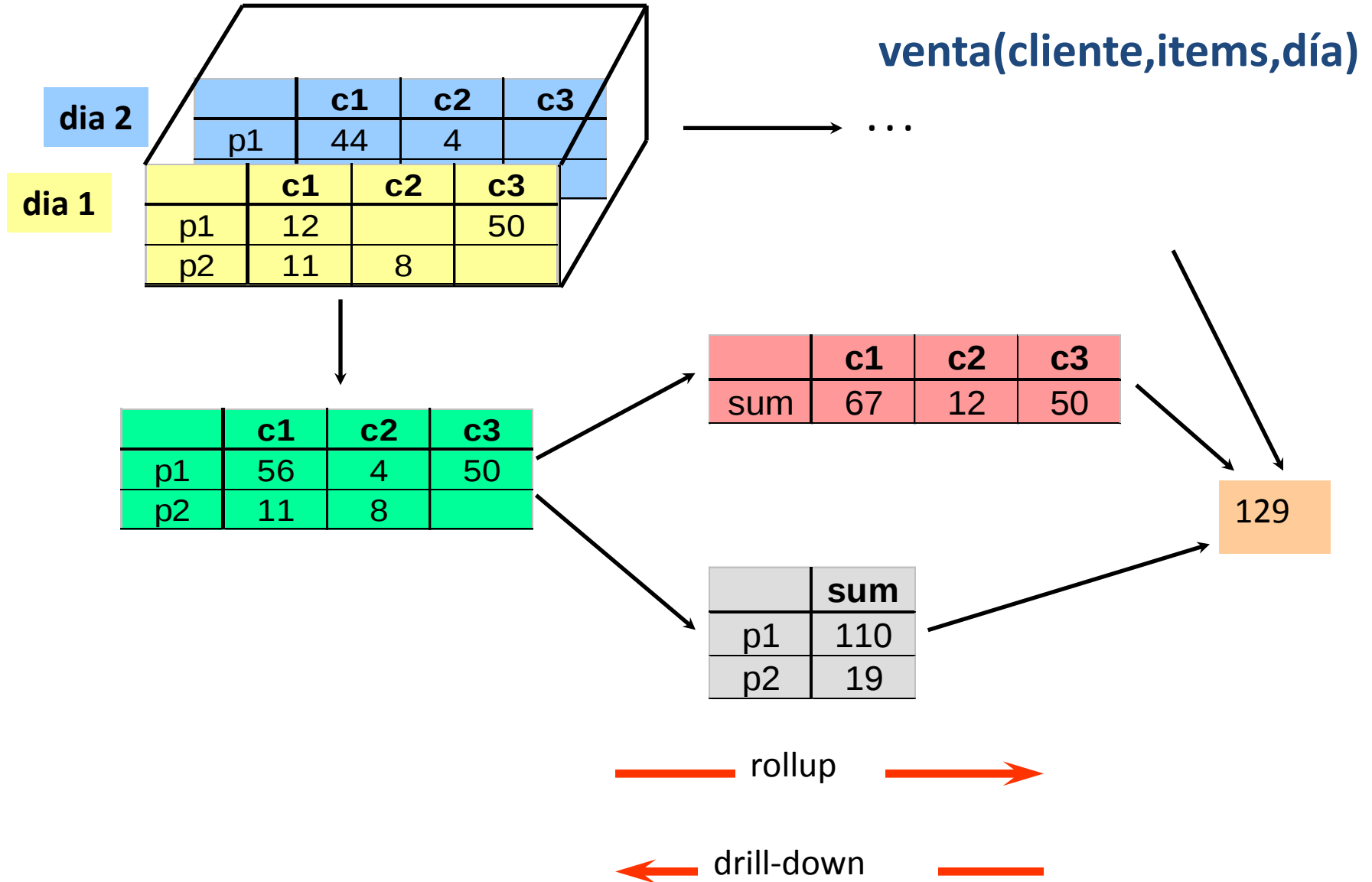
- *Rollup*: decrease nivel de detalle
- *Drill-down*: aumenta nivel de detalle
- *Slice-and-dice*: selección y proyección



- *Pivot*: re-orienta vista multidimensional

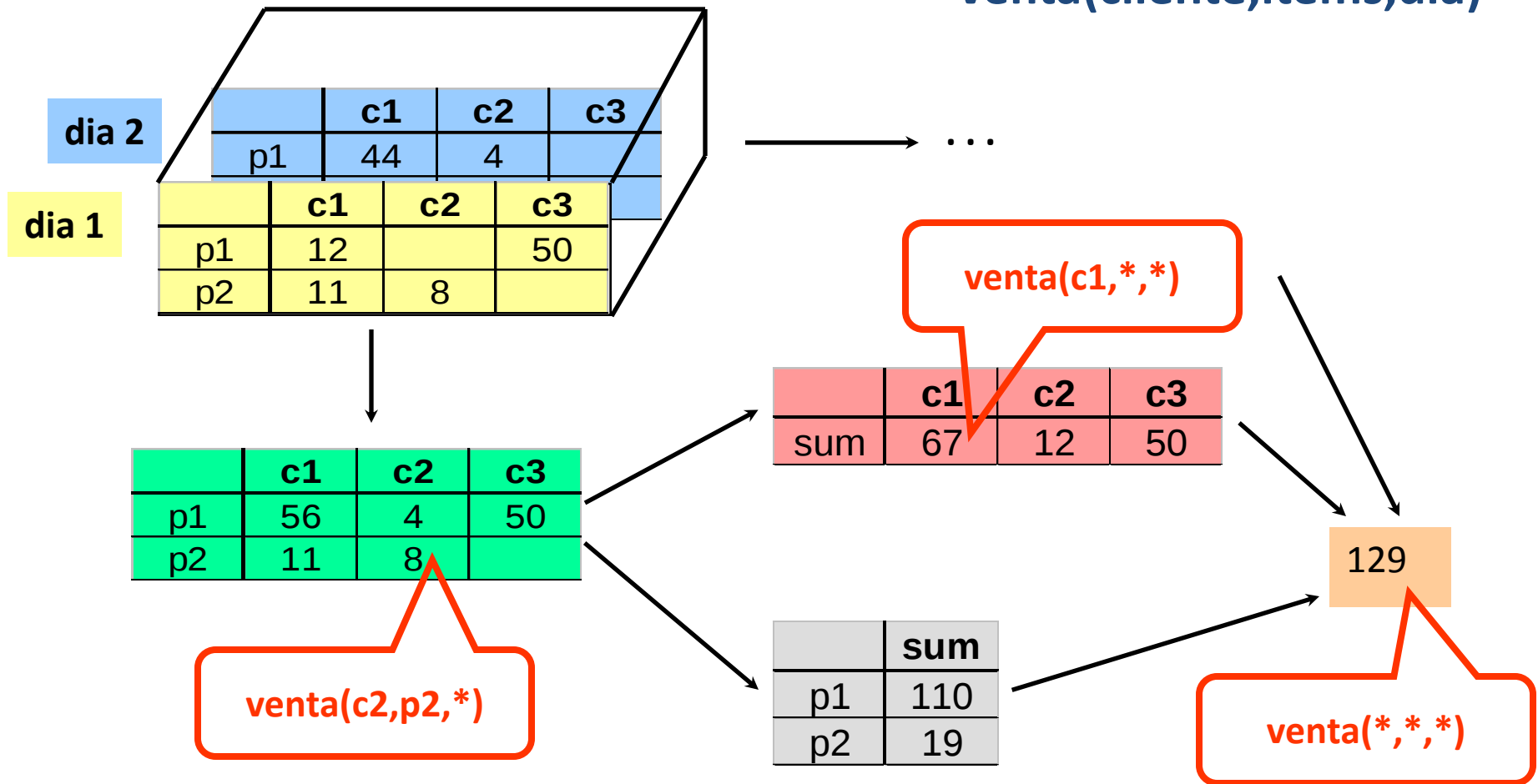


Agregación en Cubos



Agregación en Cubos

venta(cliente,item,día)



Cara de Cubos

venta(cliente,item,día)

dia 2

	*		c1	c2	c3	*
p1			56	4	50	110
p2			11	8		19
						129

dia 1

		c1	c2	c3	*
p1		12		50	62
p2		11	8		19
*		23	8	50	81

					*
					48
					48

venta(*,p2,*)

Agregación usando jerarquía

venta(cliente,items,día)

Día 2		c1	c2	c3
	p1	44	4	
Dia 1		c1	c2	c3
	p1	12		50
	p2	11	8	

↓

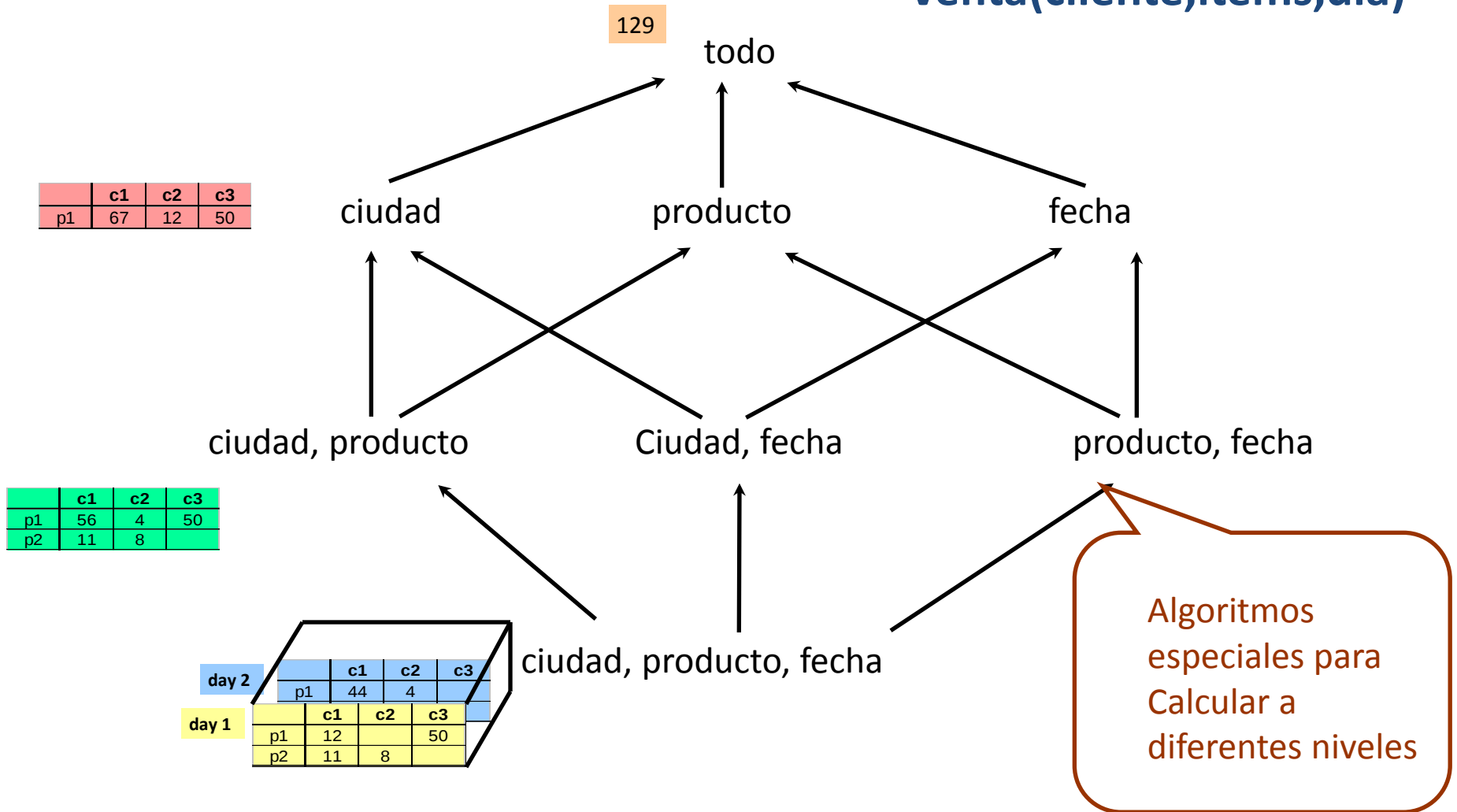
	region A	region B
p1	56	54
p2	11	8

cliente
|
region
|
país

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

Agregación en Cubos

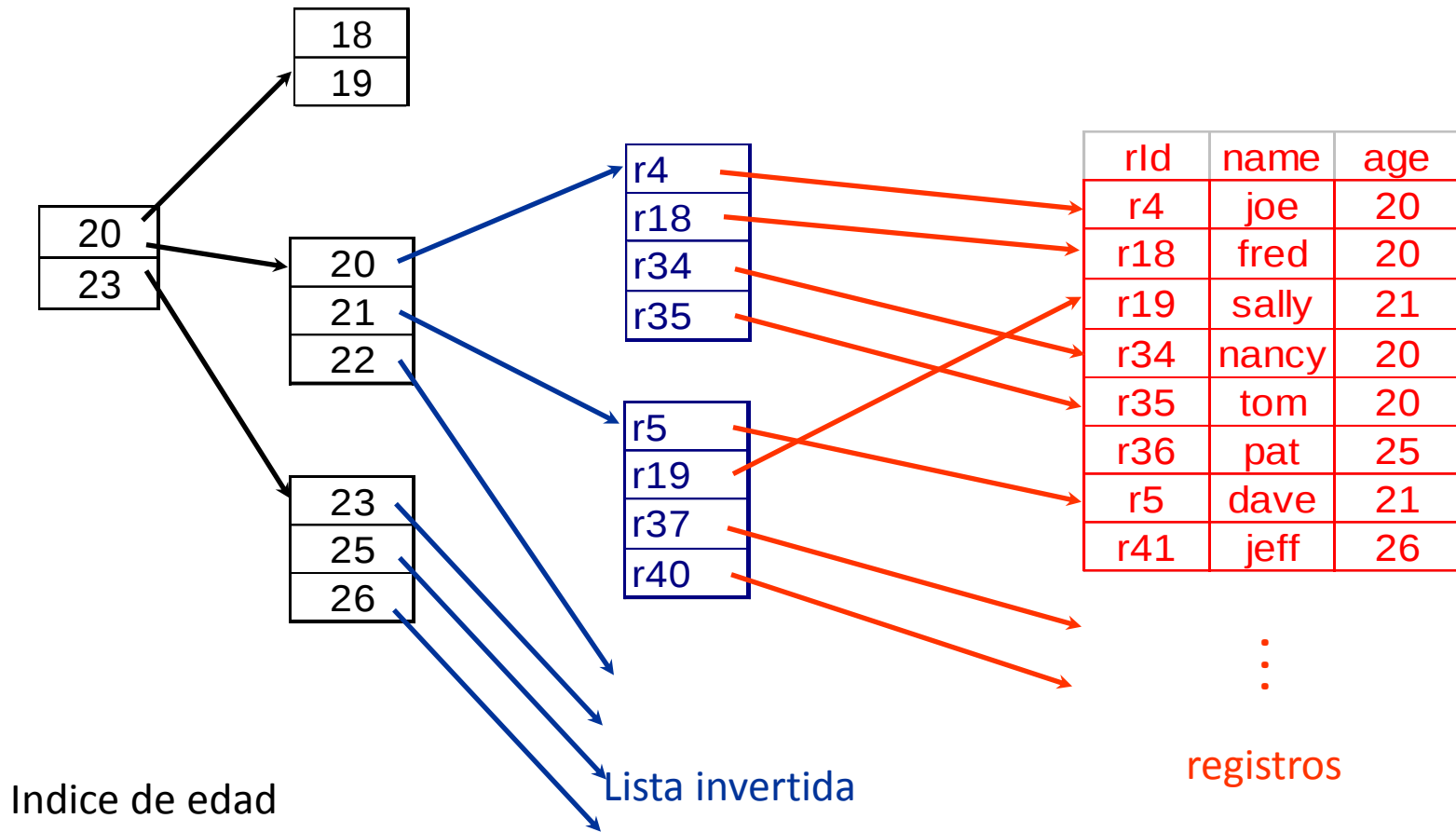
venta(cliente,item,día)



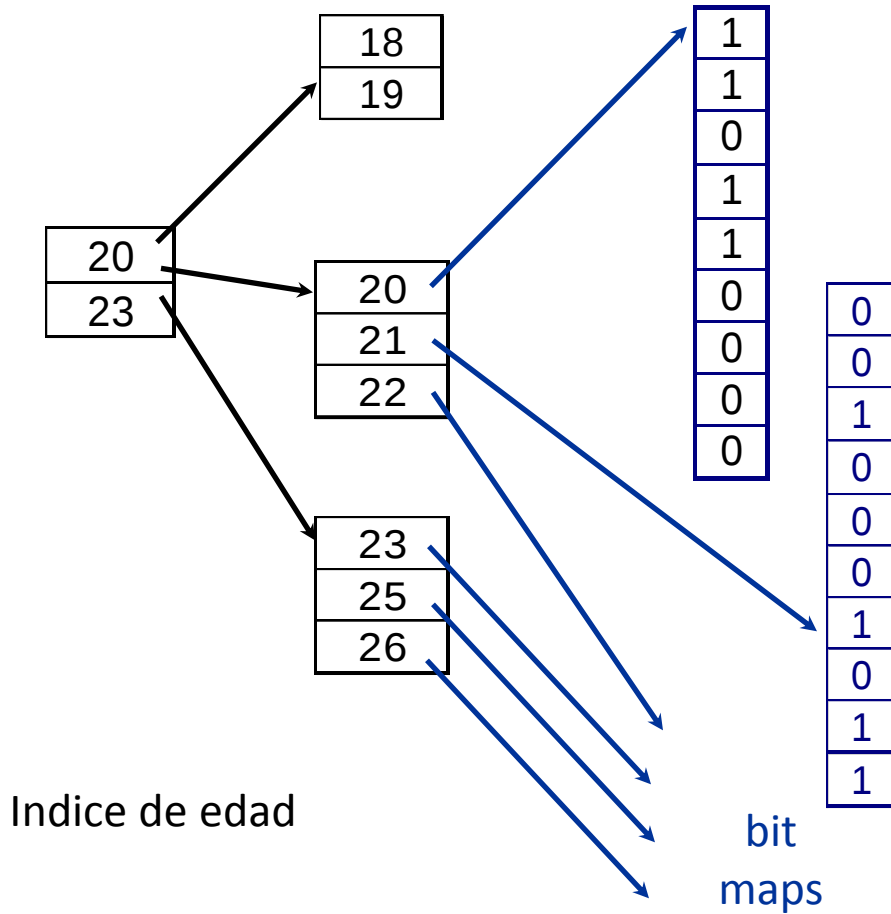
Estructuras Índices

- **Métodos de acceso tradicional**
 - Árboles B, tablas hash, ...
- **Popular en Warehouses**
 - Listas invertidas
 - Índices de mapa de bits (bit map)
 - Índices de unión (join indexes)
 - Índices de texto

Listas invertidas



Bit Maps

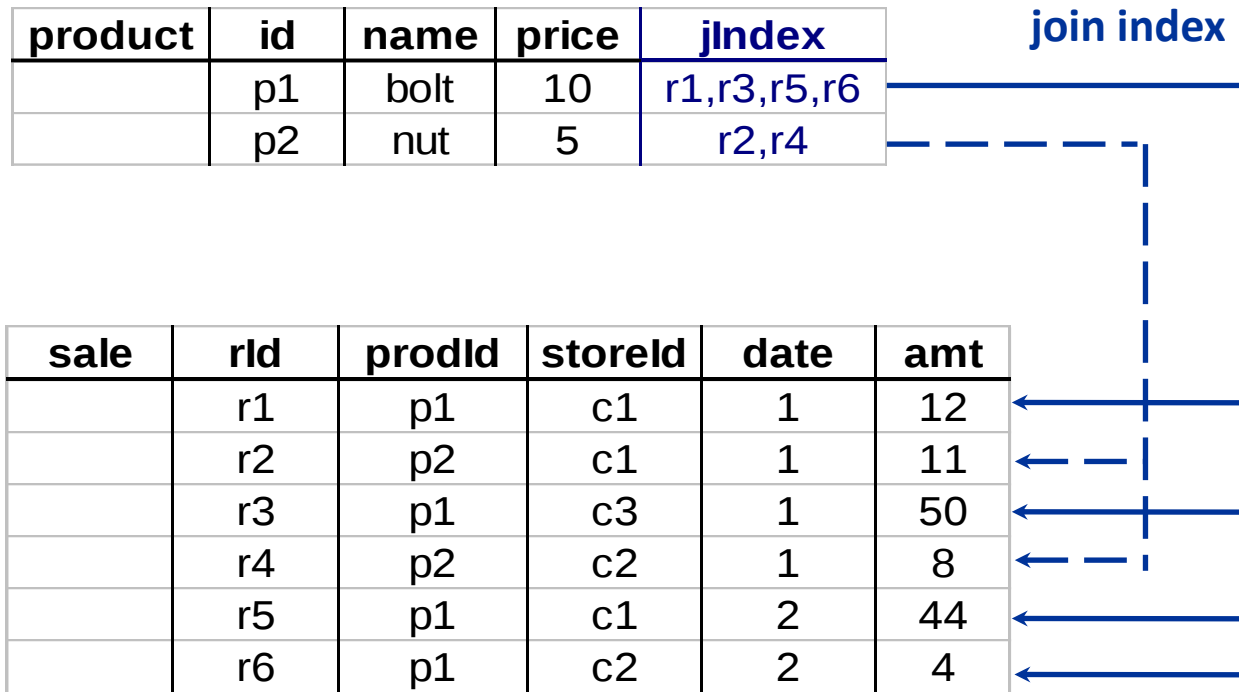


id	name	age
1	joe	20
2	fred	20
3	sally	21
4	nancy	20
5	tom	20
6	pat	25
7	dave	21
8	jeff	26

⋮

registros

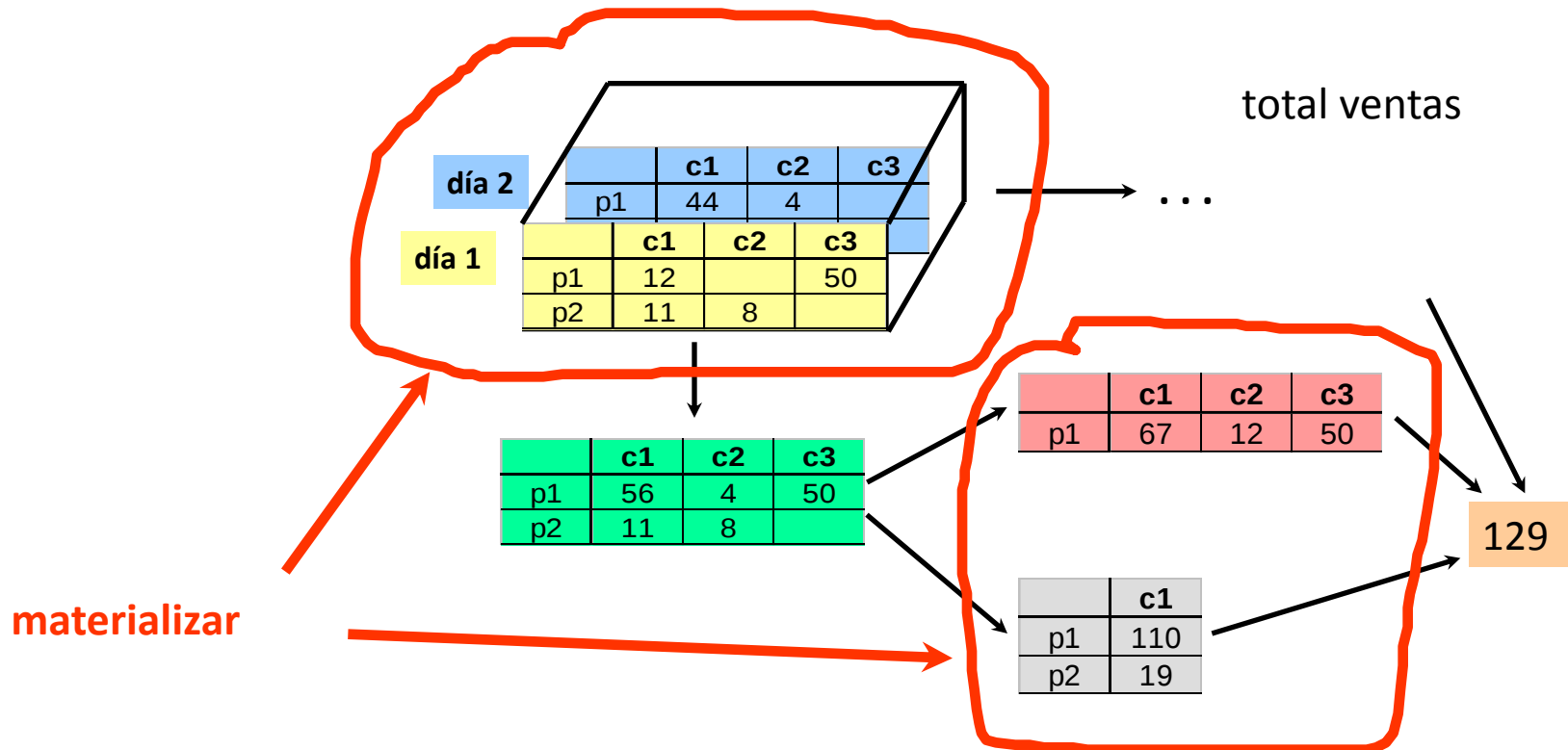
Join Indexes



- Relaciona los valores de las dimensiones de un esquema en estrella a filas de la tabla de hechos.
- Por ejemplo, la tabla de hecho ventas y la de dimensión producto
- Un índice join en producto guarda para cada producto una lista de los IDs de las tuplas que registran las ventas de ese producto

Consultas con OLAP

- Determinar cuales operaciones están **disponibles**
- Determinar que medidas es posible **materializar** en el cubo (relevantes a aplicar)



Extension de SQL

- **ROLLUP:**

- SELECT <column list>
- FROM <table...>
- GROUP BY
ROLLUP(column_list);

Hace $n+1$ agregaciones en una columna,

- **CUBO:**

- SELECT <column list>
- FROM <table...>
- GROUP BY
CUBE(column_list);

Crea n combinaciones, n será el numero de columnas del grupo

Ejemplo CUBO

Animaux

Animal	Lieu	Quantite
Chat	Paris	18
Chat	Naples	9
Chat	-	27
Chien	Paris	12
Chien	Naples	5
Chien	Rome	14
Chien	-	31
Tortue	Naples	1
Tortue	Rome	4
Tortue	-	5
-	-	63
-	Paris	30
-	Naples	15
-	Rome	18

Animal	Lieu	Quantite
Chien	Paris	12
Chat	Paris	18
Tortue	Rome	4
Chien	Rome	14
Chat	Naples	9
Chien	Naples	5
Tortue	Naples	1

**SELECT Animal, Lieu,
SUM(Quantite) as Quantite
FROM Animaux
GROUP BY ROLLUP Animal**

Ejemplo CUBO

SELECT item, city, year, SUM (amount)

FROM SALES

CUBE BY item, city, year

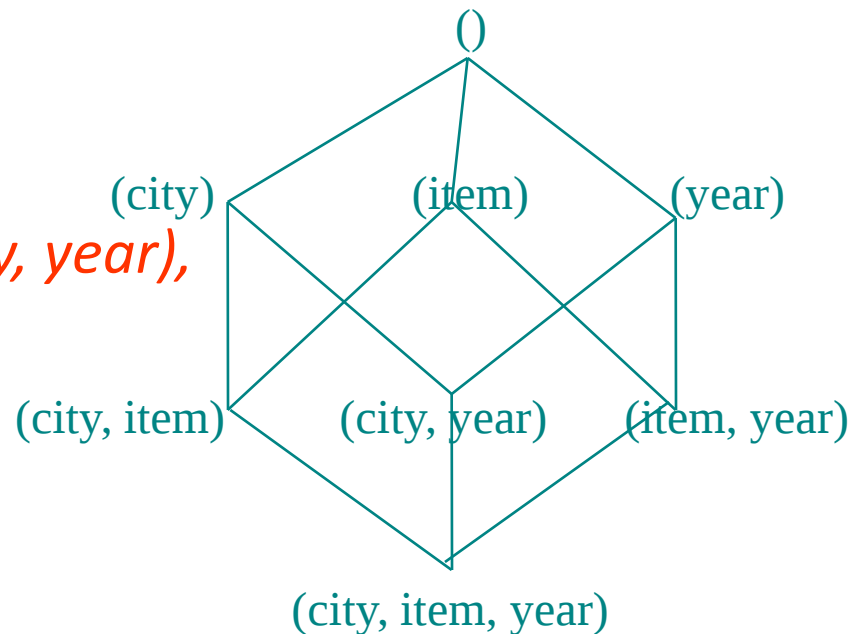
- Se debe calcular

(item, city, year),

(item, city), (item, year), (city, year),

(item), (city), (year)

()



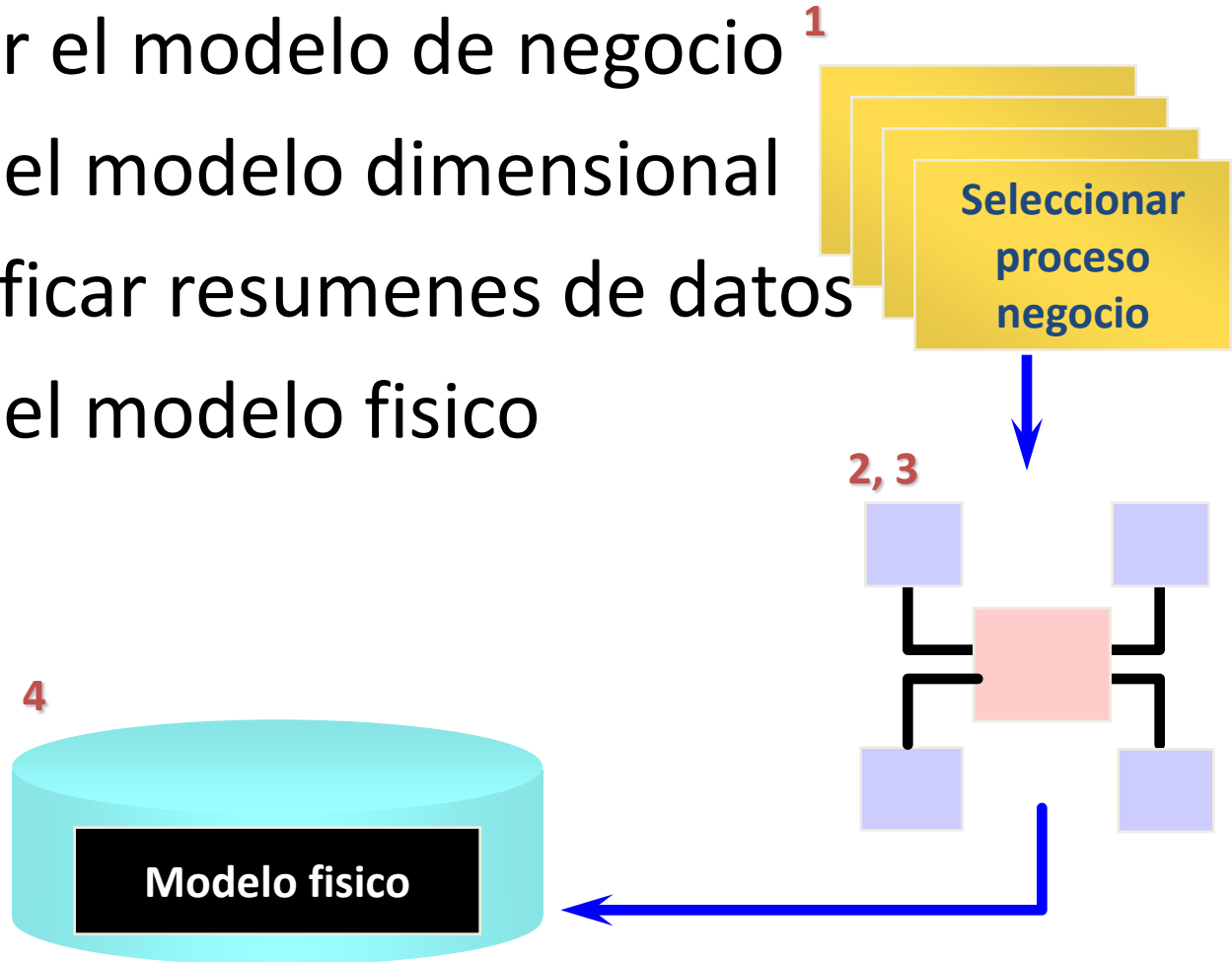
Extension de SQL

- Agrupar todos los subconjuntos de {item, región, mes}, encontrar el precio máximo en 1997 de cada grupo, y el total de ventas entre todas las tuplas de precios máximos

```
select item, región, mes, max(precio), sum(R.ventas)
from compras
where año = 1997
cube by item, region, mes: R
such that R.precio = max(precio)
```

Modelado en Data Warehouse

1. Definir el modelo de negocio¹
2. Crear el modelo dimensional
3. Identificar resúmenes de datos
4. Crear el modelo físico



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

- 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

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

Identificar resúmenes de datos

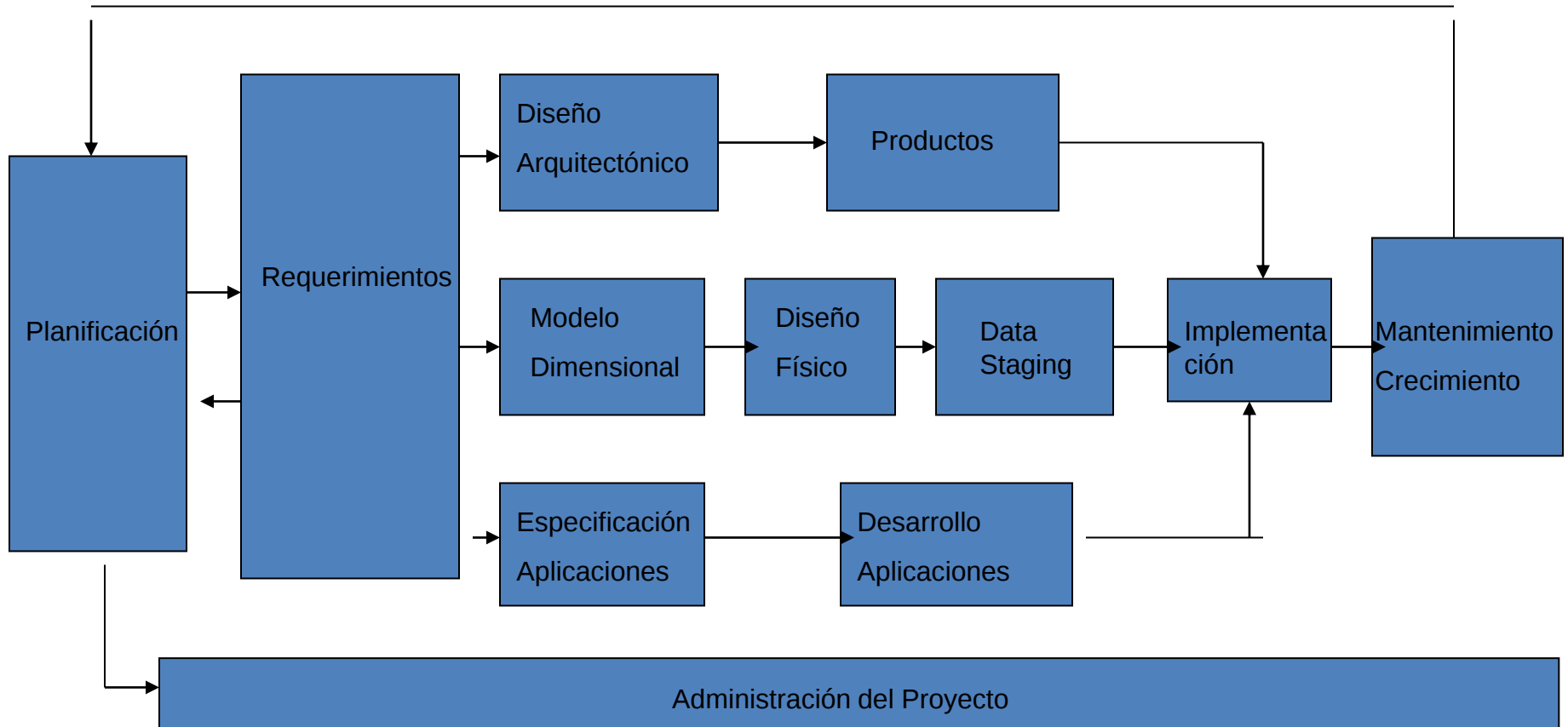
- Promedio
- Máximo
- Total
- Porcentaje

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

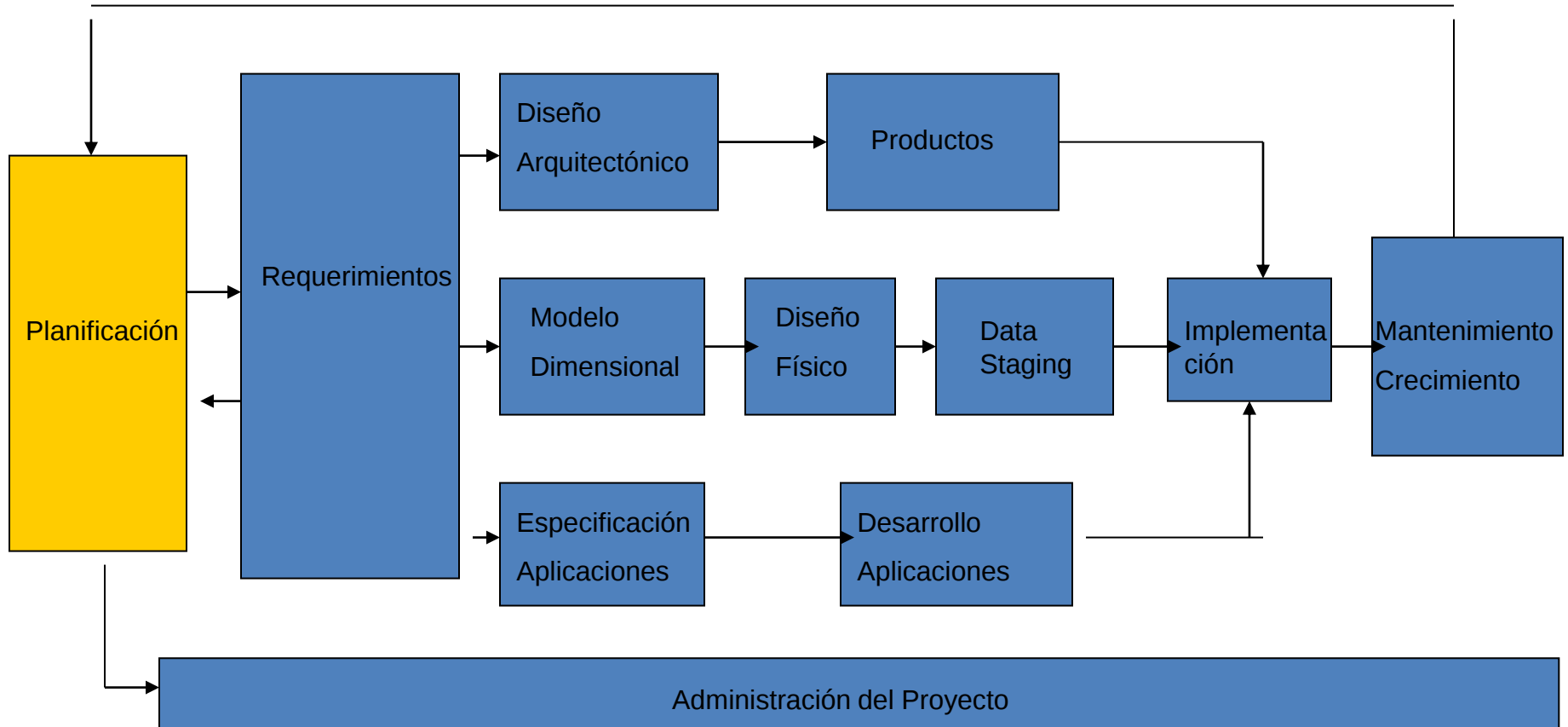
Implementación de un Data Warehouse

- Un proyecto de Data Warehouse debe basarse en satisfacer necesidades organizacionales
- Los datos se deben presentar de acuerdo con las variables organizacionales
- Un Data Warehouse está en permanente evolución
- Cada proyecto de Data Warehouse debe tener un principio y un fin

Ciclo de vida



Planificación



Planificación

- Predisposición de la organización
- Alcance
- Justificación
- Aspectos humanos
- Plan del proyecto

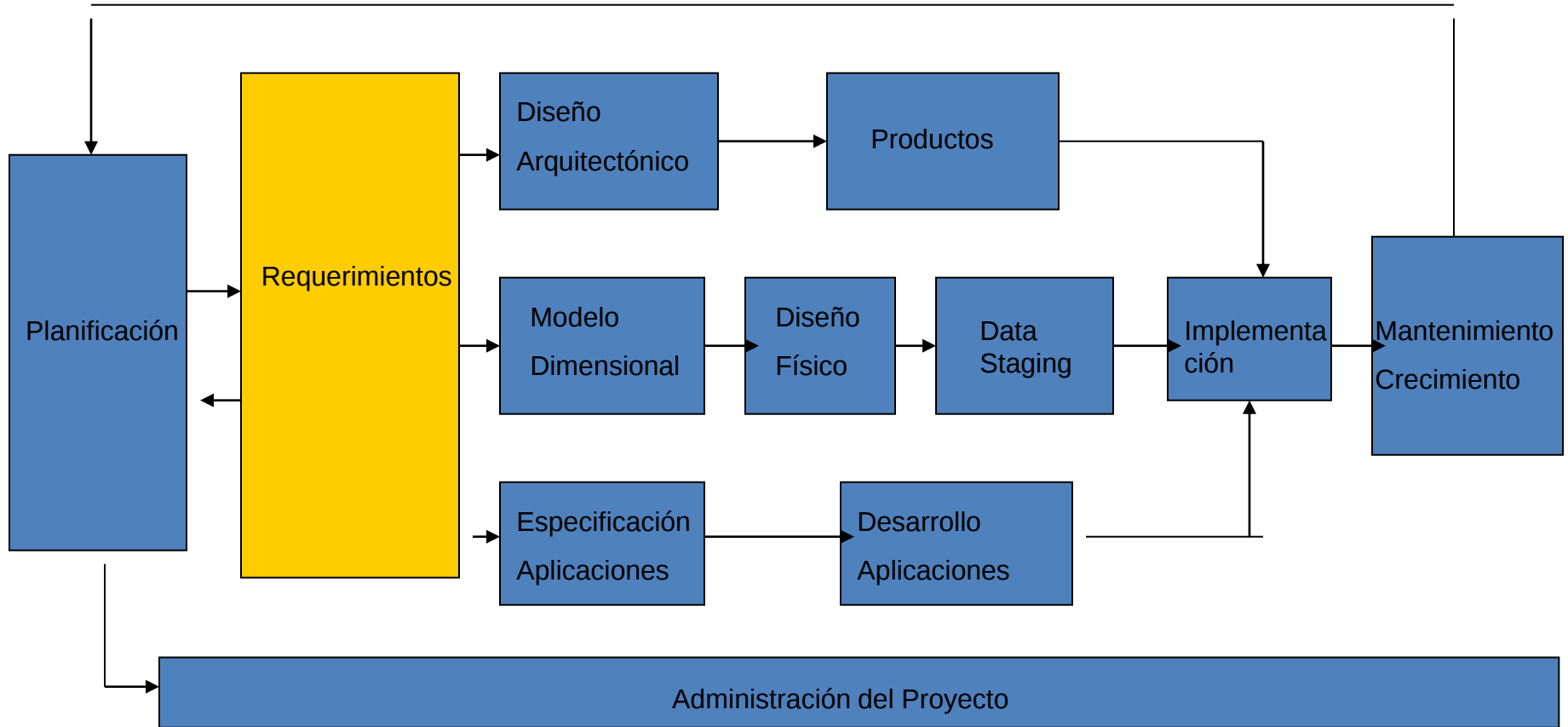
Predisposición de la organización

- Apoyo de la Gerencia (Sponsor)
- Motivación en la organización
- Participación de la gente vinculada al problema a analizar y de Sistemas
- Cultura actual de análisis de información
- Factibilidad

Puntos clave

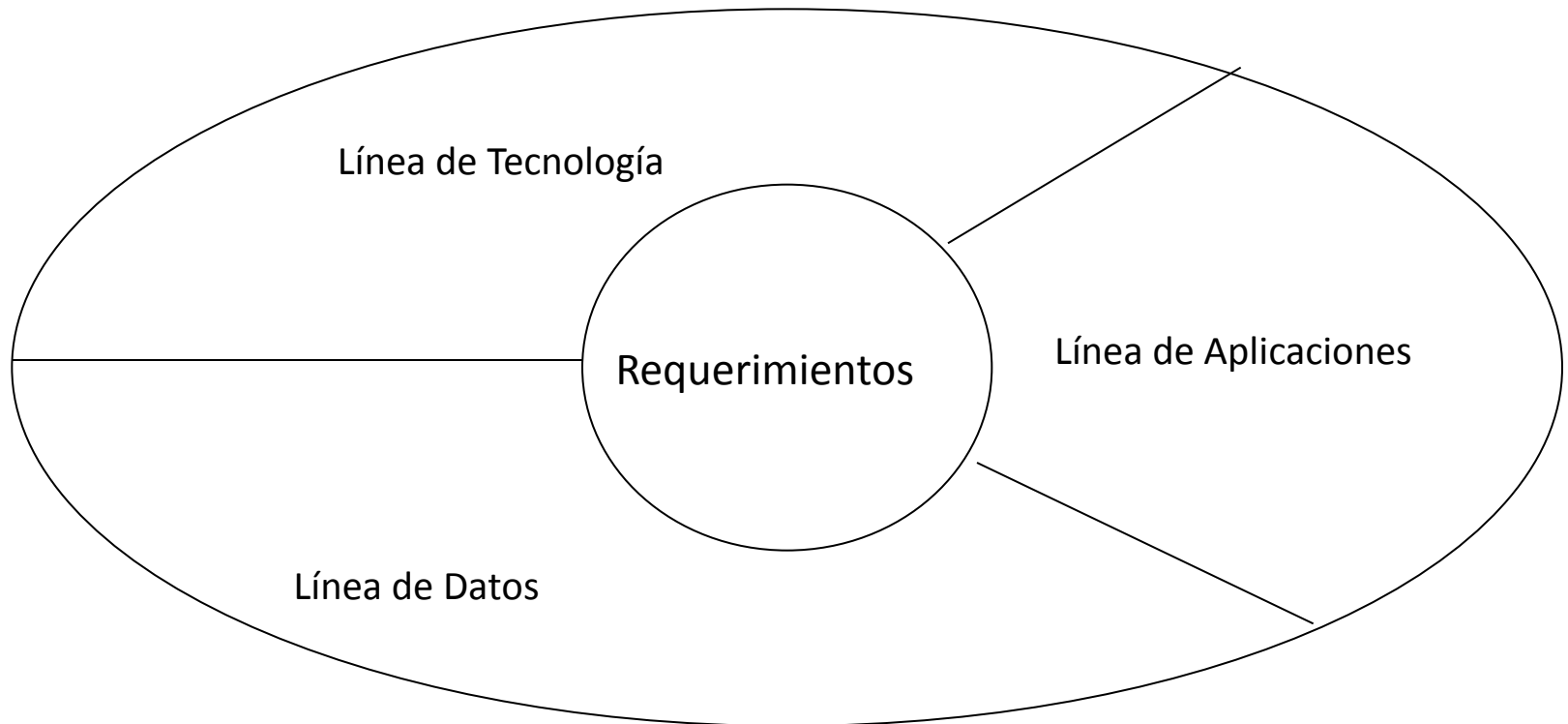
- Buscar un patrocinante bien ubicado
- Hacer un balance entre valor para la organización y manejabilidad
- Desarrollar cuidadosamente el plan del proyecto
- Ser un director de proyecto con capacidad de motivar, administrar y comunicar a todos los niveles

Requerimientos

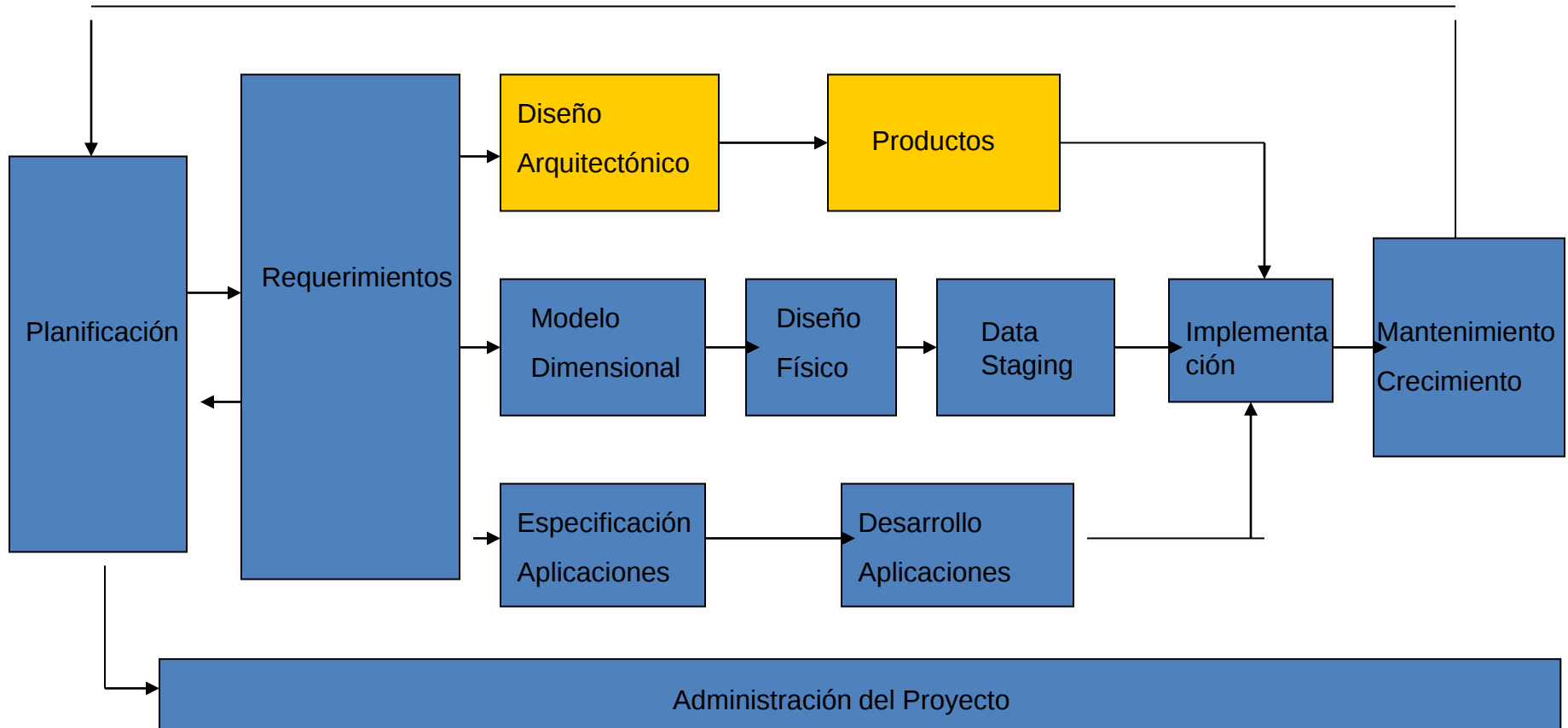


Desarrollo del proyecto

El desarrollo del proyecto se realiza en tres líneas



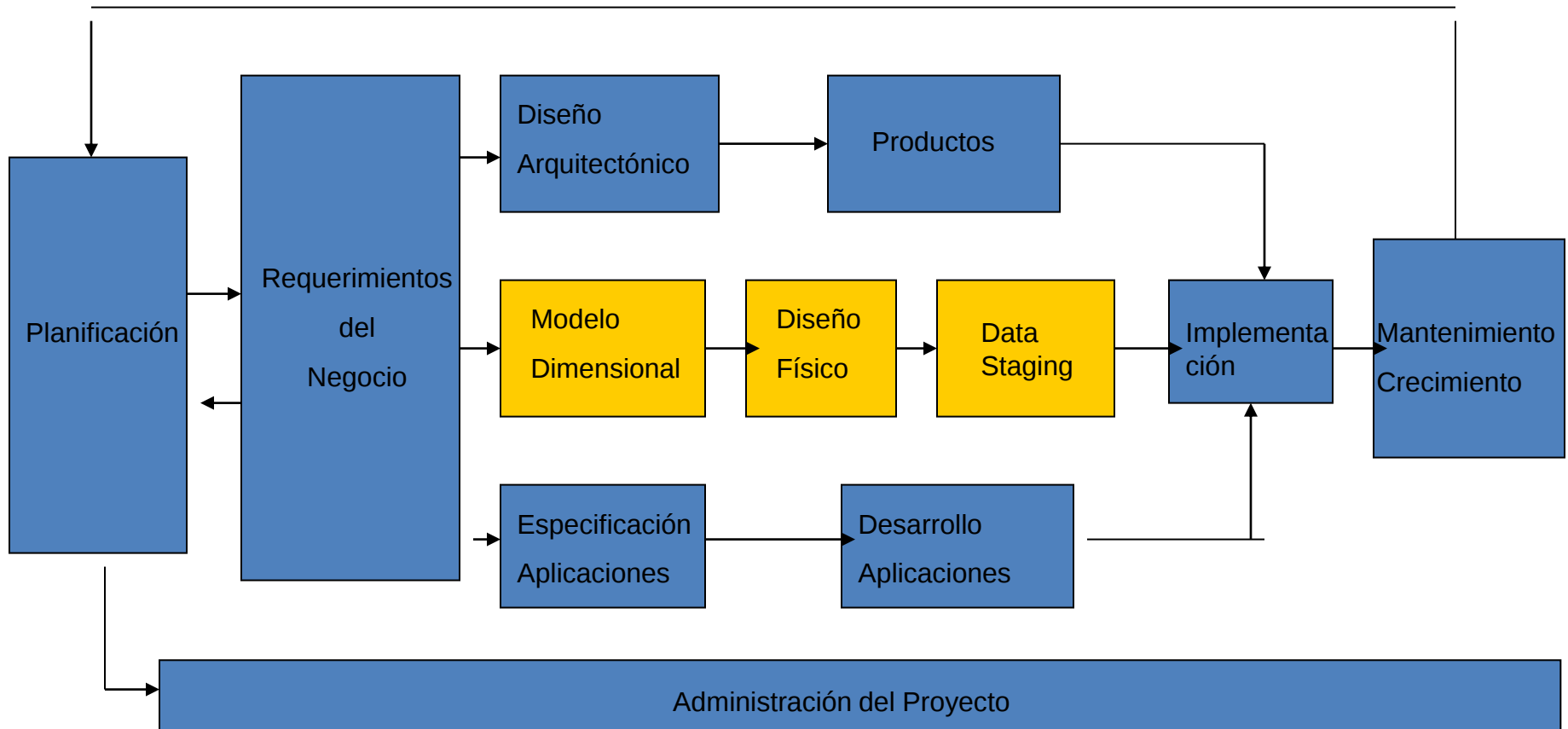
Línea de Tecnología



Selección de productos

- Política de compras
- Matriz de evaluación
- Requerimientos técnicos
- Factores de ponderación

Línea de Datos



Línea de Datos

- Modelado dimensional
 - Procesos
 - Granularidad
 - Tablas de hechos
 - Tablas de dimensión
 - Claves

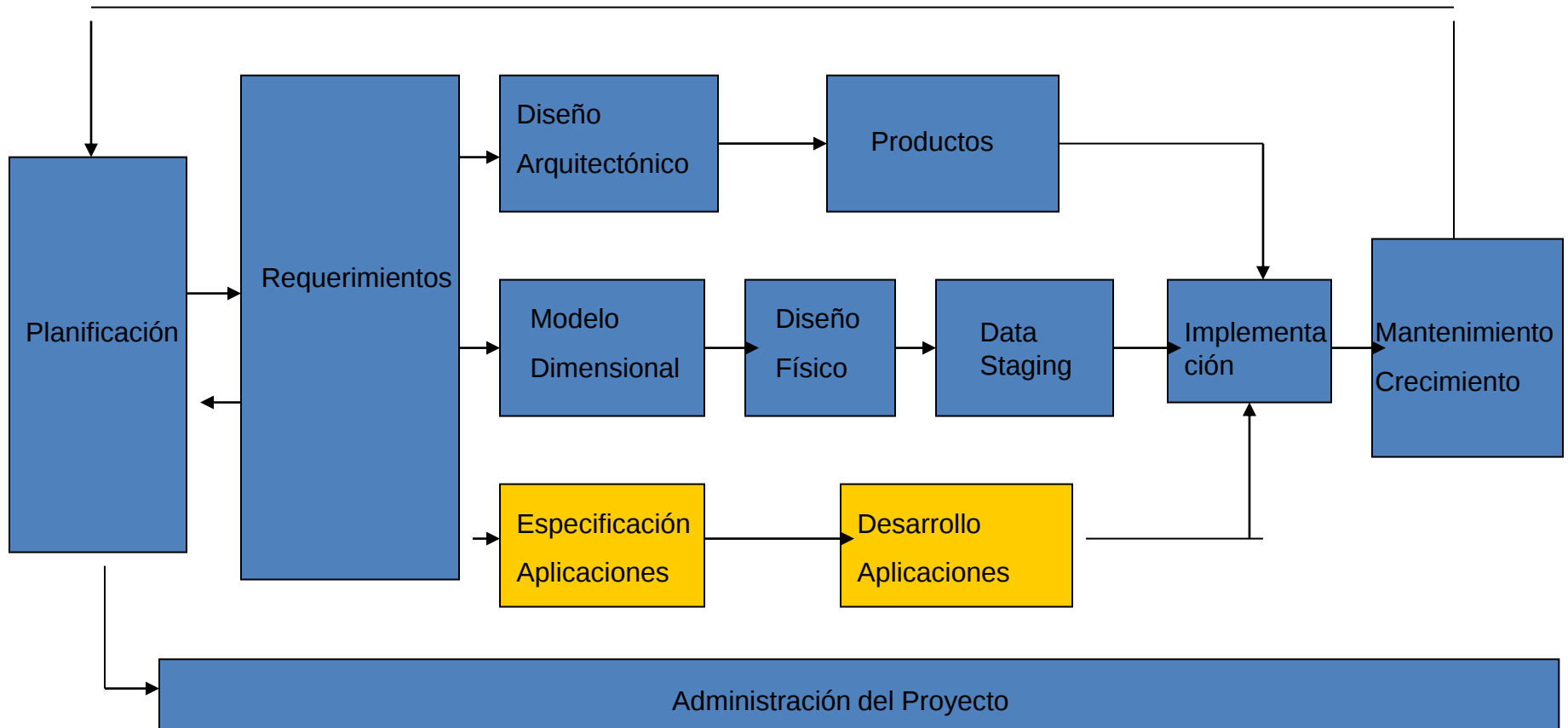
Línea de Datos

- Diseño físico de la base de datos
 - Nombres de columnas
 - Tipos de datos
 - Estrategia de consolidación/actualización
 - Qué consolidaciones se incluyen
 - Estrategia de indexación

Línea de Datos

- Diseño y desarrollo del ETL
 - Herramientas y técnicas
 - Organización de las tablas de dimensión
 - Organización de las tablas de hechos

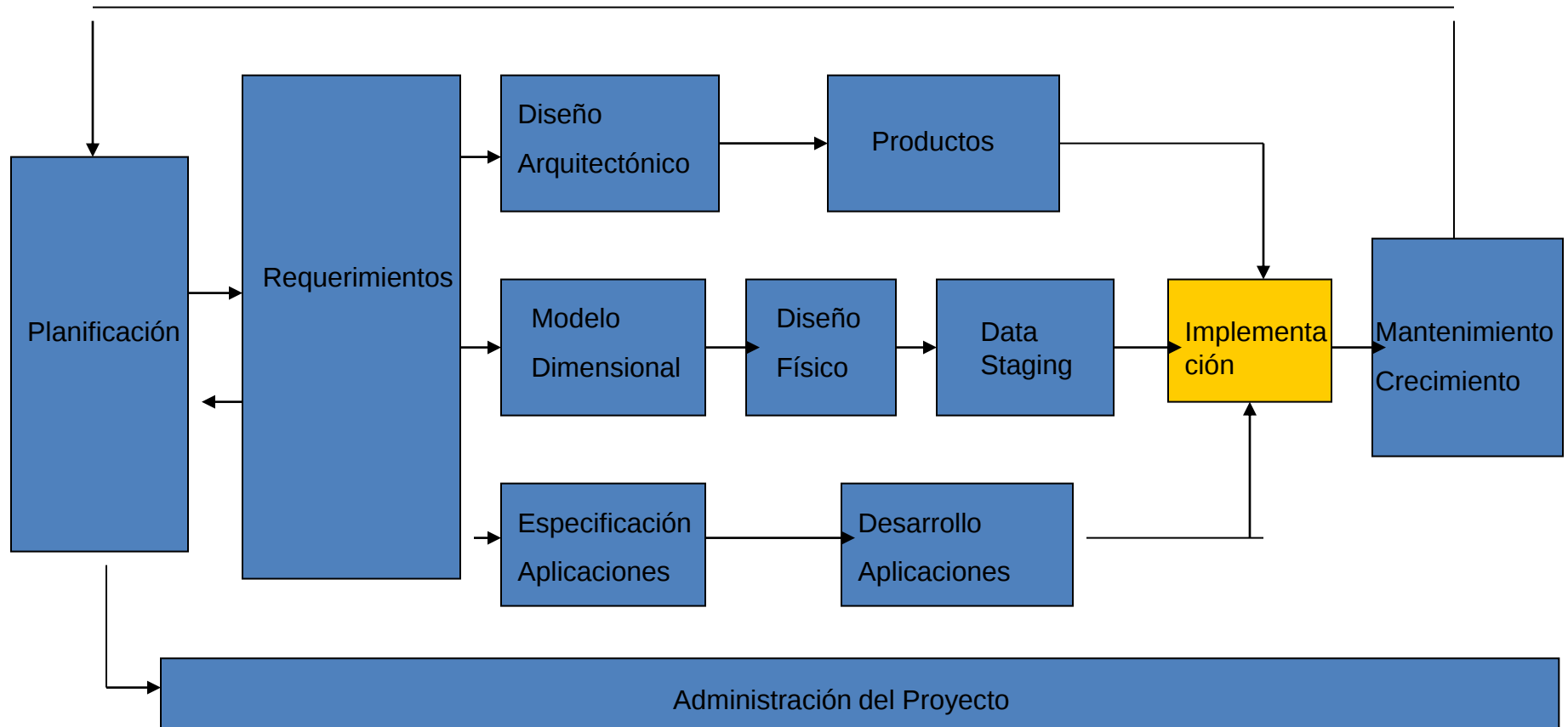
Línea de Aplicaciones



Línea de Aplicaciones

- Especificación y desarrollo de aplicaciones
 - Vías de acceso
 - Internet
 - Correo electrónico
 - Tableros de control
 - Personalización de herramientas

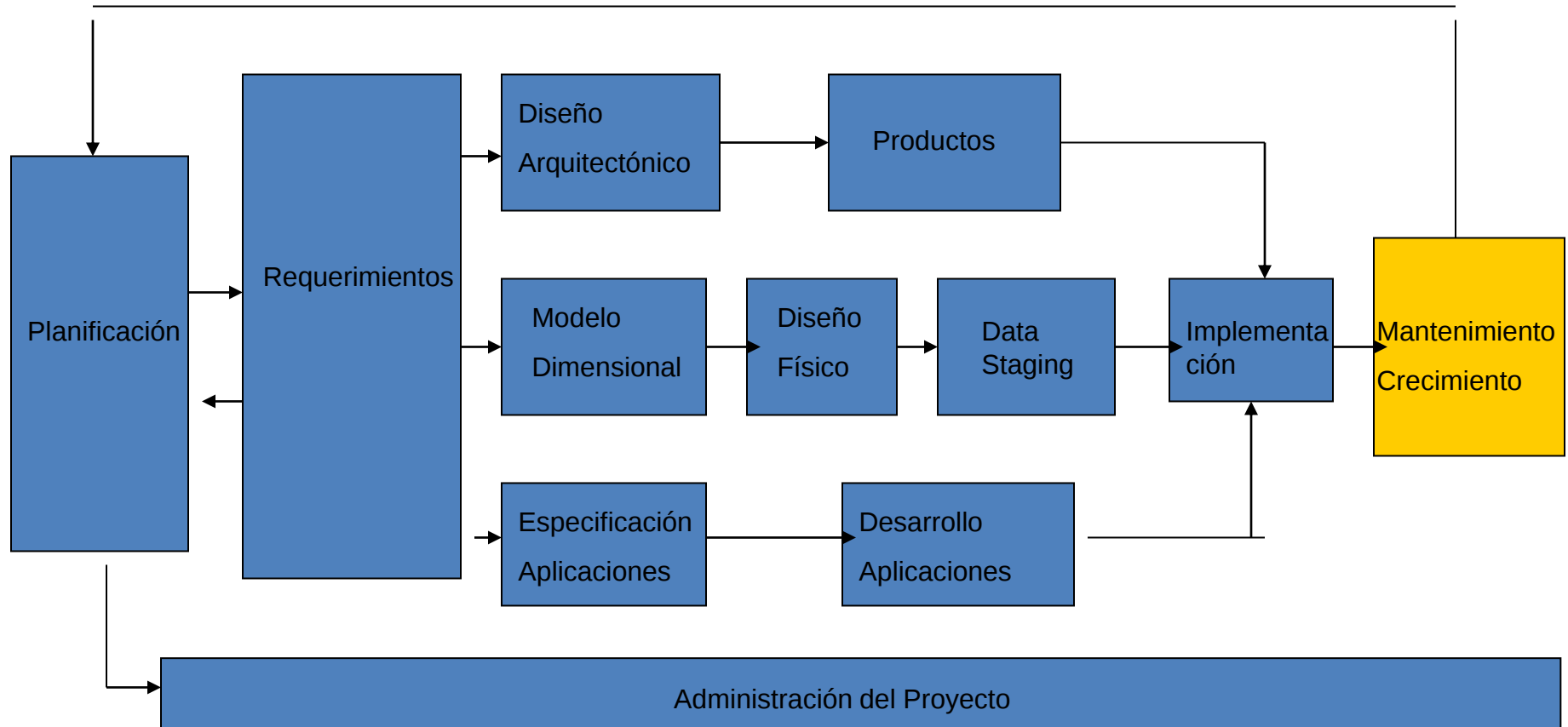
Implementación



Implementación

- Convergencia de las tres líneas
- La línea de datos es la más impredecible
- Asegurarse de que el Data Warehouse esté en condiciones
- Educación: Gestión de conocimiento en la organización

Mantenimiento y Crecimiento



Mantenimiento y Crecimiento

- **Soporte a los usuarios**
 - Si no hay consultas, posiblemente no estén usando el Data Warehouse
 - Detectar áreas de datos o de aplicaciones no cubiertas
 - Calidad del Data Warehouse
- **Educación**
 - Cursos de gestión de conocimiento
 - Usuarios calificados

Mantenimiento y Crecimiento

- **Demandas de crecimiento**
 - Nuevos usuarios
 - Nuevos datos
 - Nuevas aplicaciones
 - Mejoras de las aplicaciones existentes
- **Revisión de las prioridades establecidas**
- **Identificar los productos objetos de gestión de conocimiento organizacional**



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

Modelado de Datos (algunos comentarios finales)

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

Expresiones multidimensionales (MDX)

- Es el acrónimo de MultiDimensional eXpressions
- Es un lenguaje de consulta para bases de datos multidimensionales sobre cubos OLAP
- Fue creado en 1997 por Microsoft. No es un lenguaje estándar, sin embargo diferentes fabricantes de herramientas OLAP lo han adoptado.
- Se utiliza para generar reportes para la toma de decisiones basados en datos históricos, usando la estructura o rotación del cubo

Expresiones multidimensionales (MDX)

- **Una consulta MDX es muy similar a una consulta SQL,**
 - devuelve un conjunto de celdas, que es resultado de tomar un subconjunto de las celdas del cubo original.
- **MDX utiliza en varias situaciones las jerarquías.**
 - Por ejemplo, si una dimensión se denomina región, esta puede contener países. Los países a su vez contienen provincias y las provincias ciudades.
 - Para manejar estos componentes MDX tiene funciones como Children (hijos), cousin (primos) y parents (padres).
- **Su cliente OLAP puede manipular el cubo de distintas formas:**
 - Rotarlo
 - Rebanarlo
 - Cortarlo

Expresiones multidimensionales (MDX)

Consulta MDX Básica:

- Sintaxis:

SELECT <especificación del eje y> on columns,
 <especificación de eje x> on rows

FROM <especificación del cubo>

WHERE <especificación Slicer (rebanador)>

Expresiones multidimensionales (MDX)

```
CREATE CUBE Sales
( DIMENSION Time TYPE TIME,
    HIERARCHY [Fiscal],
        LEVEL [Fiscal Year] TYPE YEAR,
        LEVEL [Fiscal Qtr] TYPE QUARTER,
        LEVEL [Fiscal Month] TYPE MONTH OPTIONS (SORTBYKEY, UNIQUE_KEY),
    HIERARCHY [Calendar],
        LEVEL [Calendar Year] TYPE YEAR,
        LEVEL [Calendar Month] TYPE MONTH,
    DIMENSION Products,
        LEVEL [All Products] TYPE ALL,
        LEVEL Category,
        LEVEL [Sub Category],
        LEVEL [Product Name],
    DIMENSION Geography,
        LEVEL [Whole World] TYPE ALL,
        LEVEL Region,
        LEVEL Country,
        LEVEL City,
    MEASURE [Sales]
        FUNCTION SUM
        FORMAT 'Currency',
    MEASURE [Units Sold]
        FUNCTION SUM
        TYPE DBTYPE_UI4 )
```

Metadatos de un ejemplo

Nombre del cubo: CuboNW

Medidas: Total y Quantity.

Dimensiones:

- Products:

Jerarquias: Category name-Productname.

- Vw_ordenes2:

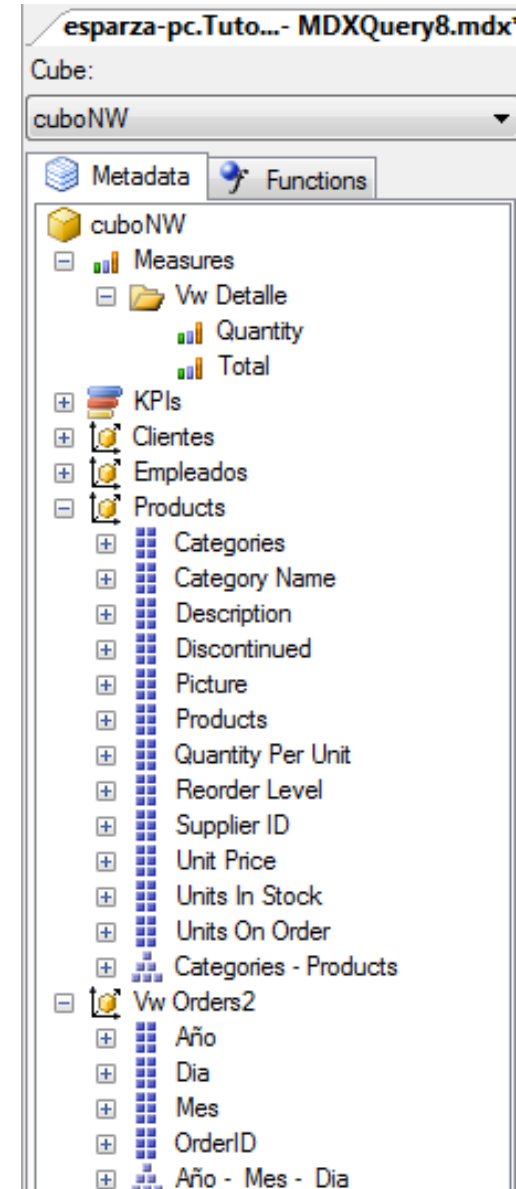
Jerarquias:Año-Mes-Dia

- Clientes.

Jerarquia: Country-Region-City-Company name

- Empleados.

Jerarquia: Empcountry-Empregion-EmpCity-Lastname



Expresión con los nombres de las categorías y todas las medidas.

```
SELECT Measures.MEMBERS ON COLUMNS,  
[category name].MEMBERS ON ROWS  
FROM [cuboNW]
```

Totales y cantidades
para todos los
productos

	Quantity	Total
All	51317	1354458.59
Beverages	9532	286526.95
Condiments	5298	113694.75
Confections	7906	177099.1
Dairy Products	9149	251330.5
Grains/Cereals	4562	100726.8
Meat/Poultry	4199	178188.8
Produce	2990	105268.6
Seafood	7681	141623.09
Unknown	(null)	(null)

Cambio de ejes de los resultados

**SELECT [category name].MEMBERS ON COLUMNS,
Measures.MEMBERS ON ROWS
FROM [cuboNW]**

Results									
	All	Beverages	Condiments	Co...	Dair...	Grai...	Mea...	Pr...	Se...
Quantity	51317	9532	5298	7906	9149	4562	4199	29...	7681
Total	1354458.59	286526.95	113694.75	177...	2513...	1007...	178...	10...	14...

Ejemplos donde se tiene en un de los renglones una dimensión y en las columnas medidas,

Lista de elementos en ejes

ahora vamos a combinar en ambos ejes dos dimensiones: año y las categorías

SELECT

{ [AÑO].members } ON COLUMNS,

[products].[category name].MEMBERS ON ROWS

FROM [cuboNW]

WHERE MEASURES.TOTAL

Es necesario
especificarle en
WHERE la medida que
se desea ver (p. e
TOTAL)

	All	1996	1997	1998
All	1354458.59	226298.5	658388.75	469771....
Beverages	286526.95	53879.2	110424	122223....
Condiments	113694.75	19458.3	59679	34557.45
Confections	177099.1	31511.6	87227.77	58359.73
Dairy Products	251330.5	44615.8	123910.8	82803.9
Grains/Cereals	100726.8	9817.6	60486.95	30422.25
Meat/Poultry	178188.8	30292.2	87621.03	60275.57
Produce	105268.6	15134.2	57718.55	32415.85



Herramientas de inteligencia de negocios

Razones

- Permitir extraer, depurar, consolidar, sintetizar y presentar **datos-información-conocimiento de forma automatizada**
- Posibilitar el **análisis y la toma de decisiones** de manera ágil, flexible y fiable, en el formato adecuado a cada uno, con posibilidad de navegación OLAP
- **Homogeneidad en la utilización de la información**

Son diferentes a las Plataformas de inteligencia de negocios



Herramientas de inteligencia de negocios

Características

- **Cubos de análisis.**
- **Vistas Ad Hoc Query:** navegación por grandes volúmenes de información. Herramientas OLAP para visualizar de manera interactiva grandes volúmenes de información desde varias perspectivas. Debe permitir generar vistas de datos agregadas para mantener a la gerencia informada
- **Cuadros de Mando por indicadores (Balanced Scorecard):** información muy resumida, para la toma de decisiones .
- **Digital Dashboards o paneles de Control Digital:** resúmenes visuales de información del negocio, con métricas e Indicadores Clave de Desempeño (KPIs).
- **Sistemas para Reporting:** Aplicaciones para generar Informes, motores de distribución de reportes para entrega de reportes y alertas, etc.
- **Minería de datos y análisis estadísticos:** modelado predictivo, para diagnóstico (descubrir relaciones causa efecto), etc.



Herramientas código abierto de inteligencia de negocios

- Eclipse BIRT Project: Generador de informes basado en Eclipse
- JasperReports
- LogiReport: Aplicación basada en Web de LogiXML
- OpenI: Aplicación Web simple orientada a OLAP.
- Pentaho
- RapidMiner (antes YALE)
- SpagoBI



Algunas Herramientas Comerciales

- ApeSoft (<http://www.apesoft.es>)
- Bitool: Herramienta de ETL y Visualización
- BiyCloud Smart: QlikView + Cloud + Social Business
- Business Objects (SAP company) | Business Objects
- IBM Cognos
- Microstrategy:
- Oracle BI
- WorkMeter
- Microsoft Office SharePoint Server y PerformancePoint Server
- JetReports





PENTAHO

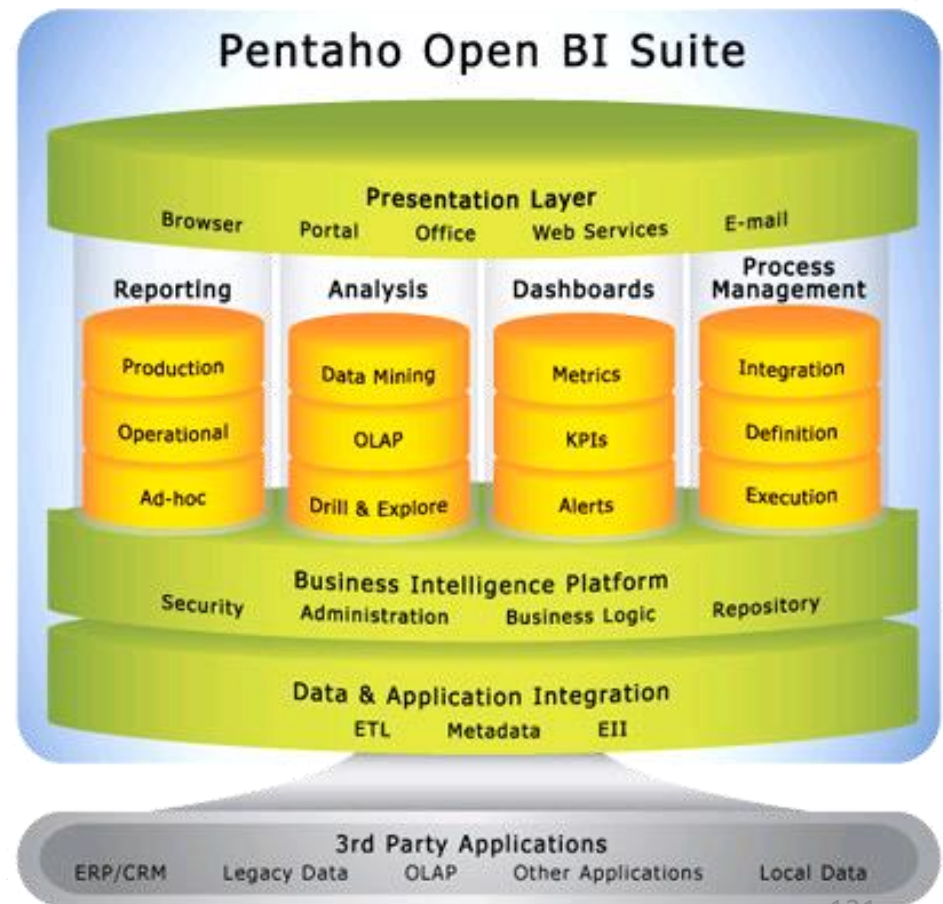
Jose Aguilar
CEMISID, Escuela de Sistemas
Facultad de Ingeniería
Universidad de Los Andes
Mérida, Venezuela

Pentaho: La solución Open Source Business Intelligence

Plataforma de IN que incluye todos los principales componentes requeridos para implementar soluciones.

Pentaho incluye herramientas para hacer:

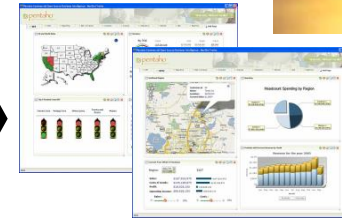
- Reportes
- Análisis
- Visualización (Dashboards)
- Manejo de Datos





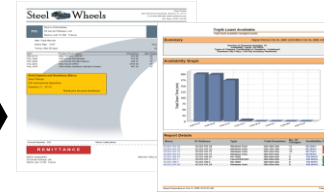
Dashboards

Plataforma Web para
publicar y visualizar la
información



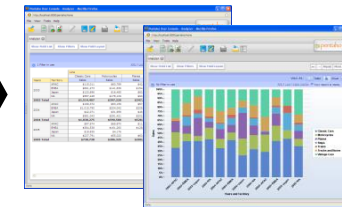
Reportes

Report Designer
Jfree Report: (Motor para
reportes)



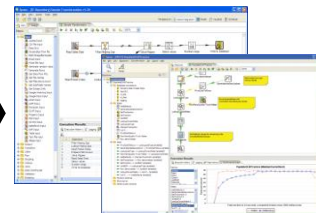
Análisis

Weka
ETL: Kettle (Spoon/Pan)



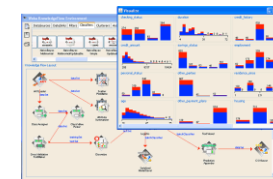
Datos

High Cube designer, WorkBench
ETL: Kettle (Spoon/Pan)
Mondrian: Motor para cubos



Minería

Weka: Motor para
minería de datos
a



Componentes independientes

PENTAHO



- Pentaho ha sido desarrollado desde el año 2004.

Bajo Java

- Open Source:
 - Tiene una comunidad de usuarios
 - Corre bien bajo múltiples plataformas (Windows, Linux, Macintosh, Solaris, etc.



What's New At Pentaho



Login

Valid Users:

Joe ([es_85] Admin) ▼

User:

joe

Password:

••••••••

Login

Reset

Version: Pentaho BI Platform 1.7.1.1117



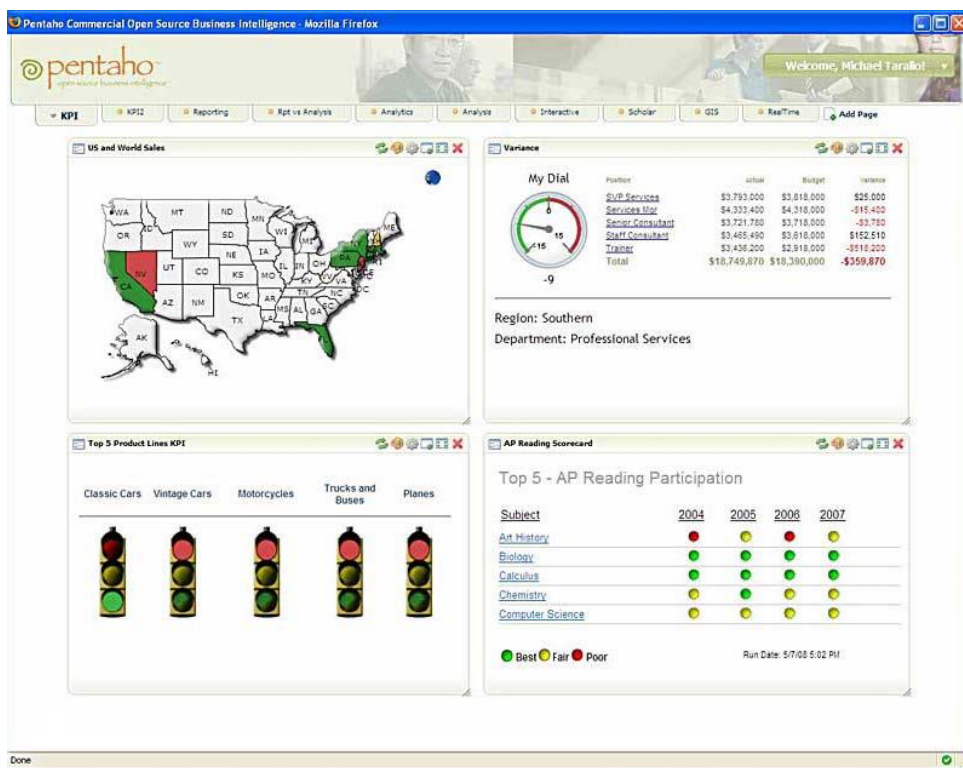
Supplied free of charge with no support, no certification, no maintenance, no warranty and no indemnity by Pentaho or its Certified Partners. [Click here](#) for Support, Certification, or Indemnity. Copyright © 2004 - 2008 Pentaho Corporation. All rights reserved.

Pentaho Dashboard

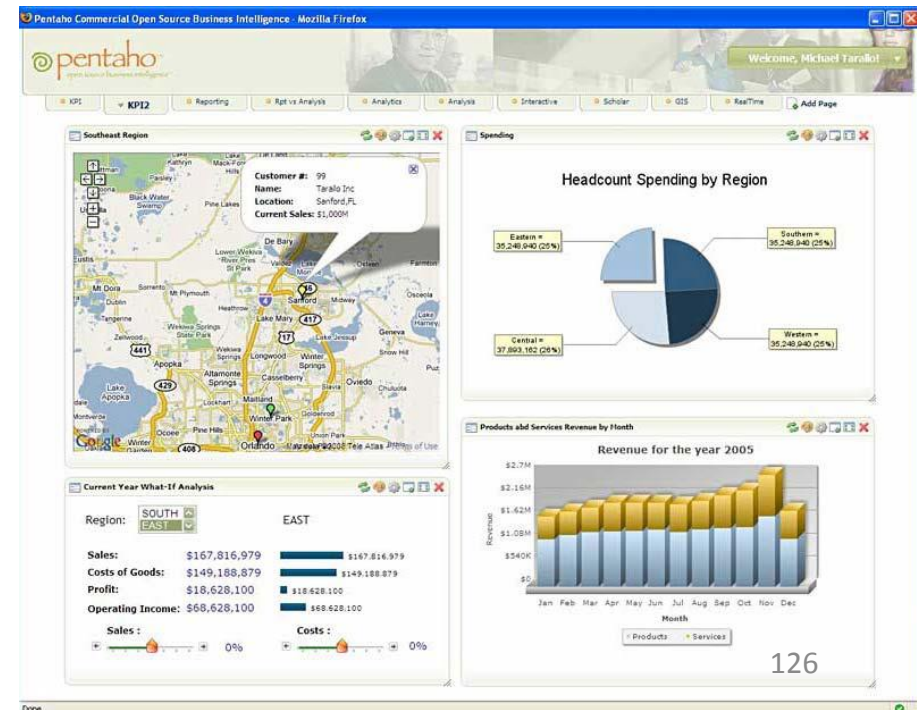


- Brinda a los usuarios la información crítica que necesitan para comprender y mejorar el desempeño organizacional.
- Proporciona una visión inmediata del rendimiento individual, departamental, o de la organización, mediante la entrega de métricas claves a través de una interfaz visual atractiva e intuitiva.

Pentaho Dashboard



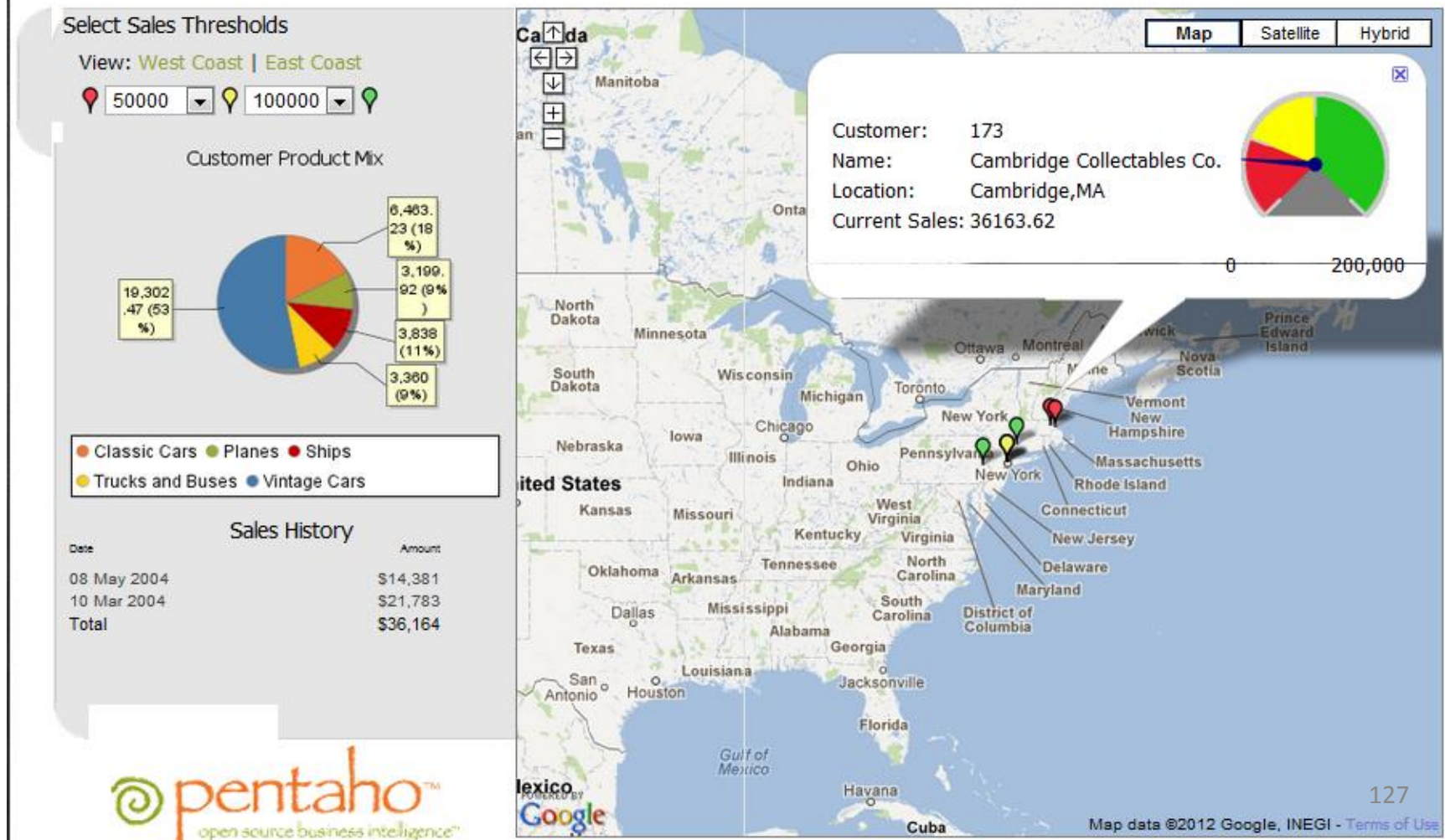
Interactive



Pentaho Dashboard: Geo Location



Pentaho Google Maps Dashboard



Reporting



La solución de Reporting que plantea Pentaho, incluida dentro de su suite, es

Jfree Report.

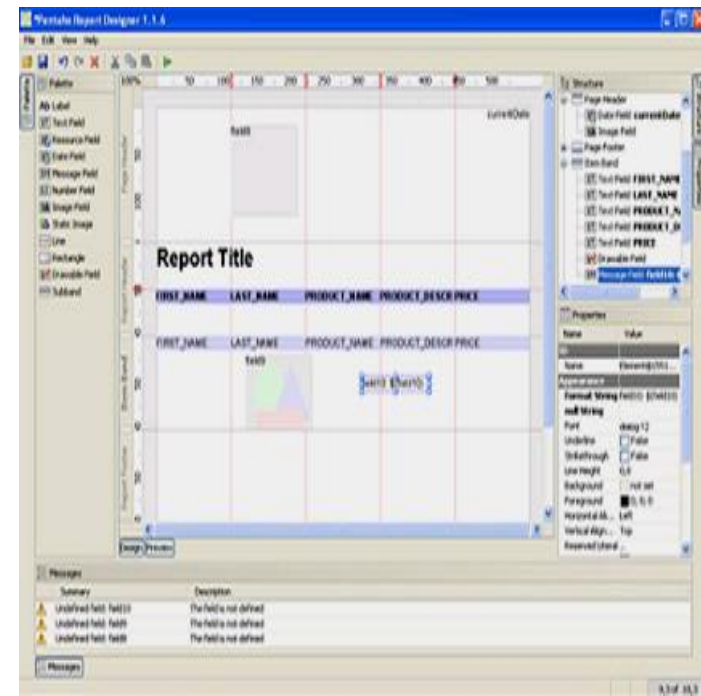
Region: Central

Department	Position	Actual	Budget	Variance
Executive Management				
	SVP Partnerships	\$367,415	\$392,100	\$24,685
	SVP WW Operations	\$476,000	\$725,887	\$249,887
	SVP Strategic Development	\$383,242	\$403,405	\$20,163
	CEO	\$549,625	\$522,250	-\$27,375
	Total	\$1,776,282	\$2,043,642	\$267,360

Department	Position	Actual	Budget	Variance
Finance				
	Controller	\$570,373	\$577,070	\$6,697
	Payroll	\$367,415	\$432,100	\$64,685
	Administrative Assistant	\$827,861	\$760,990	-\$66,871
	IS	\$570,759	\$577,346	\$6,587
	CFO	\$770,272	\$719,855	-\$50,417
	Total	\$3,106,680	\$3,067,361	-\$39,319

Department	Position	Actual	Budget	Variance
Human Resource				
	Sexual Harassment	\$530,473	\$538,570	\$8,097
	EOE	\$530,207	\$538,380	\$8,173
	HR Generalists	\$856,190	\$771,225	-\$84,965
	HR Training	\$397,473	\$443,570	\$46,097
	Administration	\$549,625	\$552,250	\$2,625
	SVP HR	\$574,896	\$570,300	-\$4,596
	Total	\$3,438,863	\$3,414,295	-\$24,568

Department	Position	Actual	Budget	Variance
Marketing & Communication				
	Graphics	\$782,375	\$728,500	-\$53,875
	Writer	\$405,985	\$459,650	\$53,665
	Analyst Relations	\$383,375	\$443,500	\$60,125
	Press Relations	\$497,296	\$524,872	\$27,576
	CMO	\$827,861	\$760,990	-\$66,871
	Product Marketing Mgr	\$693,531	\$665,040	-\$28,491



Reporting



Permite a las organizaciones fácilmente acceder, formatear y entregar reportes a los empleados, clientes y socios.

Product Lines by Territory - Report Pack

Shipped Orders by Territory - Report Pack

Product Lines within Territory

Territory: APAC -

Product Line	Total	Quantity Ordered
Ships	\$38,393.48	428
Trains	\$9,907.07	139
Classic Cars	\$411,956.24	3,852
Trucks and Buses	\$145,665.69	1,380
Planes	\$121,426.20	1,330
Vintage Cars	\$364,538.92	3,897
Motorcycles	\$189,818.23	1,852
Total	\$1,281,705.83	12,878

Territory: EMEA -

Product Line	Total	Quantity Ordered
Trucks and Buses	\$500,978.82	4,655
Trains	\$138,506.18	1,544
Planes	\$491,654.86	5,513
Ships	\$427,264.68	4,808
Vintage Cars	\$851,080.90	9,660

Shipped Orders by Territory

Countries are counted per Territory

Quantity Ordered & Total Revenue are being Totalled

Territory : NA

Country	Quantity Ordered	Total
USA	32,923	3,372,204.3
Canada	2,293	224,078.56
2	35,216	\$3,596,282.86

Territory : Japan

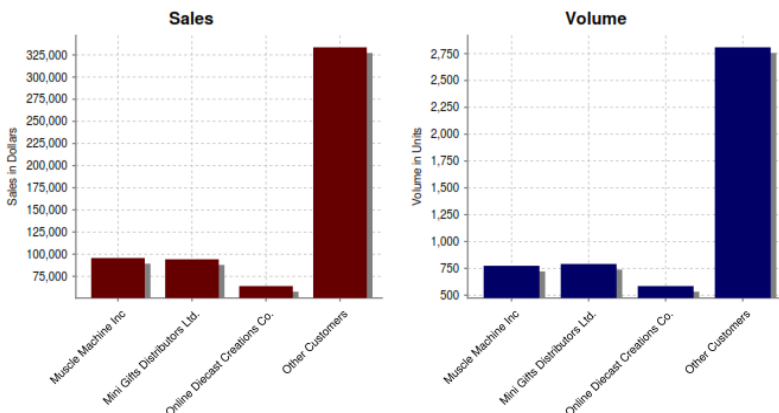
Country	Quantity Ordered	Total
Singapore	1,524	172,989.68
Philippines	961	94,015.73
Japan	1,842	188,167.81
Hong Kong	596	48,794.36
4	4,923	\$503,957.58

Territory : EMEA

Country	Quantity Ordered	Total
UK	4,584	428,472.21
Switzerland	1,078	117,713.56
Sweden	1,239	135,043.08

Steel

Customer	Sales	Volume
Muscle Machine Inc	\$ 53,981	585
Mini Gifts Distributors Ltd.	\$ 333,579	2,809
Online Diecast Creations Co.	\$ 587,428	4,959
Other Customers		
Total		



Mondrian permite

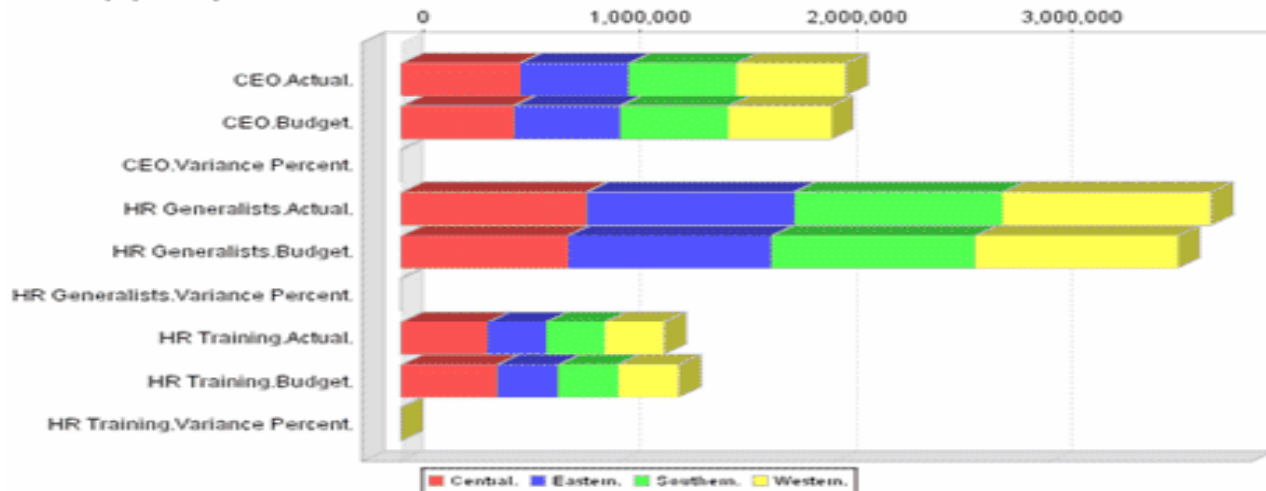
- Alto rendimiento, análisis interactivo de volúmenes grandes o pequeños de la información
- Exploración «dimensional" de datos, por ejemplo, el análisis de las ventas por línea de productos, por regiones, para un período de tiempo
- Convertir consultas multidimensionales en MDX en SQL para hacer consultas
- Consultas de alta velocidad
- Cálculos avanzados utilizando expresiones de cálculo del lenguaje MDX

Mondrian permite

		Region			
Positions	Measures	Central	Eastern	Southern	Western
CEO	Actual	549,625.00	500,000.00	500,000.00	500,000.00
	Budget	522,250.00	488,750.00	498,750.00	478,750.00
	Variance Percent	-5.24%	-2.30%	-.25%	-4.44%
HR Generalists	Actual	856,190.00	961,000.00	961,000.00	961,000.00
	Budget	771,225.00	940,158.00	940,158.00	938,158.00
	Variance Percent	-11.02%	-2.22%	-2.22%	-2.43%
HR Training	Actual	397,473.00	271,200.00	271,200.00	271,200.00
	Budget	443,570.00	279,674.00	279,674.00	277,674.00
	Variance Percent	10.39%	3.03%	3.03%	2.33%

Slicer: [(All)=All Departments]

Slicer: (All)=All Departments





OLAP en Pentaho

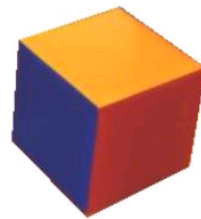
Permite realizar todas las **típicas funcionalidades de un sistema OLAP** a una gran velocidad

Hace uso de diversas Bases de Datos:

- Oracle
- DB2
- SQL-Server
- MySQL
- Postgre



OLAP en Pentaho



Pentaho OLAP

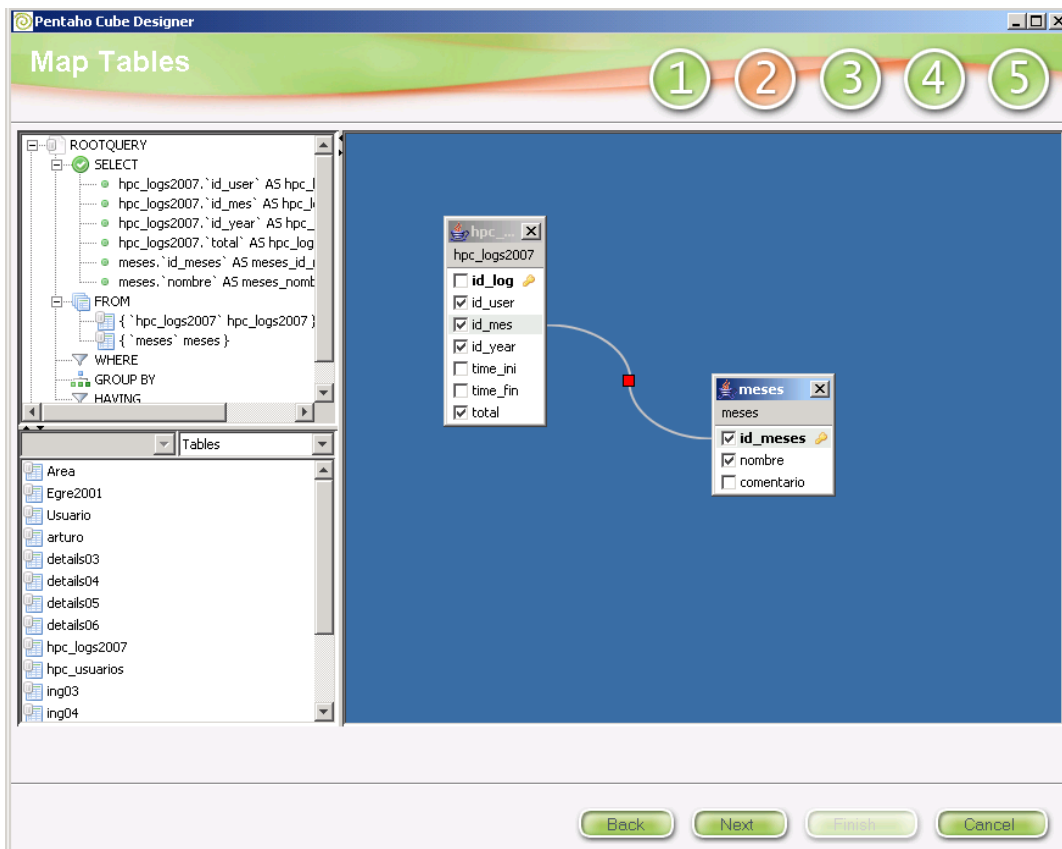
Puntos Acumulados



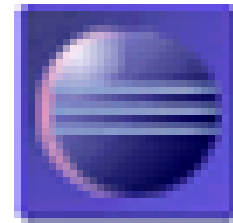
Permite diseñar cubos dinámicos, para un análisis de la información mucho más rápida y detallada, que apoya la Toma de Decisiones dentro de una organización



OLAP en Pentaho



Cube Designer
incorpora “arrastrar y
soltar” para un
manejo más fácil.





OLAP en Pentaho

Create Dimensions

1
2
3
4
5

Create Dimensions

Create dimensions from the source fields using "Add New Dimension" button. You can add levels by using the Arrow buttons. Once dimensions are created, dimension/hierarchy/level to view/change the attributes at the bottom section. Click "Add Property" button to add member properties for a level.

Source fields

hpc_logs2007.id_user
 hpc_logs2007.id_mes
 hpc_logs2007.id_year
 hpc_logs2007.total
 meses.id_meses
meses.nombre

Add New Dimension

==>

<==

Dimensions

hpc_logs2007.id_user

hpc_logs2007.id_user

hpc_logs2007.id_year

hpc_logs2007.id_year

meses.nombre

hpc_logs2007.id_user

hpc_logs2007.id_year

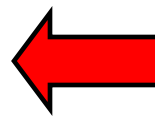
meses.nombre

Attributes and Properties

Add Property

Remove Property

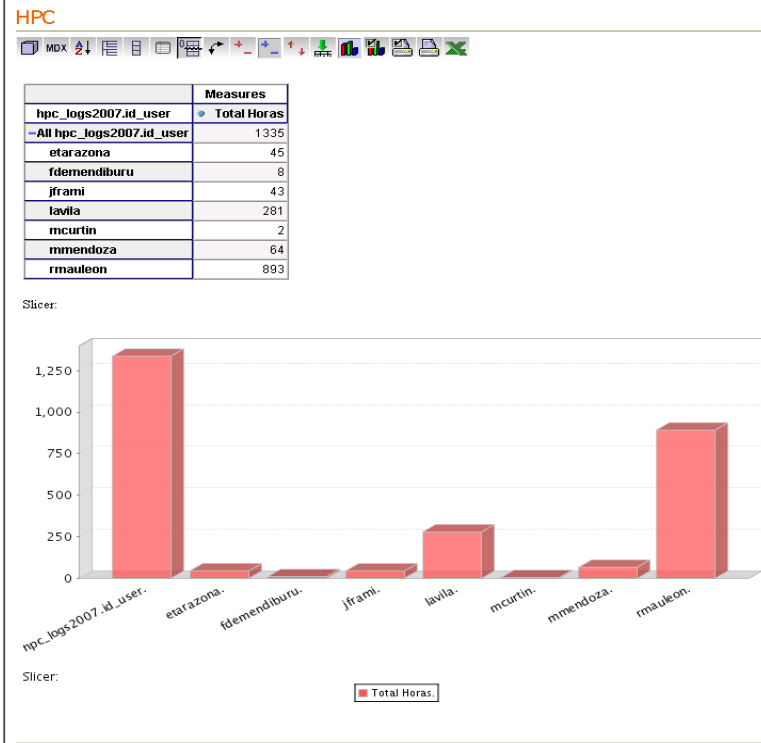
Property	Value
name	meses.nombre
nameColumn	meses.nombre
uniqueMembers	false



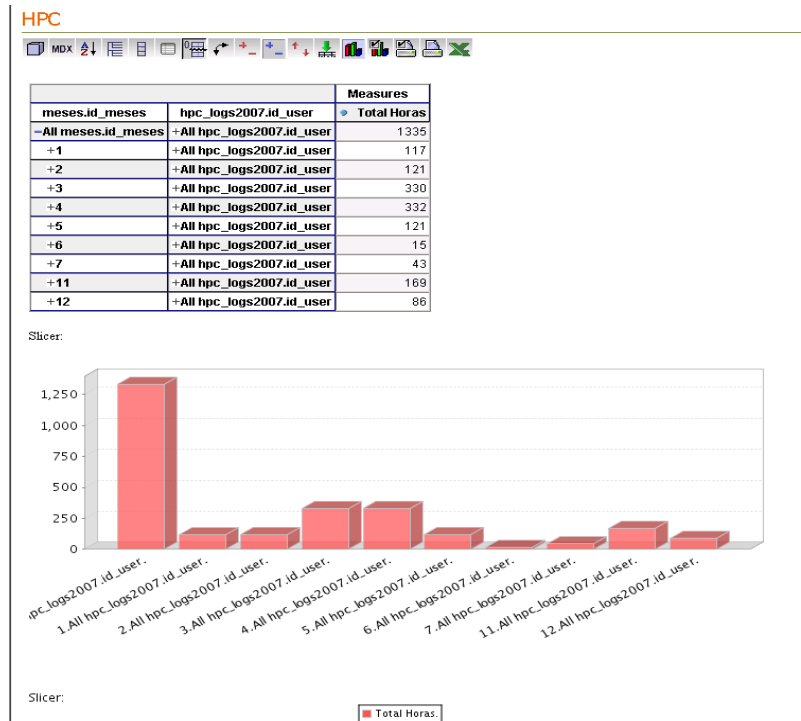
**Crea las
dimensiones del
Cubo que luego
seran mostradas en
la Suite de Pentaho**



Gráficas de Resultados



Estadística por Usuario



Estadística por Mes



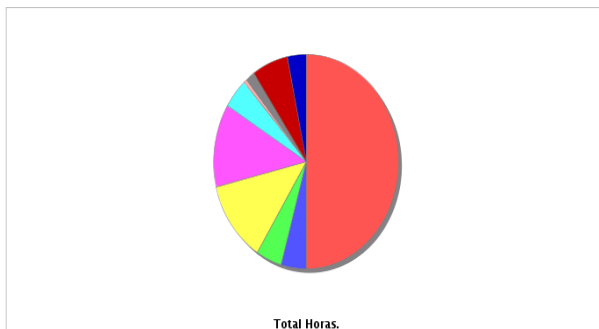
Gráficas de Resultados

HPC



meses.id_meses	hpc_logs2007.id_user	Measures
-All meses.id_meses	+All hpc_logs2007.id_user	Total Horas
+1	+All hpc_logs2007.id_user	117
+2	+All hpc_logs2007.id_user	121
+3	+All hpc_logs2007.id_user	330
+4	+All hpc_logs2007.id_user	332
+5	+All hpc_logs2007.id_user	121
+6	+All hpc_logs2007.id_user	15
+7	+All hpc_logs2007.id_user	43
+11	+All hpc_logs2007.id_user	169
+12	+All hpc_logs2007.id_user	86

Slicer:



Total Horas.

Slicer:

All meses.id_meses: All hpc_logs2007.id_user, 1 All hpc_logs2007.id_user, 2 All hpc_logs2007.id_user, 3 All hpc_logs2007.id_user,
 4 All hpc_logs2007.id_user, 5 All hpc_logs2007.id_user, 6 All hpc_logs2007.id_user, 7 All hpc_logs2007.id_user,
 11 All hpc_logs2007.id_user, 12 All hpc_logs2007.id_user.

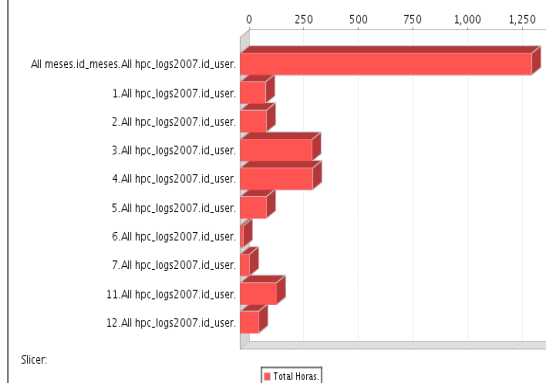
*Todos los gráficos
y estadísticas se
generan
dinamicamente,
sólo con solicitar
la información los
gráficos van
cambiando*

HPC



meses.id_meses	hpc_logs2007.id_user	Measures
-All meses.id_meses	+All hpc_logs2007.id_user	Total Horas
+1	+All hpc_logs2007.id_user	117
+2	+All hpc_logs2007.id_user	121
+3	+All hpc_logs2007.id_user	330
+4	+All hpc_logs2007.id_user	332
+5	+All hpc_logs2007.id_user	121
+6	+All hpc_logs2007.id_user	15
+7	+All hpc_logs2007.id_user	43
+11	+All hpc_logs2007.id_user	169
+12	+All hpc_logs2007.id_user	86

Slicer:



Slicer:

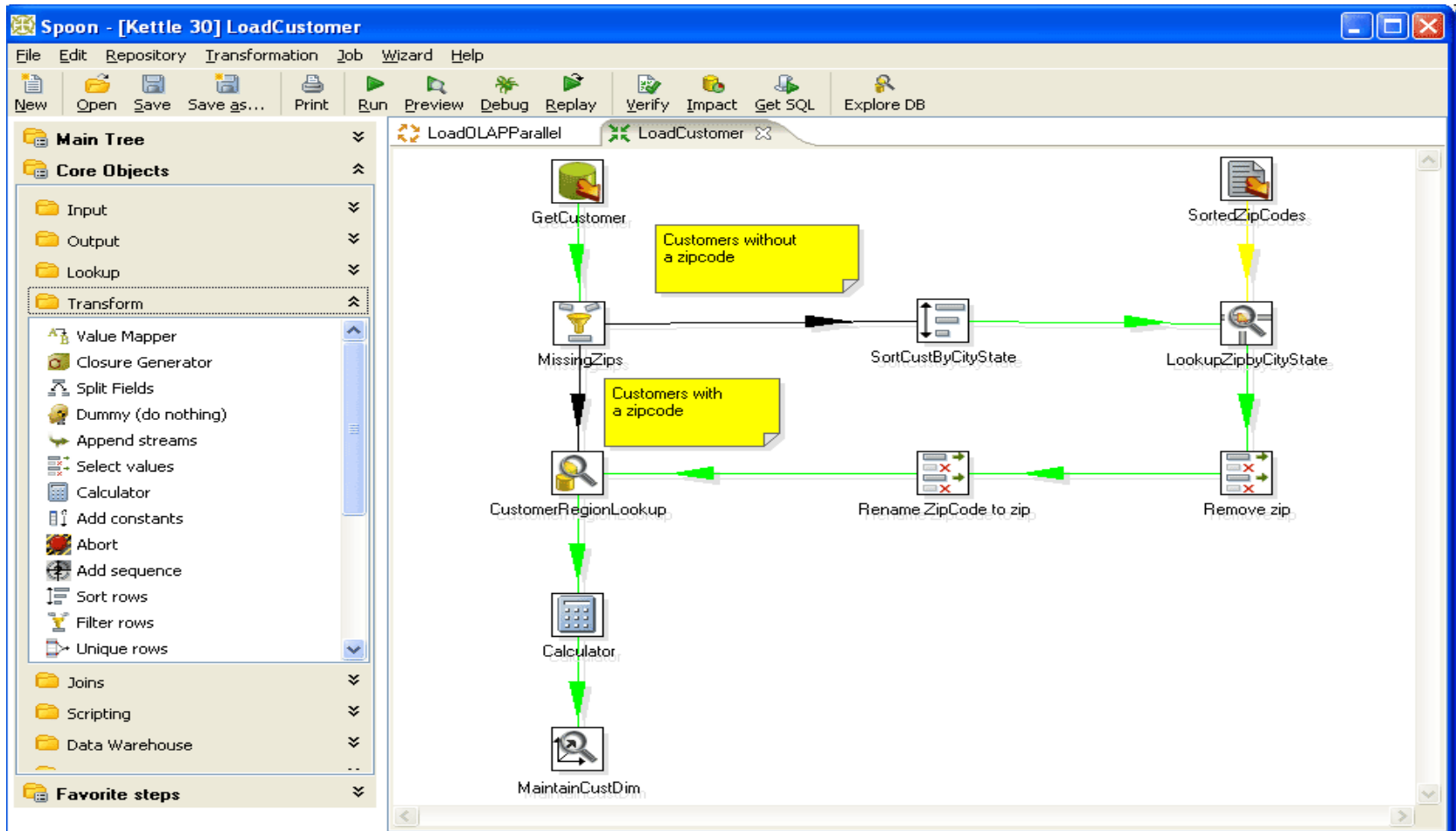
Kettle

- tiene capacidades de acceder, limpiar e integrar datos desde cualquier lugar de la organización.
- ofrece una poderosa mecanismo de extracción, transformación y carga (ETL), con un diseño intuitivo y una arquitectura escalable basada en estándares probados.

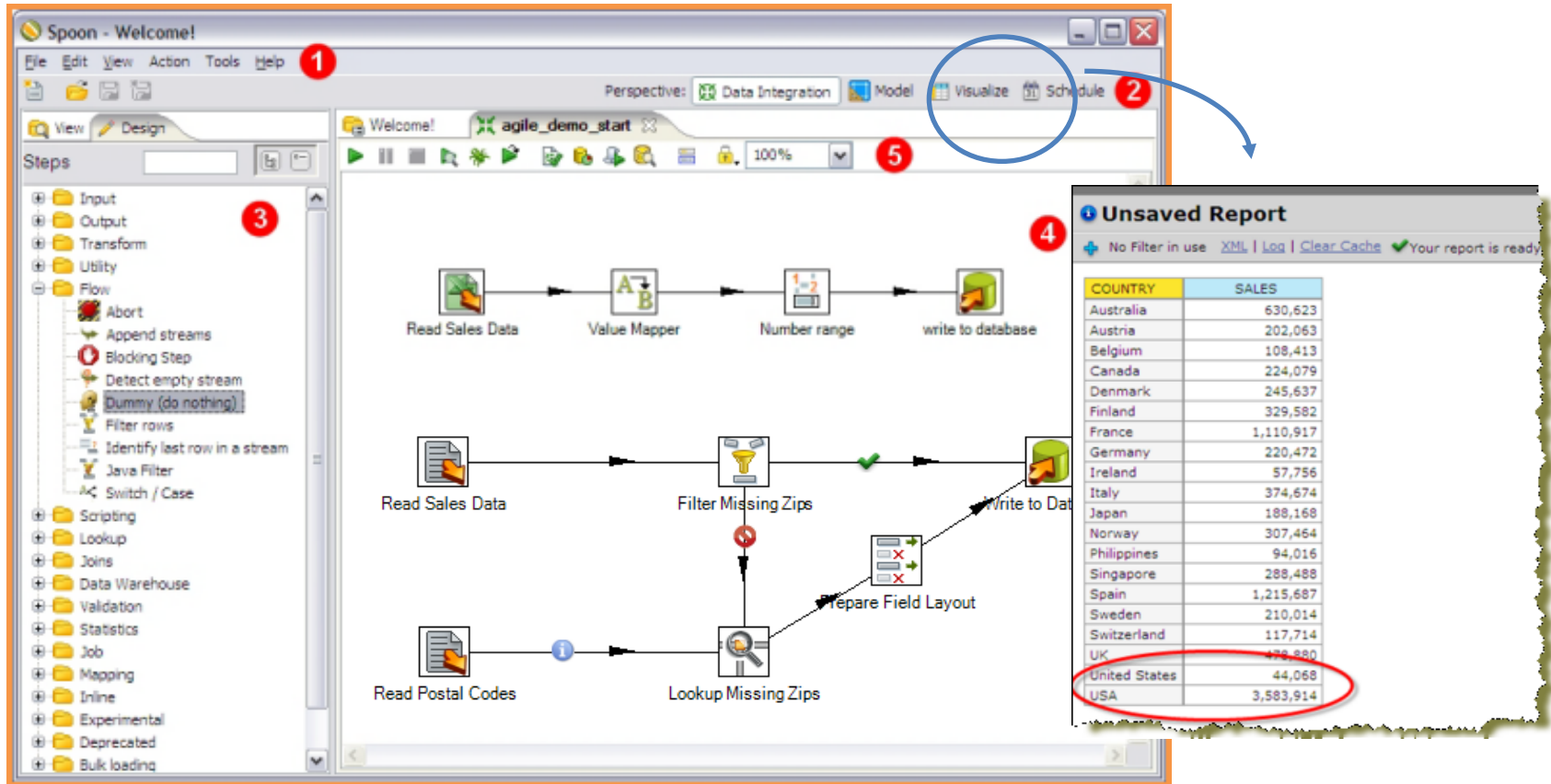
Suite

- Kitchen : Job Execution
- Spoon : GUI Designer
- Carte : Cluster Web Server
- Pan : Transformation Execution

Kettle



Integración de datos

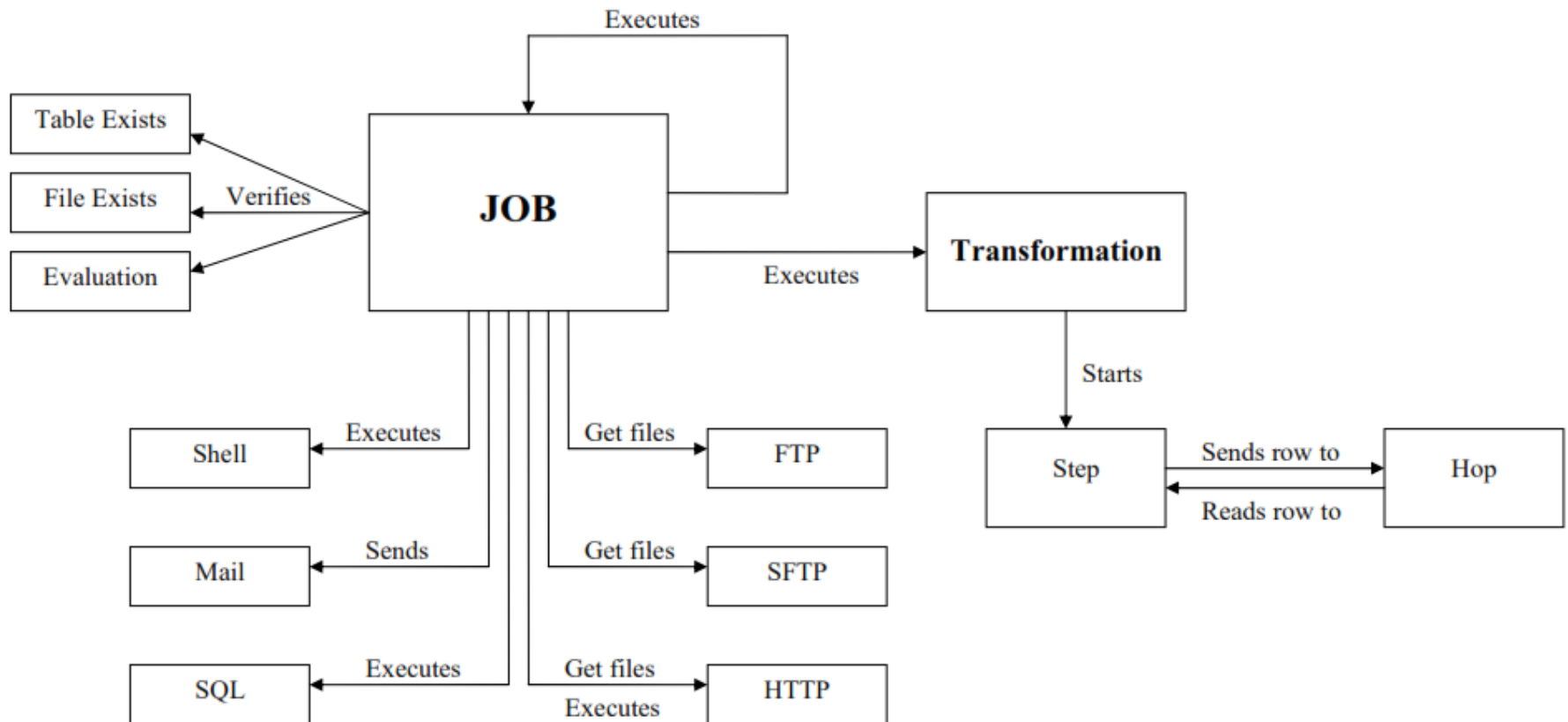


The screenshot displays the Pentaho Spoon interface for data integration. The main workspace shows a job design with two parallel paths. The top path consists of 'Read Sales Data', 'Value Mapper', 'Number range', and 'write to database'. The bottom path consists of 'Read Postal Codes', 'Lookup Missing Zips', 'Filter Missing Zips', 'Repare Field Layout', and 'Write to Dat'. The 'Filter Missing Zips' step is highlighted with a red circle and a blue arrow pointing to the 'Visualize' button in the top toolbar. The 'Visualize' button is also circled in blue. The 'Steps' panel on the left lists various transformation and job steps, with 'Dummy (do nothing)' selected. The 'Unsaved Report' window on the right shows a table of sales data by country, with the 'United States' and 'USA' rows circled in red.

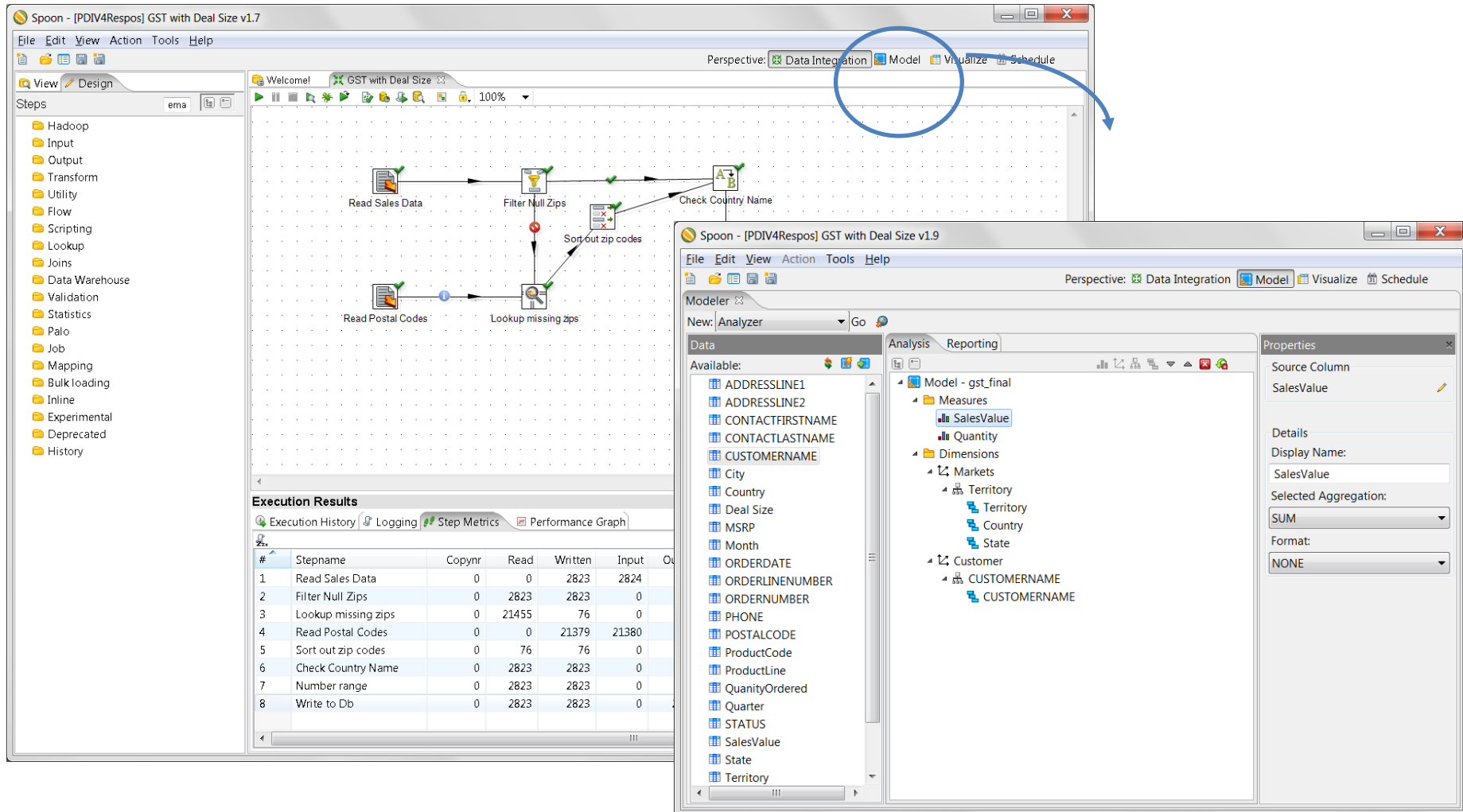
Unsaved Report

COUNTRY	SALES
Australia	630,623
Austria	202,063
Belgium	108,413
Canada	224,079
Denmark	245,637
Finland	329,582
France	1,110,917
Germany	220,472
Ireland	57,756
Italy	374,674
Japan	188,168
Norway	307,464
Philippines	94,016
Singapore	288,488
Spain	1,215,687
Sweden	210,014
Switzerland	117,714
UK	458,880
United States	44,068
USA	3,583,914

Modelo de Kettle



Integración de datos: crear metadatos



The screenshot displays the Pentaho Spoon interface for a data integration job titled "GST with Deal Size v1.7". The job is composed of several steps: Read Sales Data, Filter Null Zips, Sort out zip codes, Read Postal Codes, Lookup missing zips, Check Country Name, and Write to Db. The job is currently in the "Model" perspective, which is highlighted by a blue circle and an arrow.

The "Modeler" window is open, showing the "Data" tab. The "Available" list on the left contains various fields from the source data, including ADDRESSLINE1, ADDRESSLINE2, CONTACTFIRSTNAME, CONTACTLASTNAME, CUSTOMERNAME, City, Country, Deal Size, MSRP, Month, ORDERDATE, ORDERLINENUMBER, ORDERNUMBER, PHONE, POSTALCODE, ProductCode, ProductLine, QuantityOrdered, Quarter, STATUS, SalesValue, State, and Territory.

The "Analysis" tab is also visible, showing a hierarchical structure for the "Model - gst_final" job. The "Measures" section includes "SalesValue", and the "Dimensions" section includes "City", "Country", "State", and "Customer". The "Properties" window on the right shows the "Source Column" set to "SalesValue" and the "Selected Aggregation" set to "SUM".

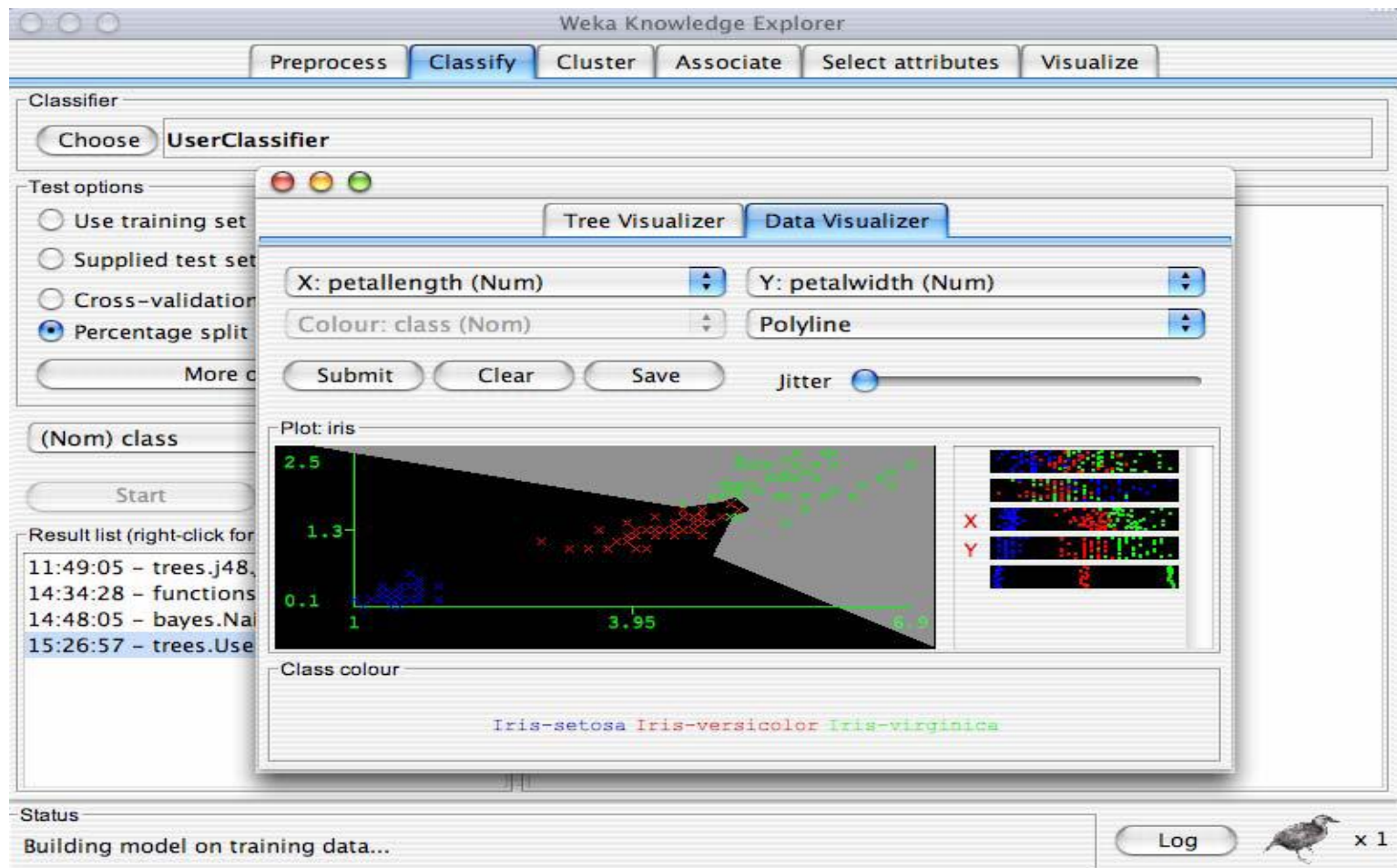
The "Execution Results" window at the bottom shows the job's execution history, including the number of rows read, written, and input for each step.

#	Stepname	Copynr	Read	Written	Input	Output
1	Read Sales Data	0	0	2823	2824	
2	Filter Null Zips	0	2823	2823	0	
3	Lookup missing zips	0	21455	76	0	
4	Read Postal Codes	0	0	21379	21380	
5	Sort out zip codes	0	76	76	0	
6	Check Country Name	0	2823	2823	0	
7	Number range	0	2823	2823	0	
8	Write to Db	0	2823	2823	0	

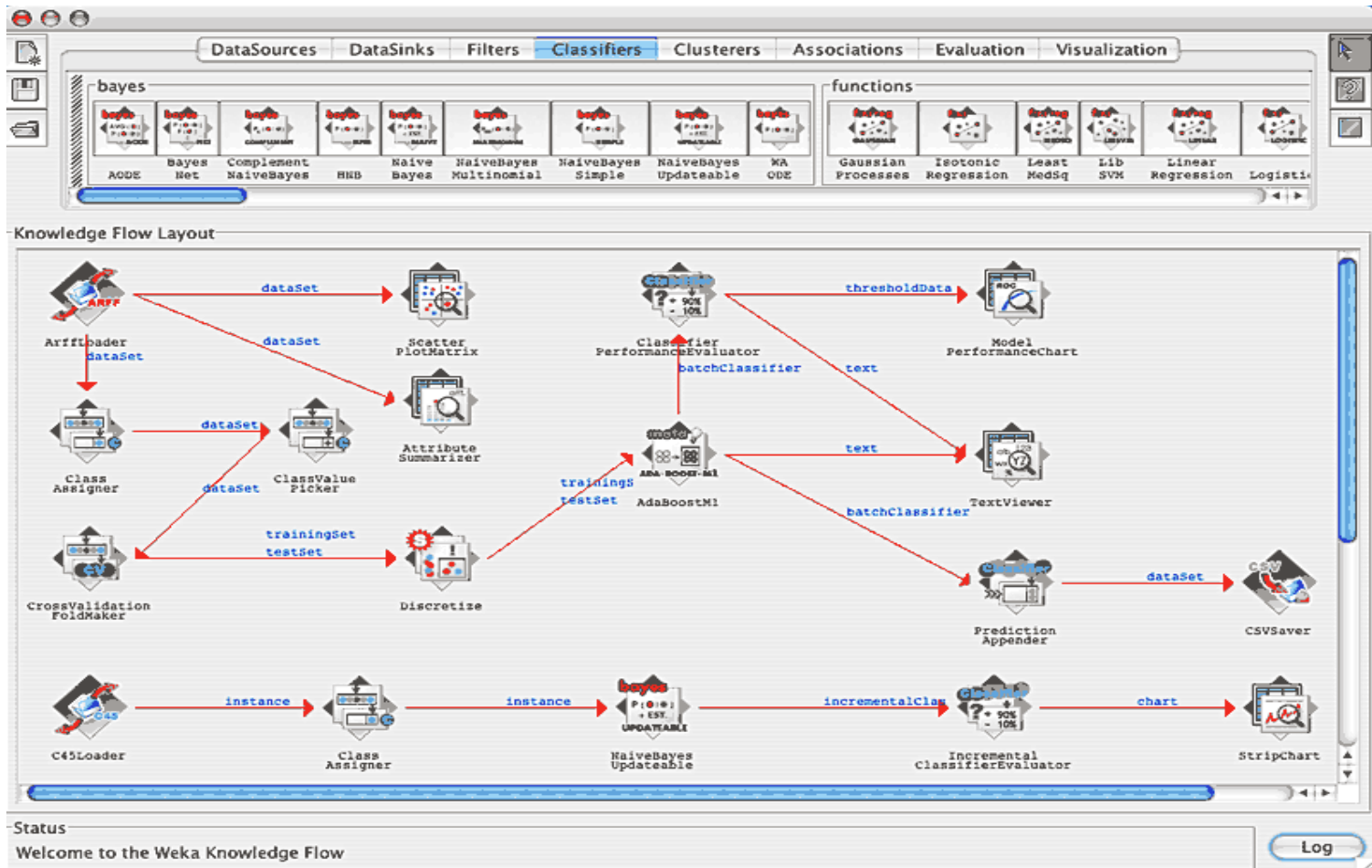
Weka

- Descubrir relaciones ocultas en los datos y descubrir indicadores de desempeño futuro.
- Análisis predictivo y exploración de correlaciones en los datos para mejorar el desempeño organizacional

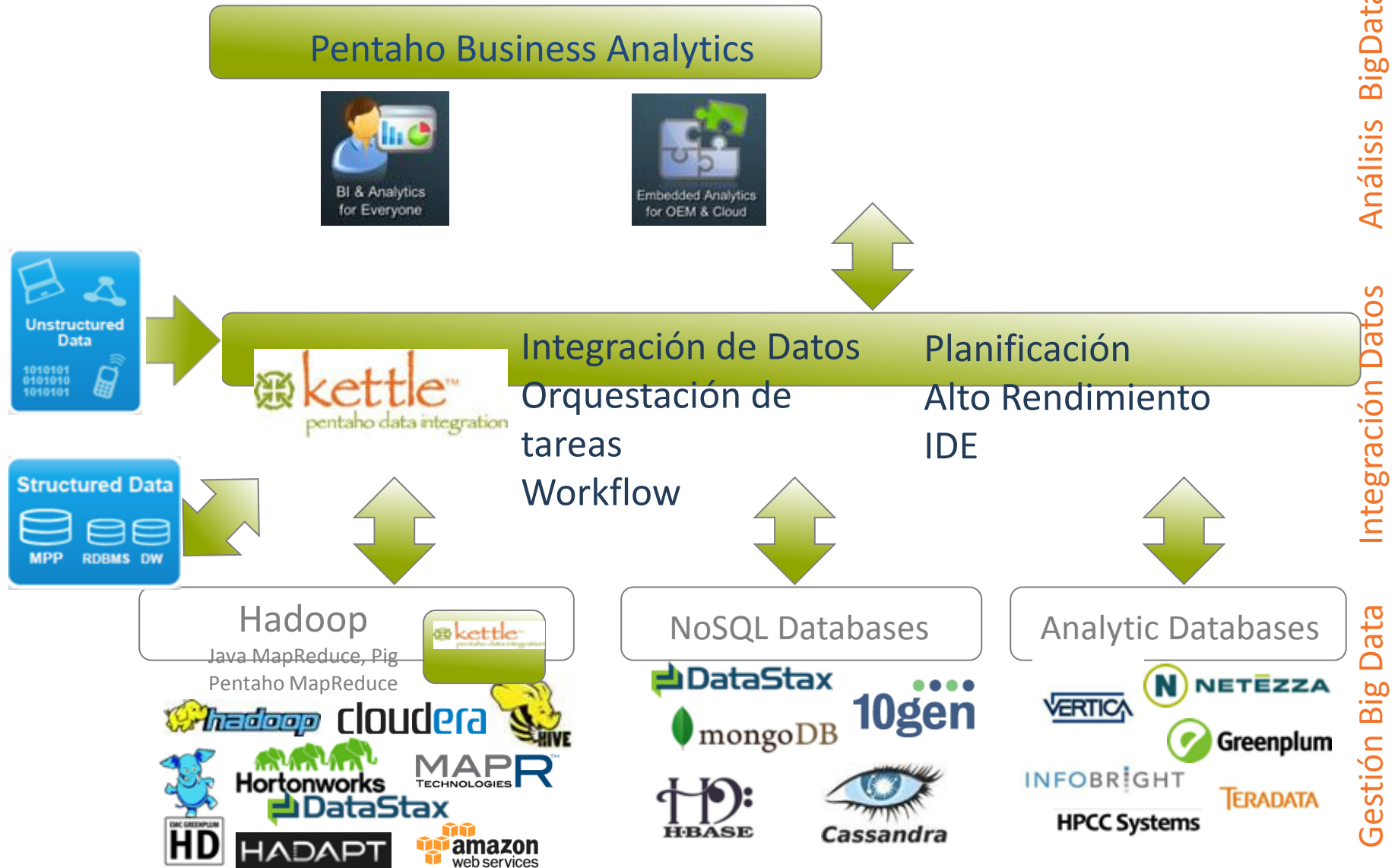
Weka



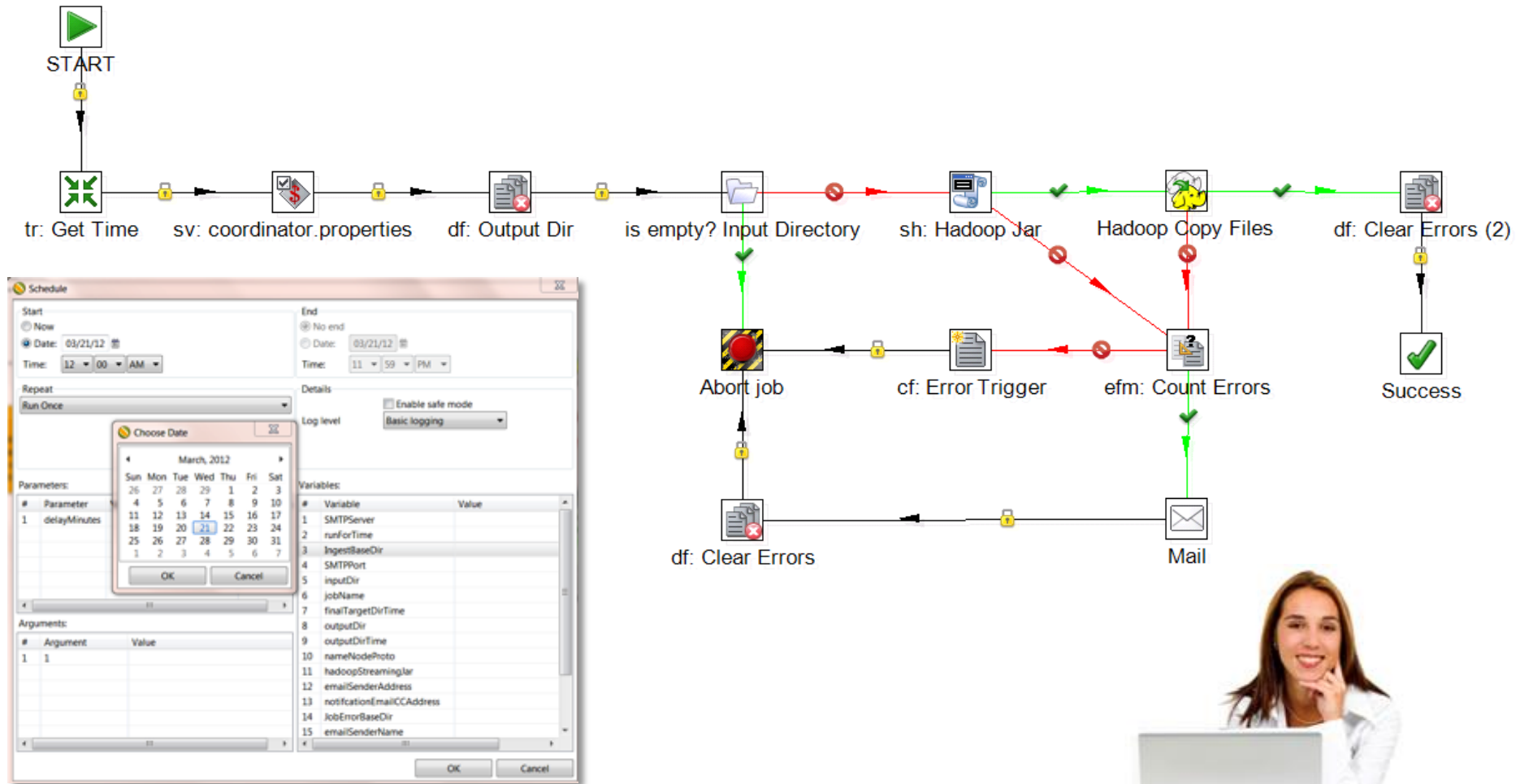
Weka



Pentaho y Big Data



Orquestación virtual de tareas con diferentes fuentes de datos



Planificación



Iteration 2

- Definir las fuentes de datos disponibles
- Diseñar el modelo de datos dimensional para el proyecto de IN
- Especificar y realizar las operaciones ETL para construir el DW (Data Marts)
- Diseñar las operaciones OLAP requeridas para el proyecto de IN: métricas e indicadores clave de desempeño, reportes y alertas.
- Hacer pruebas sobre el Data Marts (tanto de las operaciones diseñadas como otras consultas)

