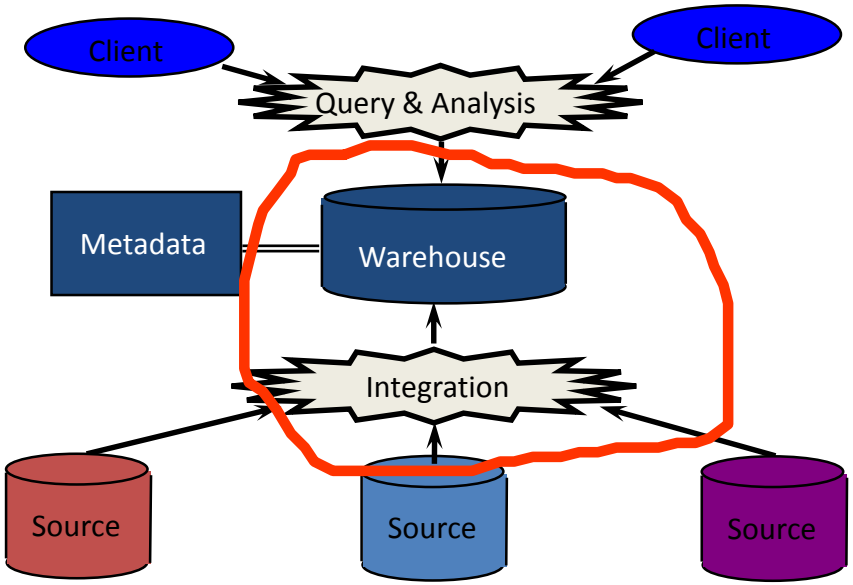
UNIVERSIDAD
DE LOS ANDES
MERIDA VENEZUELA

Datawarehousing: ETL

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

Integración

- Selección de los datos
- Transformación de datos
- Carga de datos



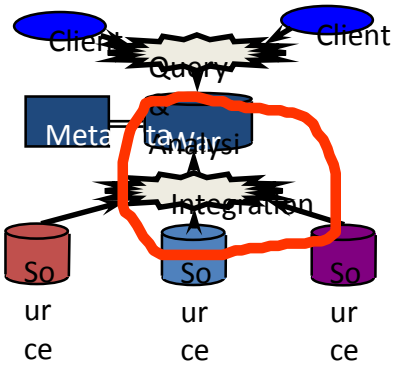
Proceso ETL

ETL (Extracción, Transformación y Carga)

Extracción: Obtención de información de las distintas fuentes, tanto internas como externas.

Transformación: Filtrado, limpieza, depuración, homogeneización y agrupación de la información.

Carga: Organización y actualización de los datos y los metadatos en el DW.

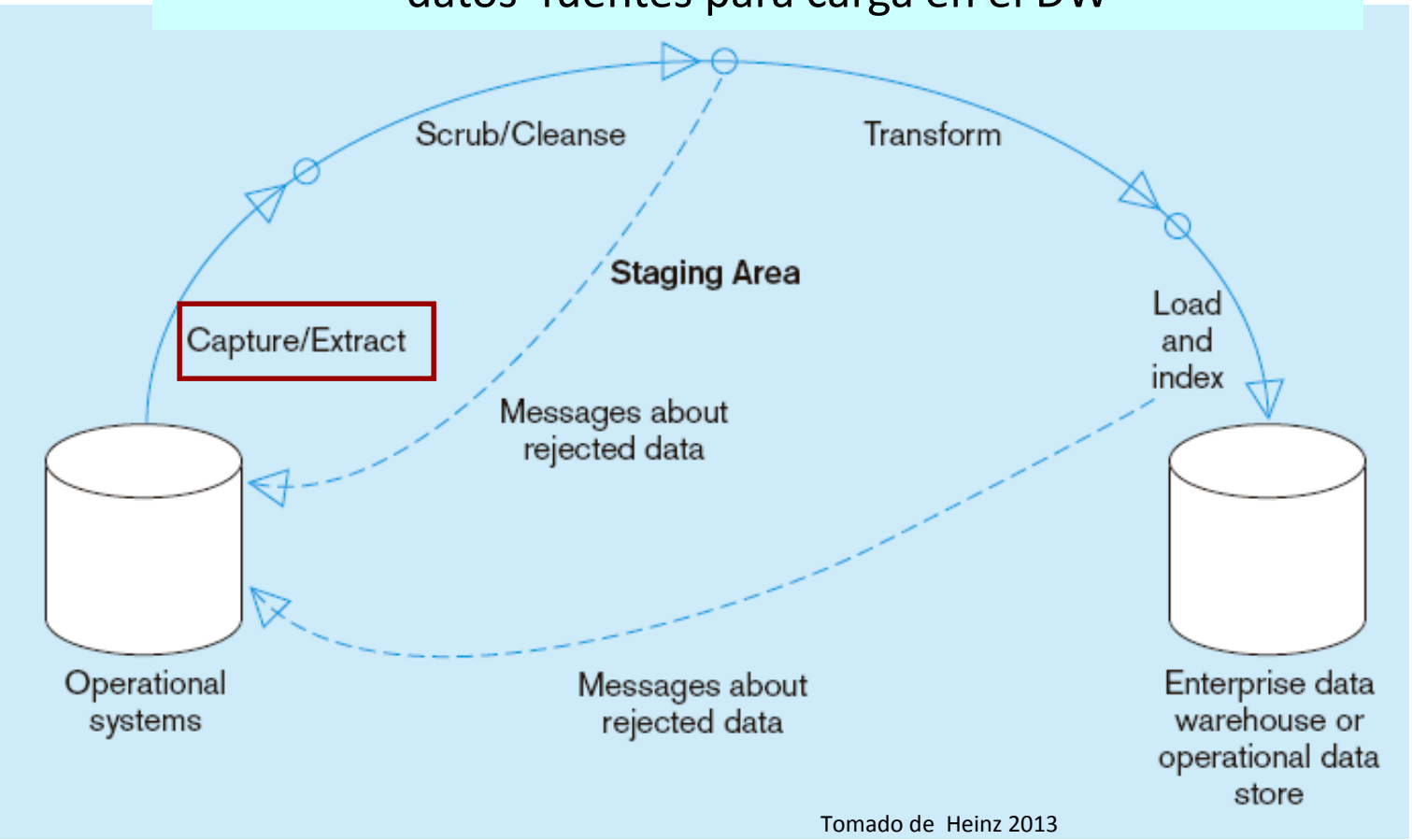


Extracción

Obtener datos de múltiples, heterogéneas fuentes externas

- Periodicidad
- Claves:
  - Manipular los datos sin interrumpir ni paralizar los OLTP, ni tampoco el DW.
  - No depender de la disponibilidad de los OLTP.
  - Almacenar y gestionar los metadatos que se generarán en los procesos ETL.
  - Facilitar la integración de las diversas fuentes, internas y externas.

Captura / Extrae... obtiene un subconjunto de los datos fuentes para carga en el DW



Tomado de Heinz 2013

Extracción estática: captura los datos en un momento puntual

Extracción Incremental: captura cambios que se van produciendo

ETL: Técnicas de Monitoreo

- Instantáneas cada cierto tiempo
- Disparadores de base de datos (reglas (triggers))
- Logs de registros
- Envío de datos
- Envío de transacciones
- Consultas a las BDs

Métodos de Extracción

- **Extracciones a granel**
  - Todo el DW se actualiza periódicamente
  - Pesado para las conexiones de red entre el origen y destino
  - Más fácil de configurar y mantener

- **Extracciones basadas en intercambio**
  - Sólo los datos que han sido recién insertadas o actualizadas en los sistemas de origen se cargan en el DW
  - Menos carga en la red pero requiere una programación más compleja para determinar cuando un nuevo registro de DW debe insertarse o cuando un registro DW tiene que actualizarse

Limpieza de datos

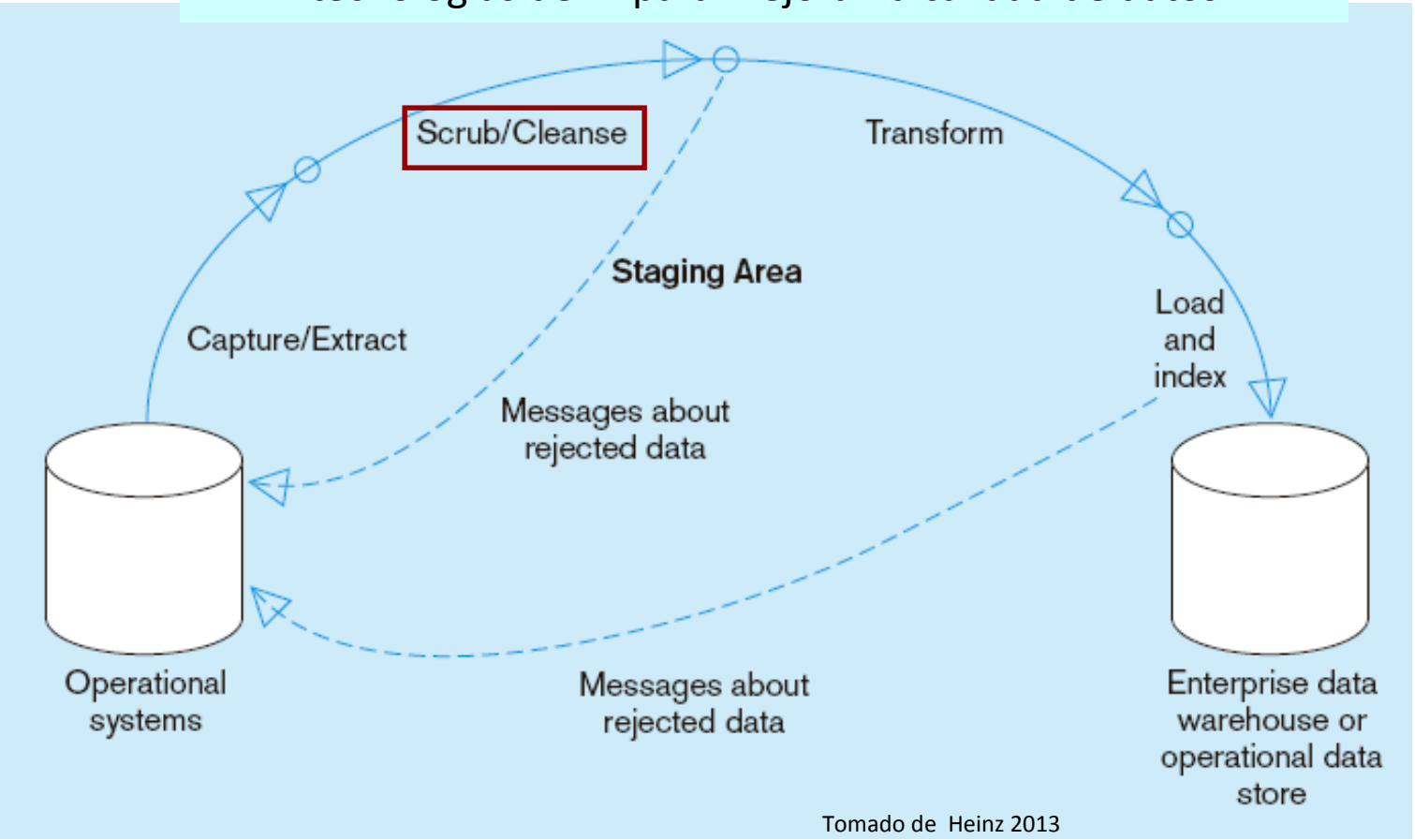
Sistemas de origen tiene "datos sucios" que deben ser limpiados

Razones de datos “sucios"

- valores ficticios
- La falta de datos
- datos encriptados
- datos contradiciéndose
- El uso inapropiado de datos
- Violación de Reglas de Negocio
- Claves principales reutilizados
- Identificadores no únicos

detectar errores en los datos y rectificarlos

Limpieza... utiliza reconocimiento de patrones y tecnologías de IA para mejorar la calidad de datos



Tomado de Heinz 2013

Solución de errores: faltas de ortografía, fechas erróneas, uso de campo incorrecto direcciones no coinciden, datos faltantes, datos duplicados, inconsistencias

También decodifica, reformatea, convierte, genera claves, fusiona, detecta errores registro, localiza datos faltantes

Pasos en la Limpieza de datos

1. Análisis (parsing)
2. Corrección
3. Estandarización
4. Mapeo (matching)
5. Consolidación

Análisis (Parsing)

Localiza e identifica **elementos individuales** en los archivos de origen y luego los aísla.

- **Ejemplos**
  - Análisis del primer nombre, segundo nombre y apellido;
  - Analizar número y nombre de la calle;
  - Analizar la ciudad y el estado.

Corregir

Corrige los componentes de **datos individuales** utilizando algoritmos sofisticados y fuentes de datos secundarias.

Acciones típicas con Datos Anómalos (Outliers):

- Ignorarlos.
- Eliminar la columna.
- Filtrar la columna.
- Filtrar la fila errónea, ya que a veces su origen, se debe a casos especiales.
- Reemplazar el valor.

Acciones contra Datos Faltantes (Missing Values):

- Ignorarlos.
- Eliminar la columna.
- Filtrar la columna.
- Filtrar la fila errónea, ya que a veces su origen, se debe a casos especiales.
- Reemplazar el valor.
- Esperar hasta que los datos faltantes estén disponibles.

Estandarización

Aplica **rutinas de conversión** para transformar los datos en formatos preferidos (y coherentes), utilizando reglas de estándarización.

- **Ejemplos:** la sustitución de un apodo, uso de un nombre de calle preferido, etc.

Mapeo (matching)

búsqueda de **coincidentes en registros** en los datos analizados, corregidos y estandarizados, basados en reglas predefinidas para eliminar la duplicación.

- **Ejemplos:** identificación de nombres y direcciones similares.

Transformación de datos

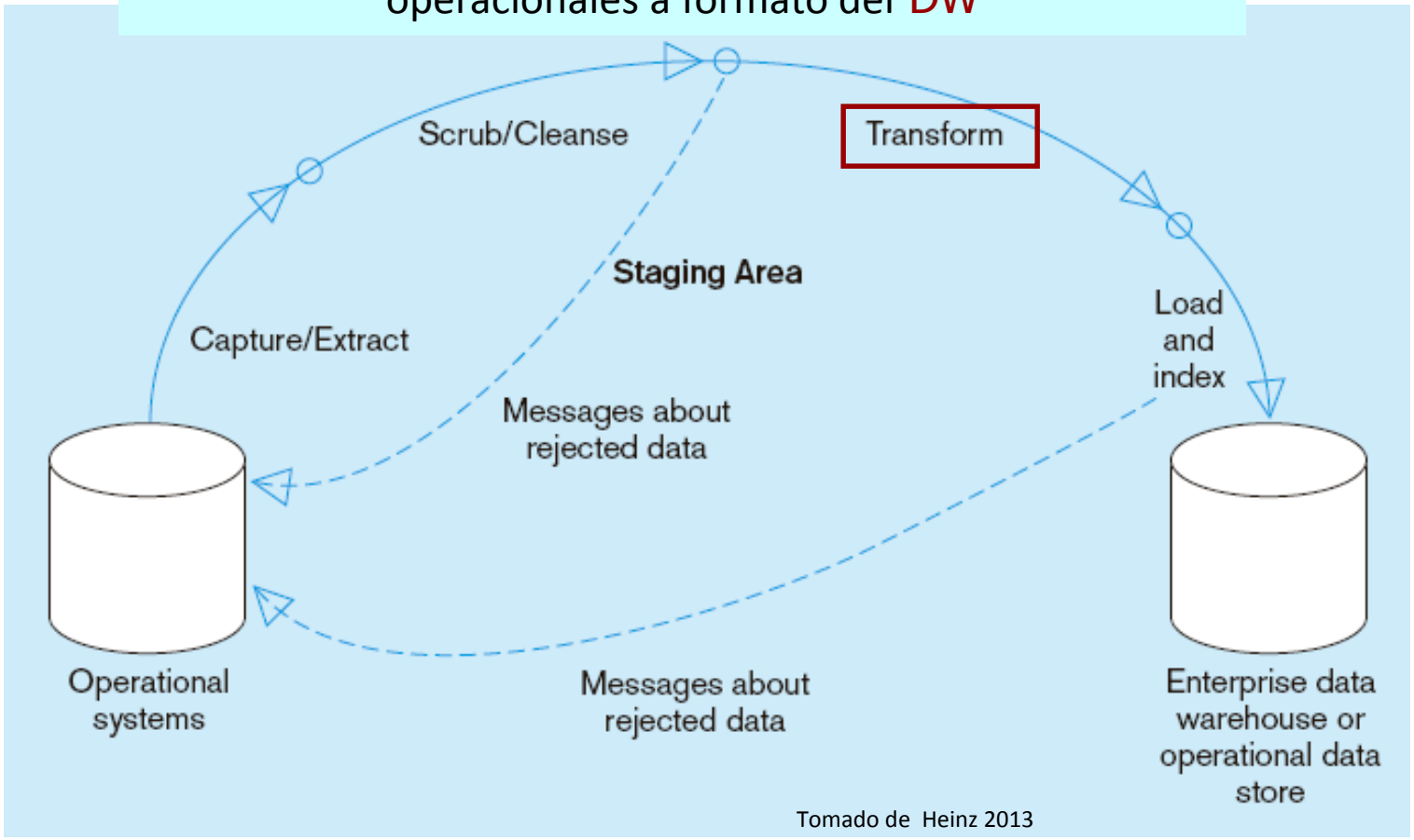
Transforma los datos de acuerdo con reglas y normas establecidas

- Clásicos problemas:
  - Codificación.
  - Medida de atributos.
  - Convenciones de nombramiento.
  - Fuentes múltiples,

ordena **calcula**

resume **consolida**

Transforma… convierten los datos desde las BD operacionales a formato del DW



Tomado de Heinz 2013

A nivel de registro:

- Partición de Datos (selección)
- Juntar datos (combinación)
- Resumir datos (agregación)

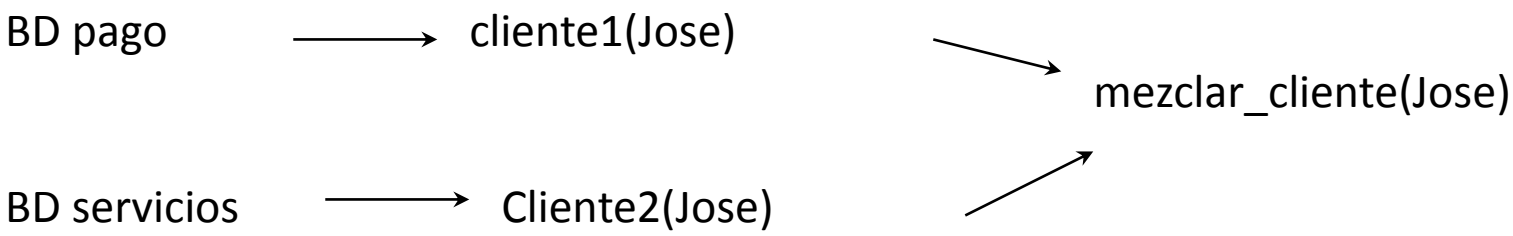
A nivel de campo:

- de un solo campo: de un campo a un campo
- multi-campo: de muchos campos a uno, o uno a muchos campos

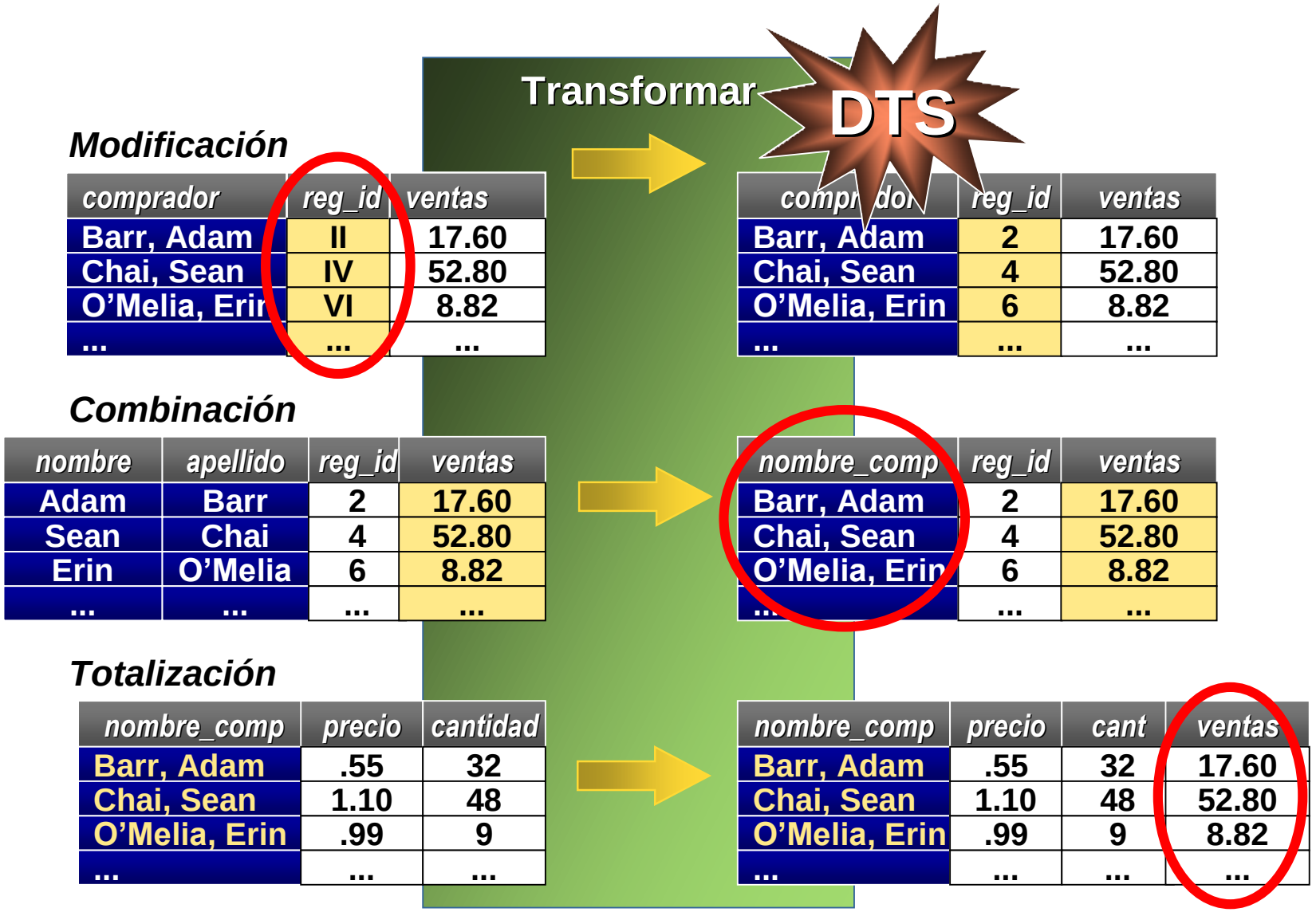
ETL: Transformación de datos

Posibles tareas

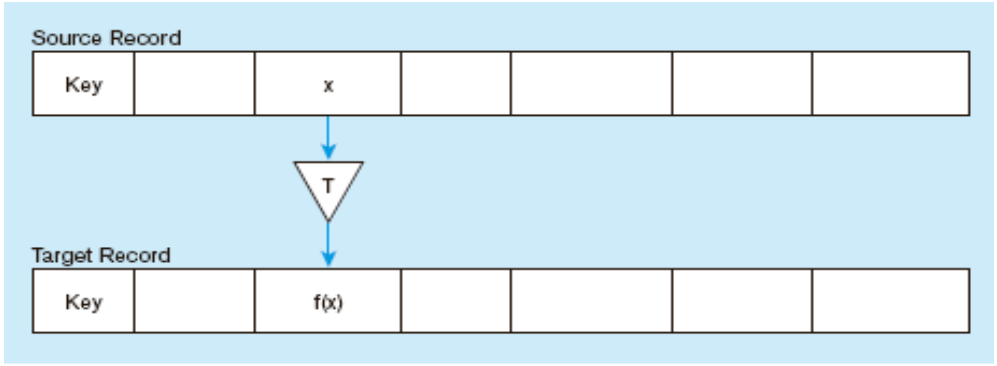
- **Migrar** (por ejemplo, yen a dólares)
- **Refinar**: utilizar el conocimiento específico de dominio (por ejemplo, números de seguro social)
- **Fusionar** (por ejemplo, lista de correo con la de clientes)



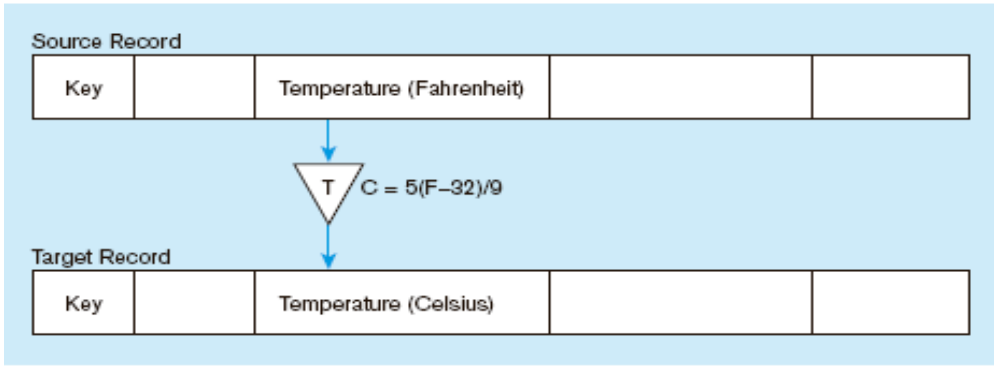
Proceso ETL: Tranformar



Transformación de un solo campo



Una función de transformación traduce los datos de una forma antigua a una nueva



Transformación algorítmica: utiliza una fórmula o expresión lógica

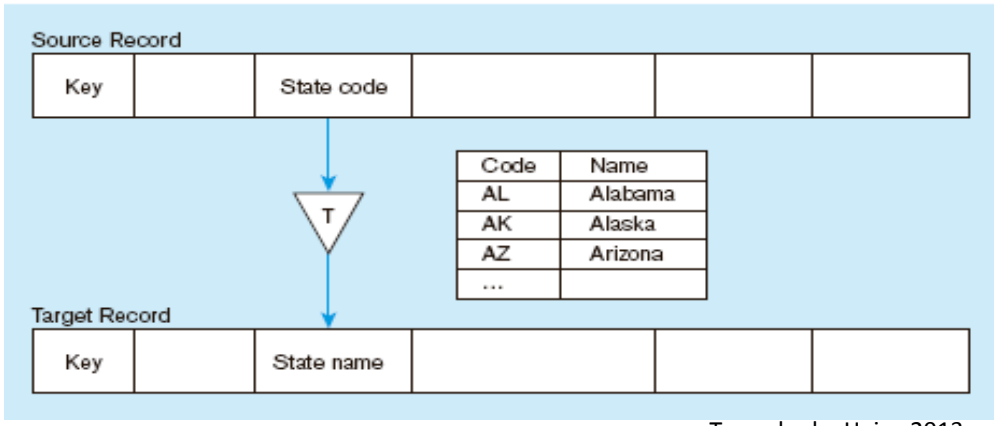
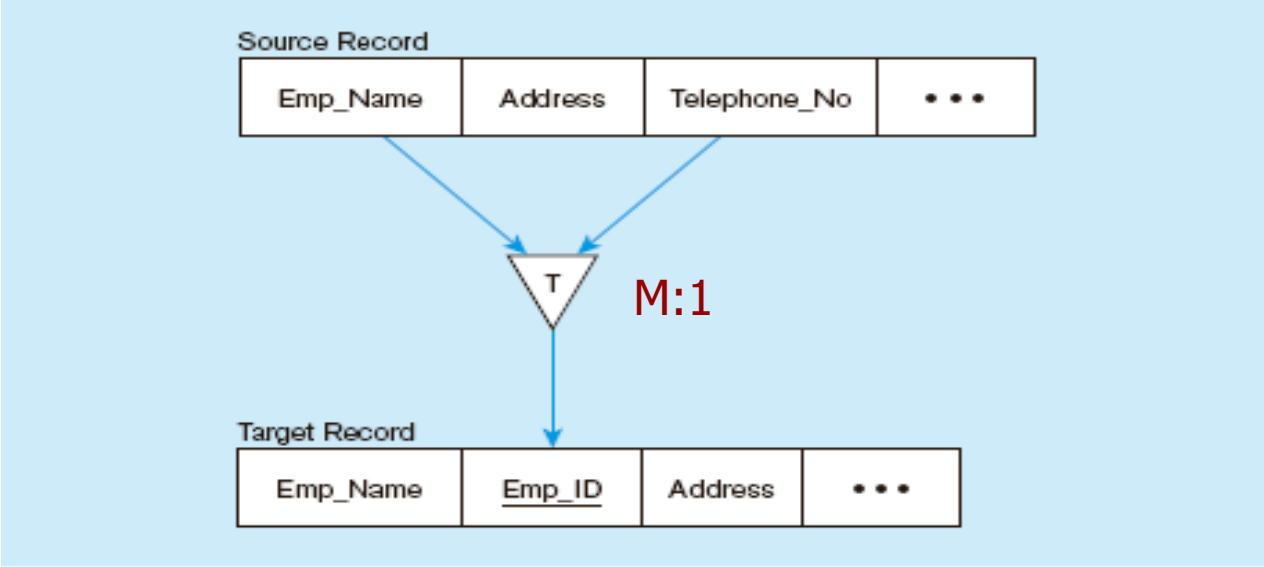
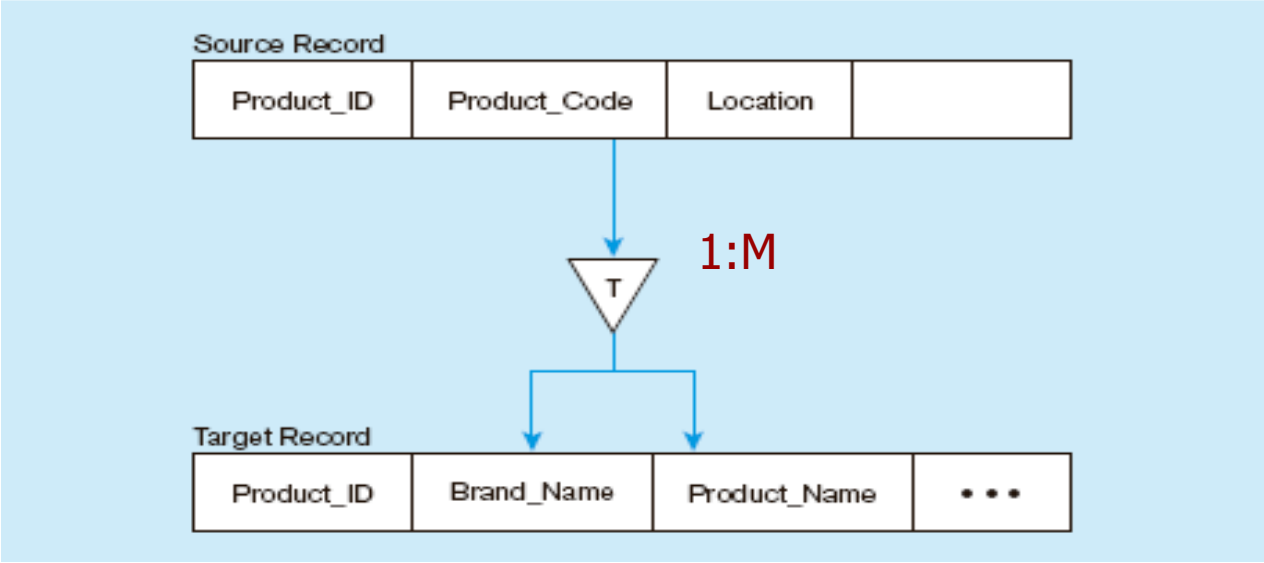


Table lookup: utiliza una tabla separada basada en códigos

Tomado de Heinz 2013

Transformación multicampos



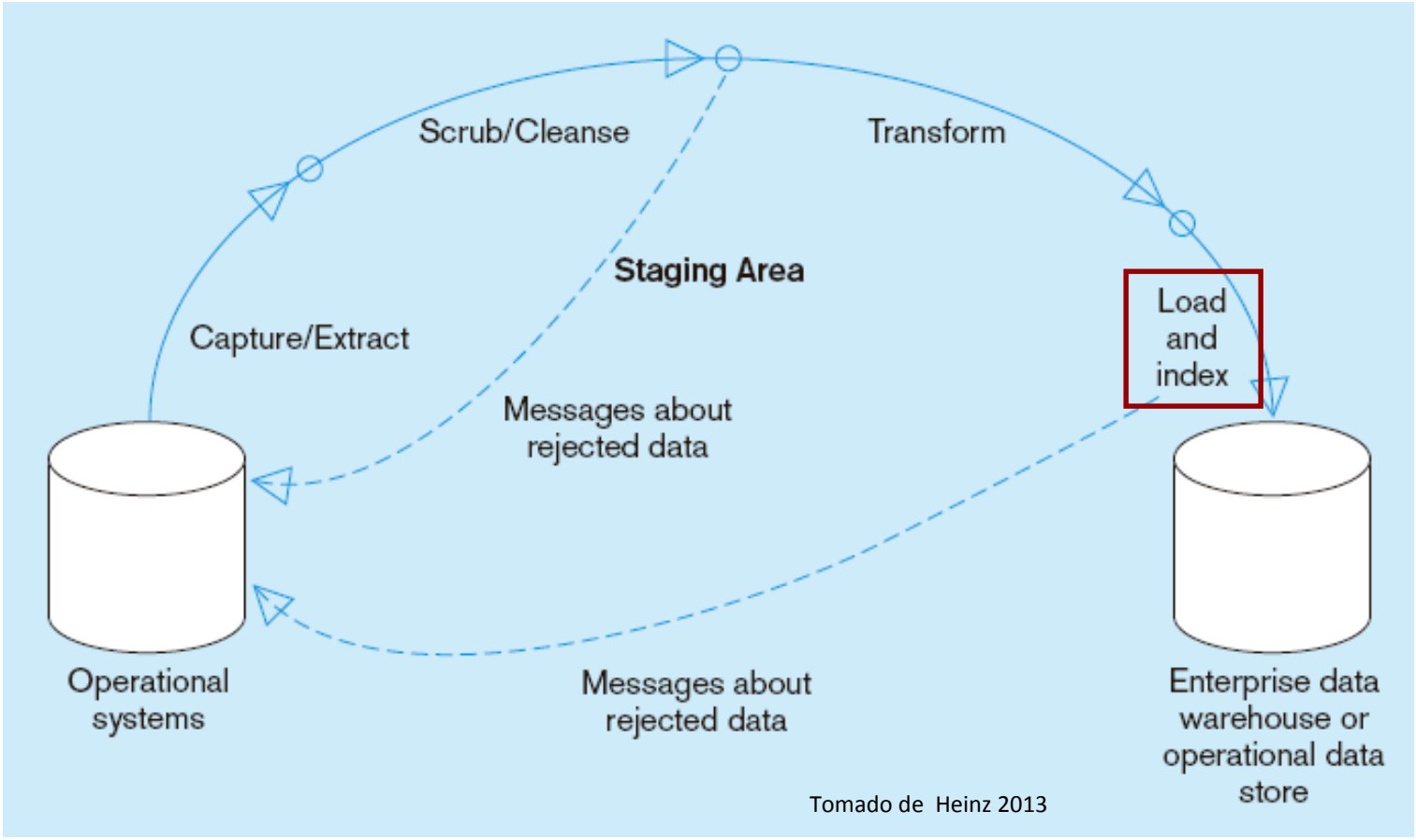


Carga de Datos

Los datos físicamente se almacena en el almacén de datos

- La carga ocurre en una "ventana de carga“
- La tendencia cada vez mayor es actualizaciones en tiempo real

Cargar/Indizar… Transforma datos y crea índices



Tomado de Heinz 2013

Modo de Actualización 1:
reescritura masiva de datos en destino a intervalos periódicos

Modo de actualización 2:
solamente cambios en los datos de origen se escriben en el warehouse

UNIVERSIDAD
DE LOS ANDES
MERIDA VENEZUELA

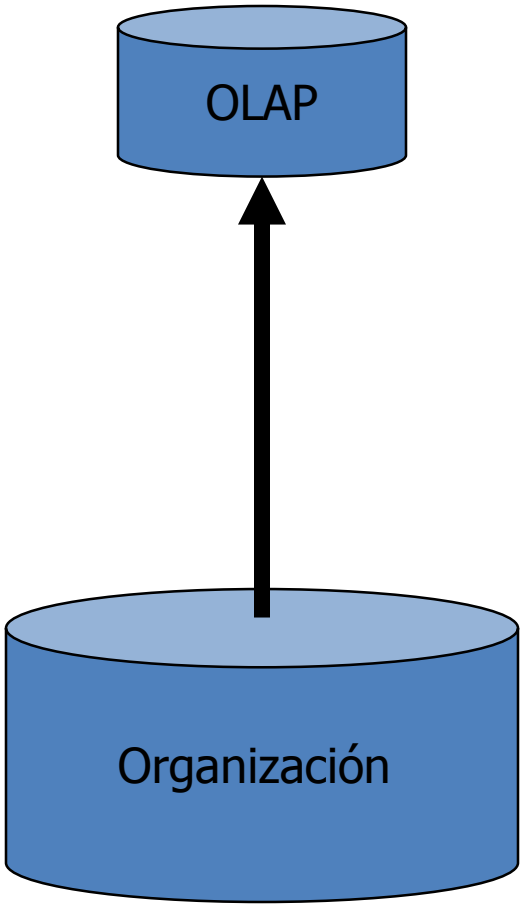
UTPL
UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA

Modelado de Datos

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

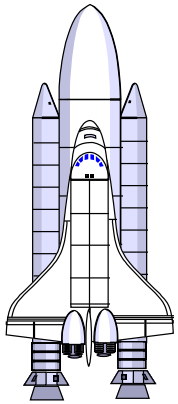
Modelado de Datos y de Información

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.

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 l**a 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

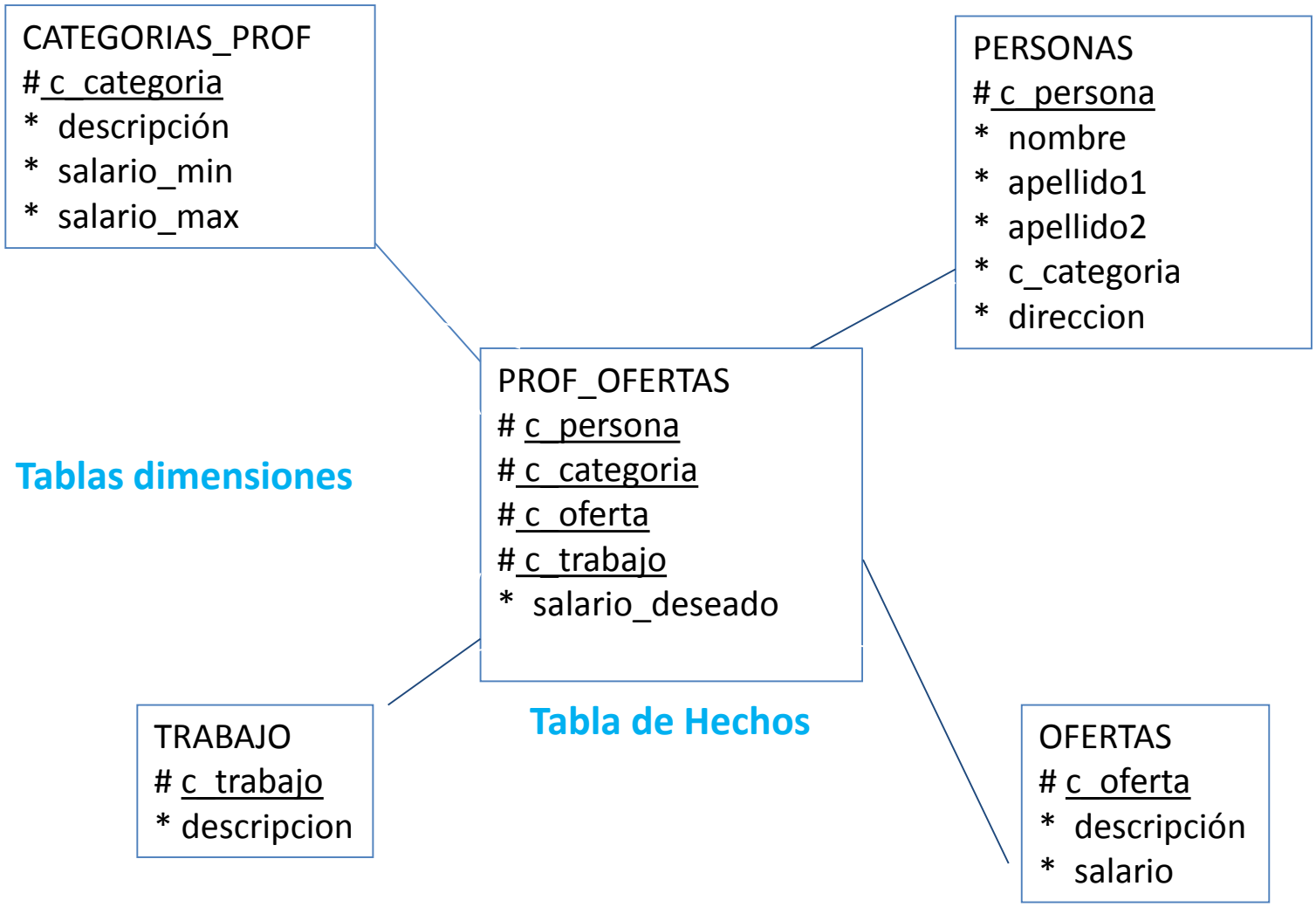
- En estrella
- Constelación
- Copo de nieve

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: Componentes

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

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 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

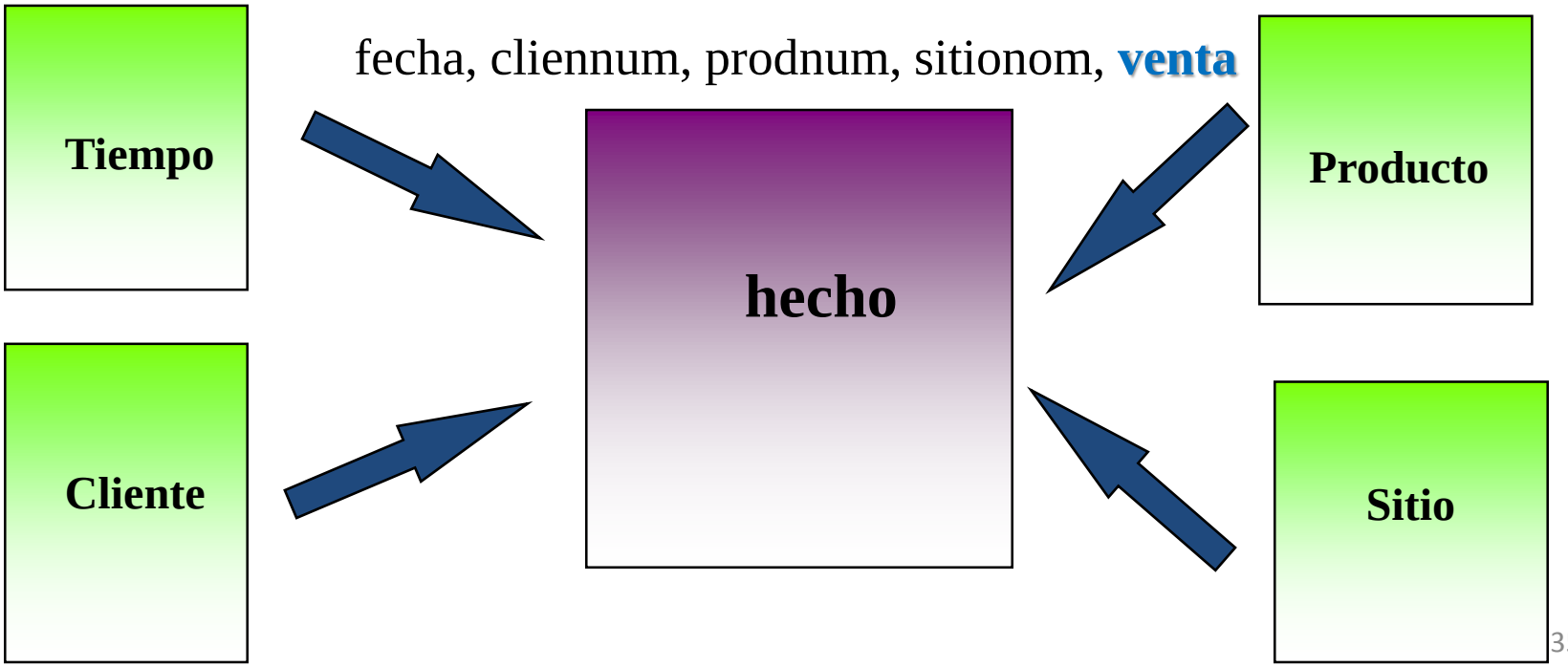
- El modelo estrella consolida hechos en relación a dimensiones o filtros.

- 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

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



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

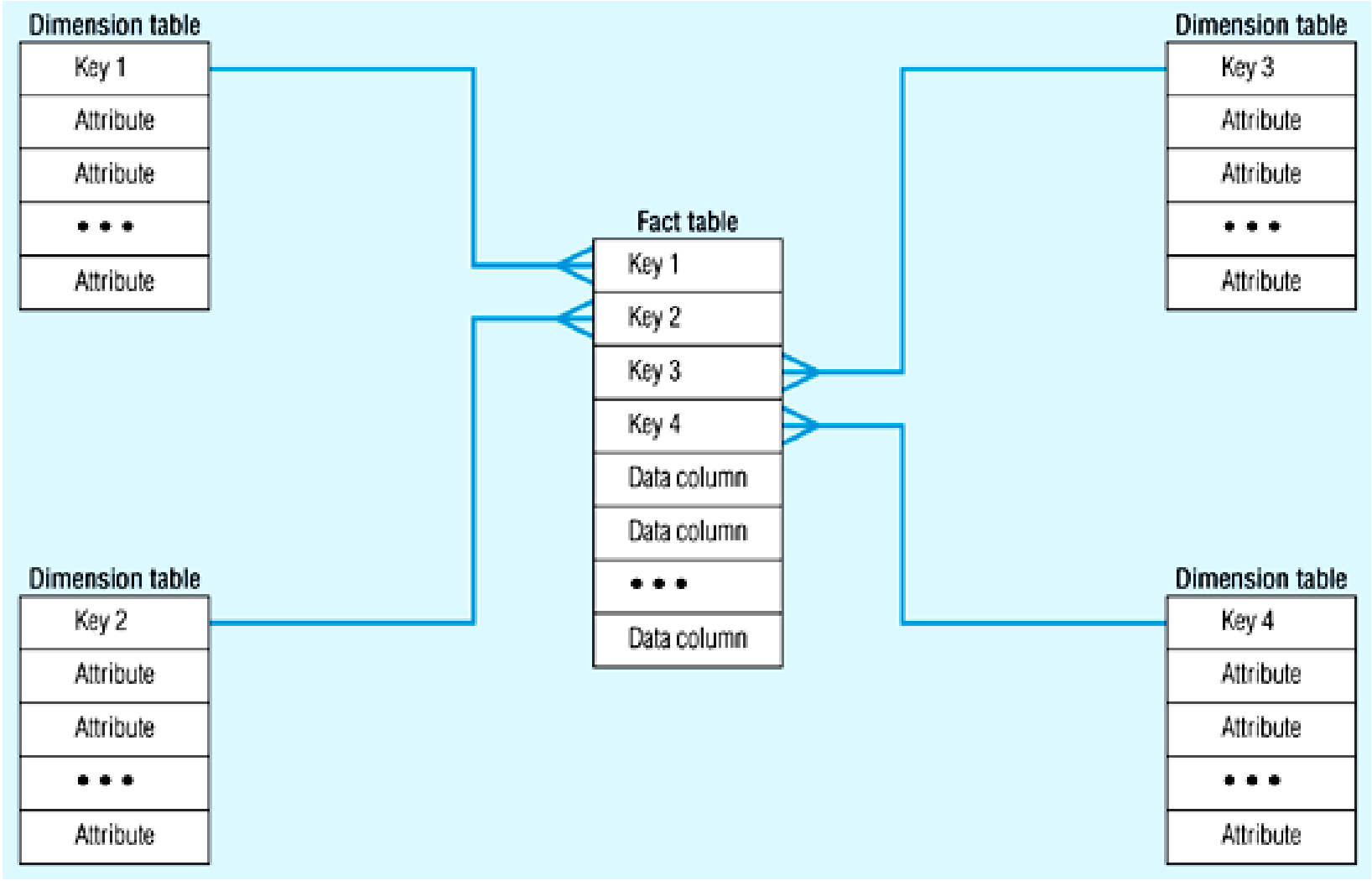
Esquema en estrella: Tabla de Dimensiones

- 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, inclusos 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

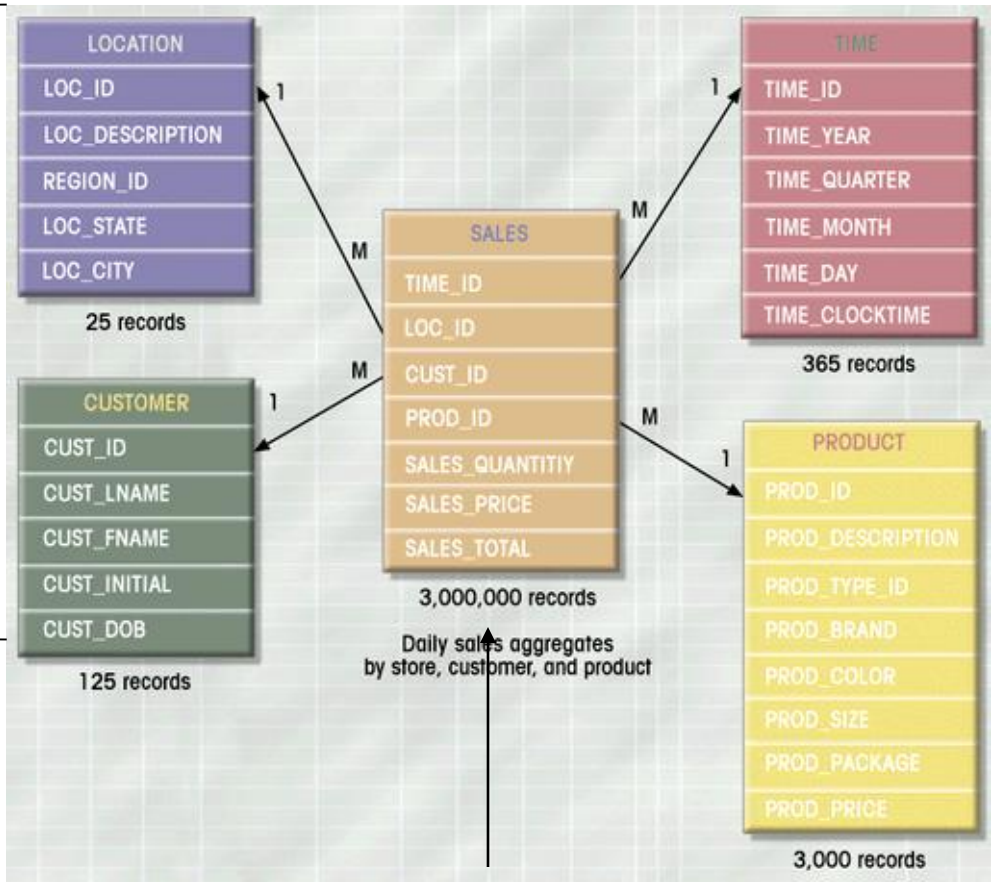
- Tablas de dimensiones contienen atributos
- Los atributos se utilizan para buscar, filtrar o clasificar los hechos
- Dimensiones proporcionan características descriptivas acerca de los hechos a través de sus atributos
- Debe definir los atributos comunes que se utilizará para reducir la búsqueda, agrupar 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

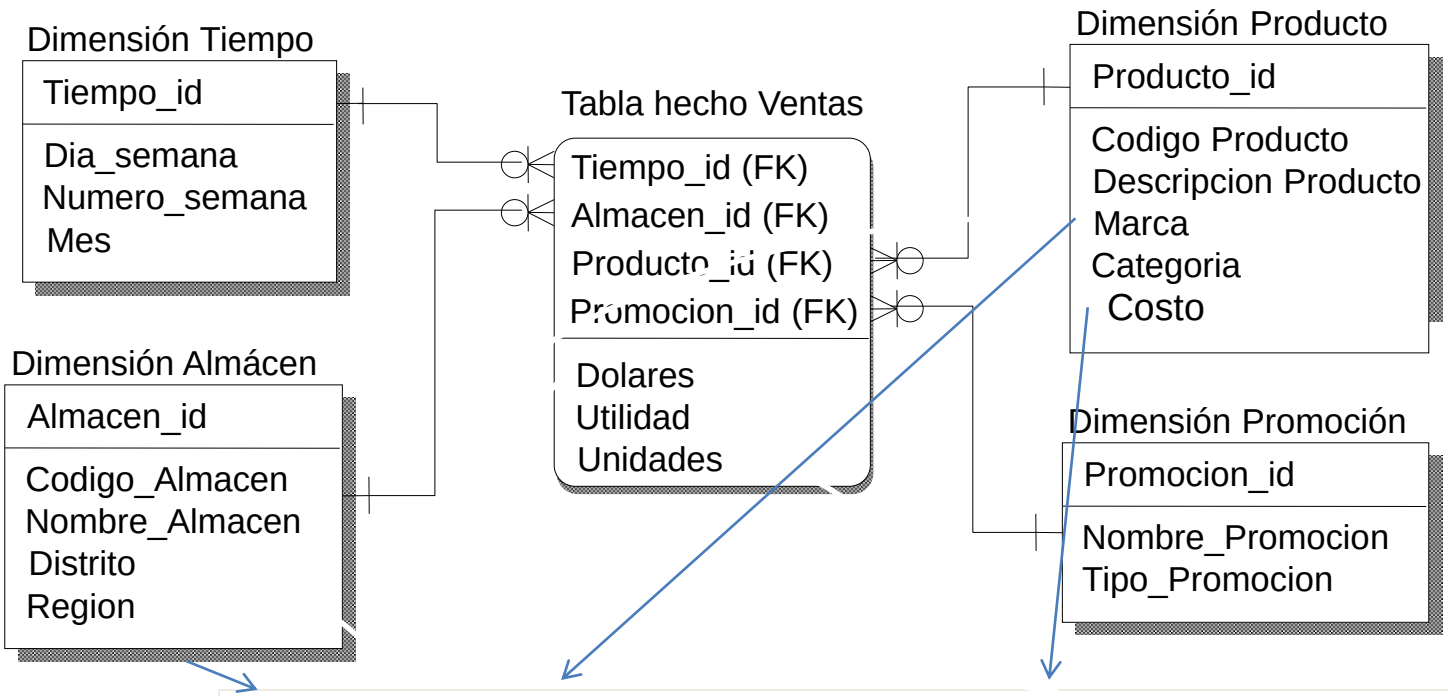
Tablas de Dimensiones



Relación M-1

Tabla de hechos

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(s) pregunta(s) específica(s) de IN**

Esquema Copo de nieve

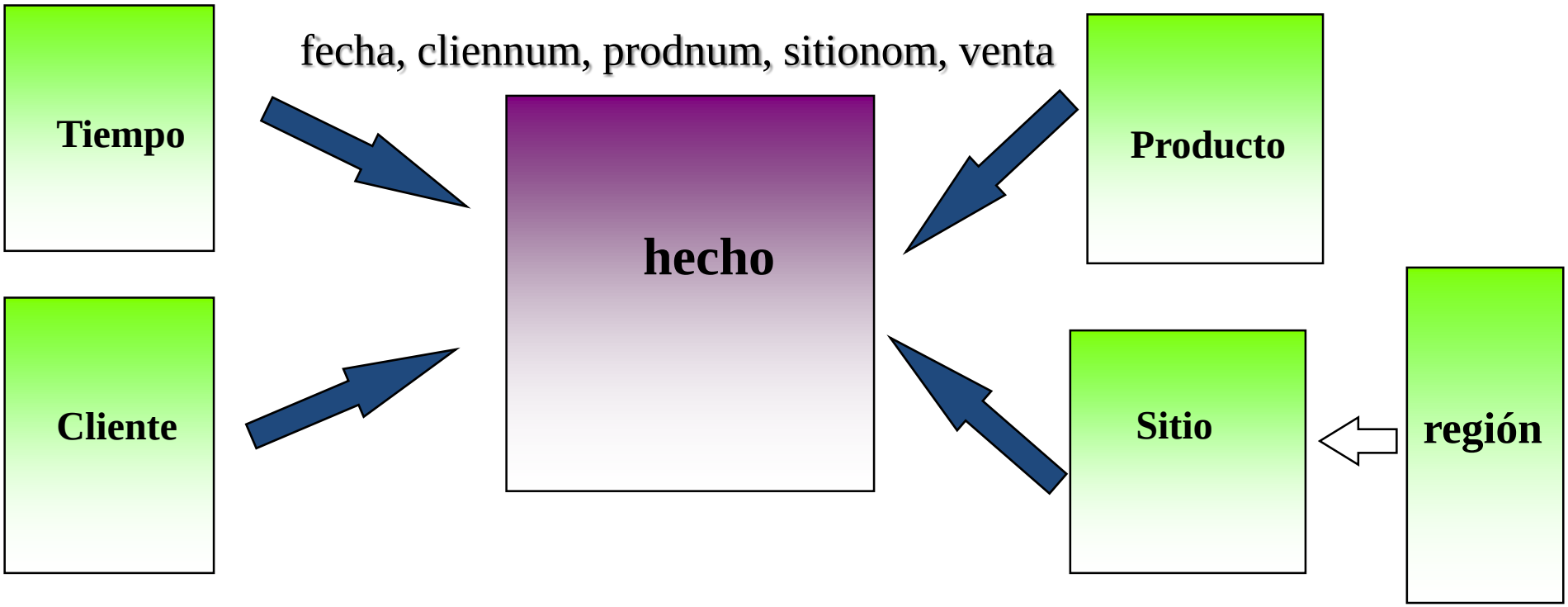
Un refinamiento del esquema en estrella donde alguna jerarquía dimensional se despliega en un conjunto de tablas de dimensiones más pequeñas

- 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

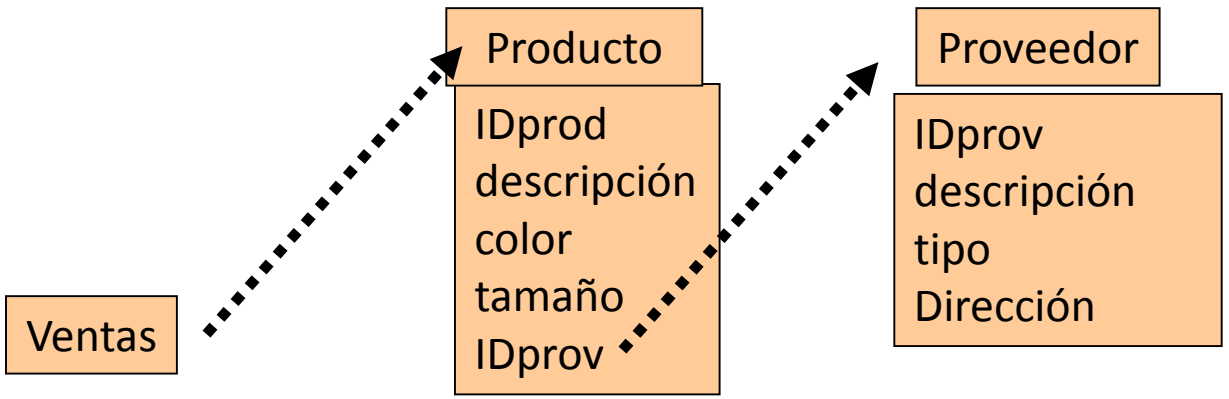
Representa jerarquía dimensional



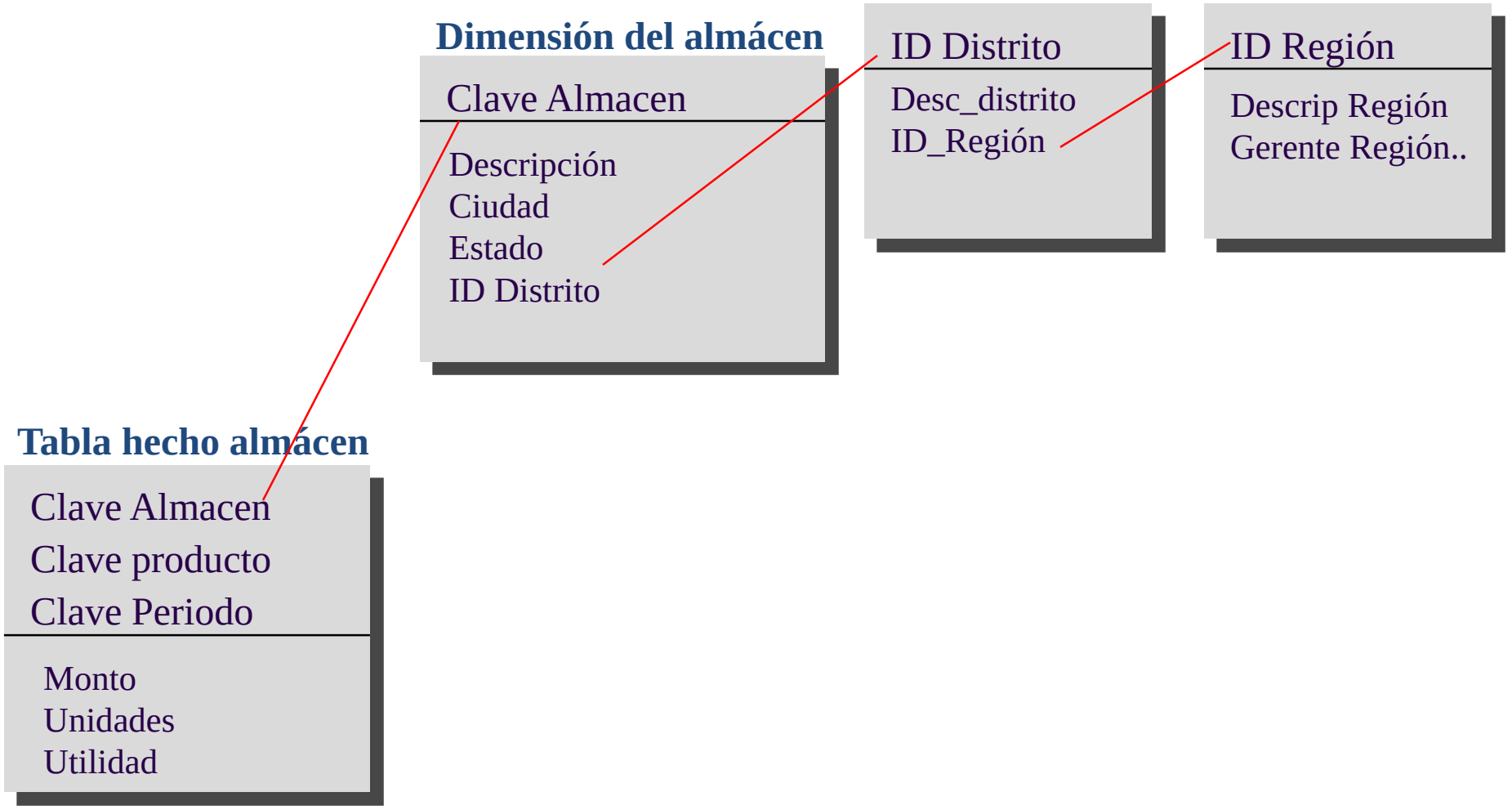
Esquema Copo de nieve

Beneficios

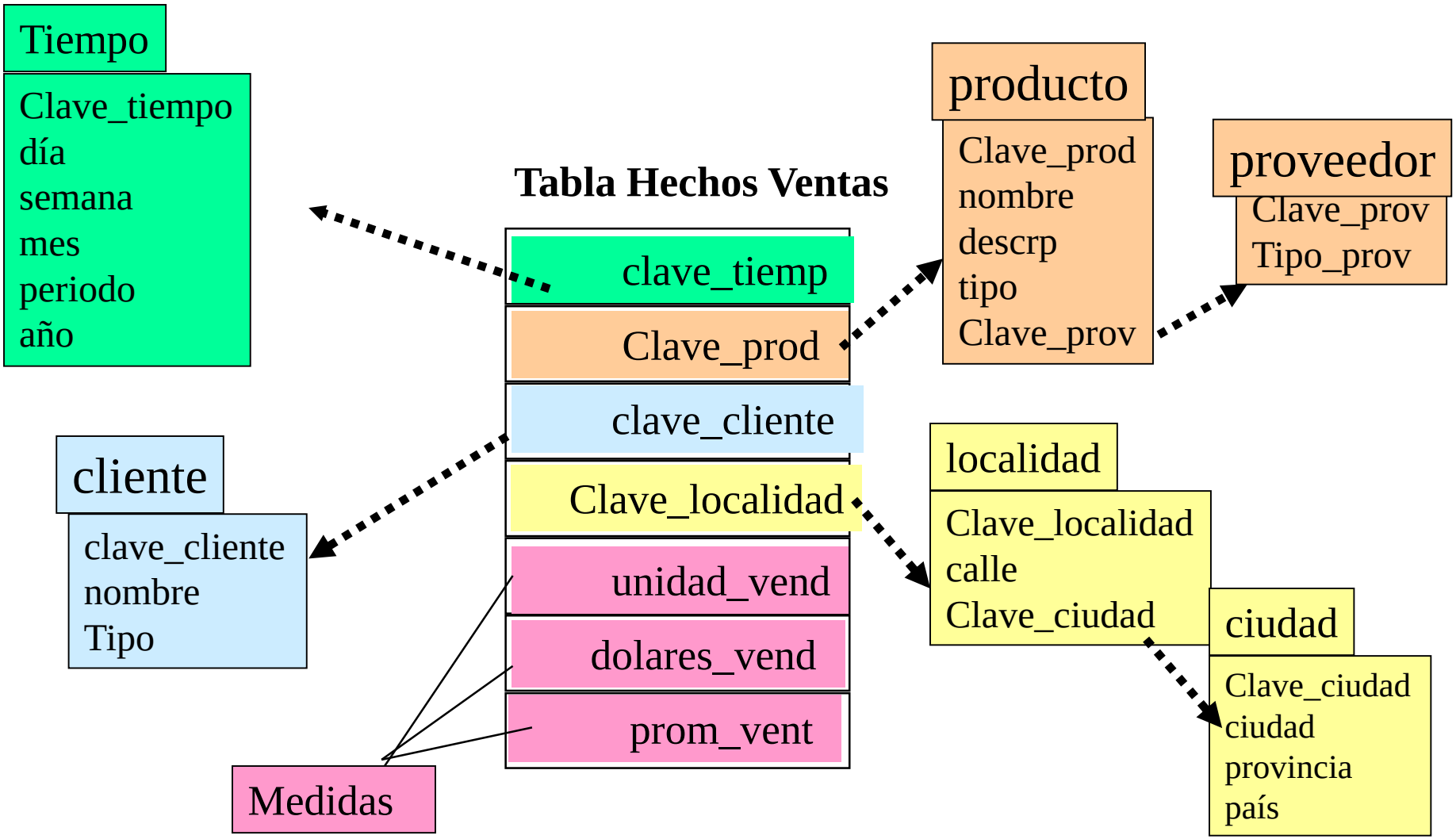
- Evita la duplicación
- Conduce a las constelaciones (varias tablas de hechos con dimensiones compartidas)



Esquema Copo de nieve



Esquema Copo de nieve



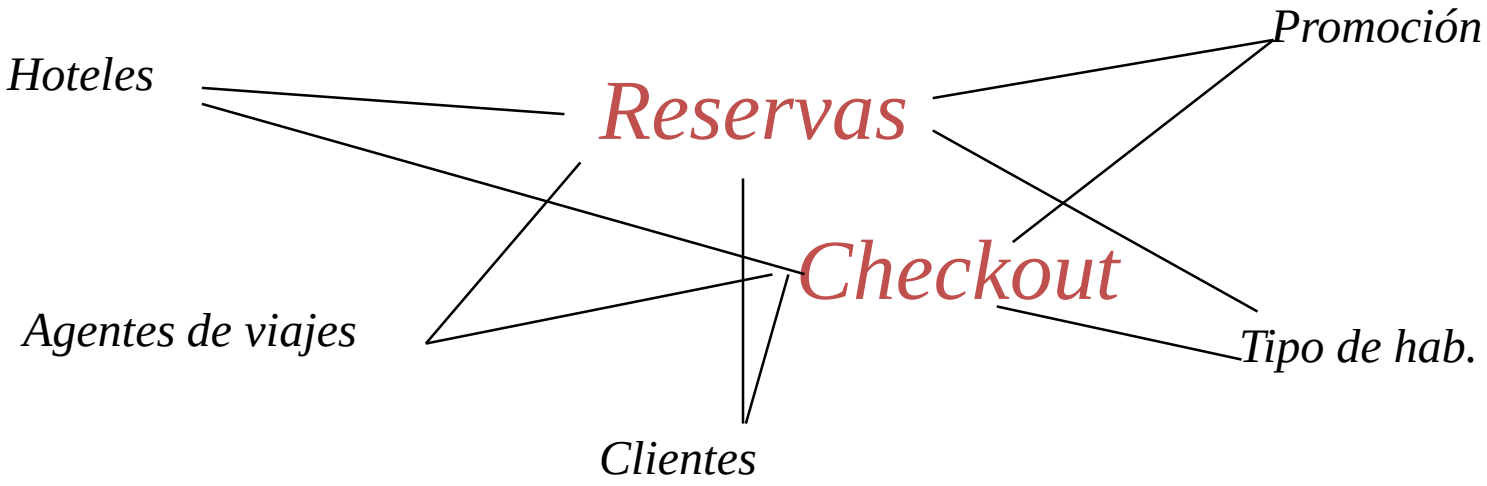
Esquema de Constelación de hechos

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

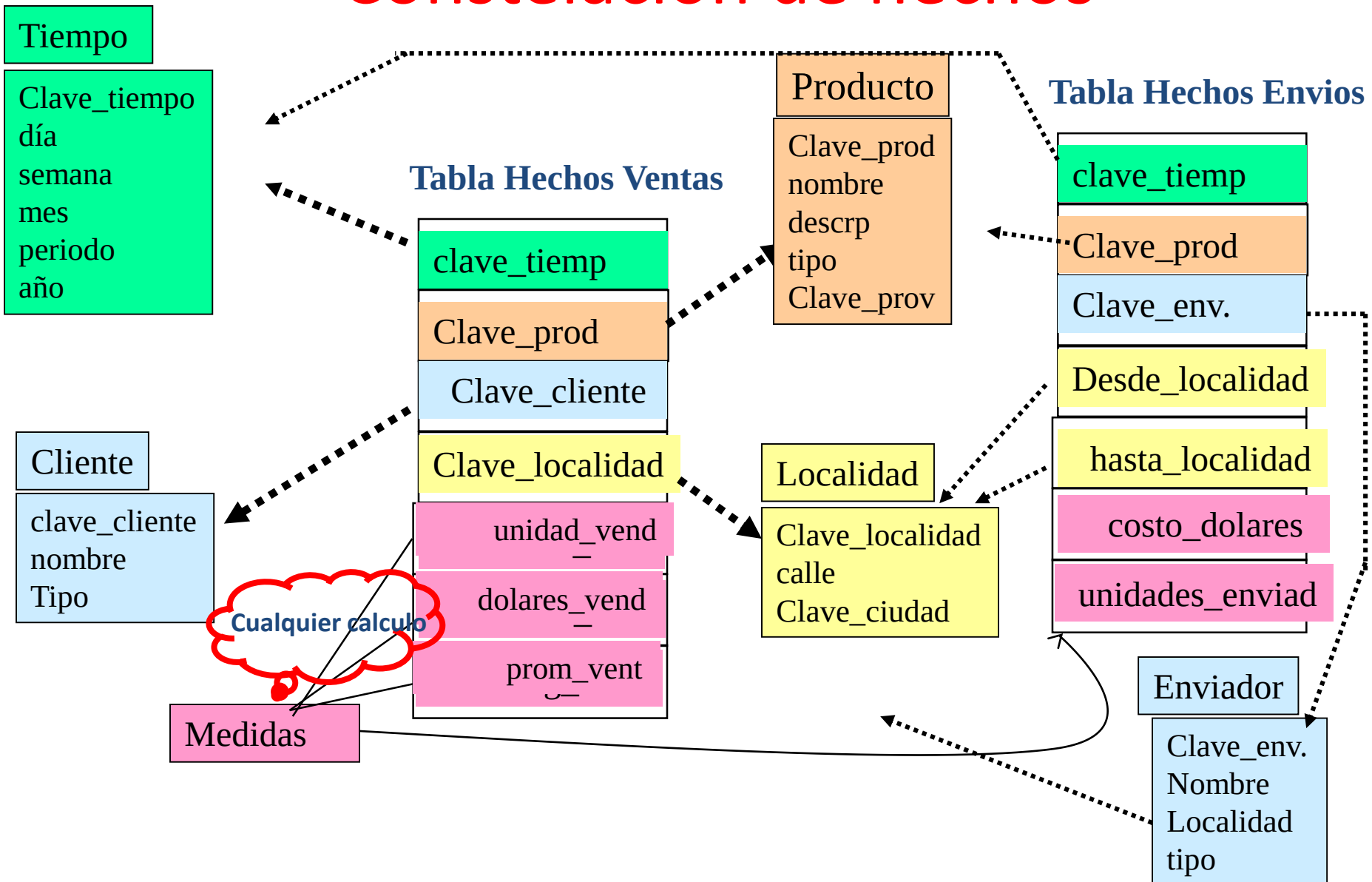
- vistos como una **colección de estrellas**, por lo tanto, llamados **esquema de galaxia o constelación de hecho**

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

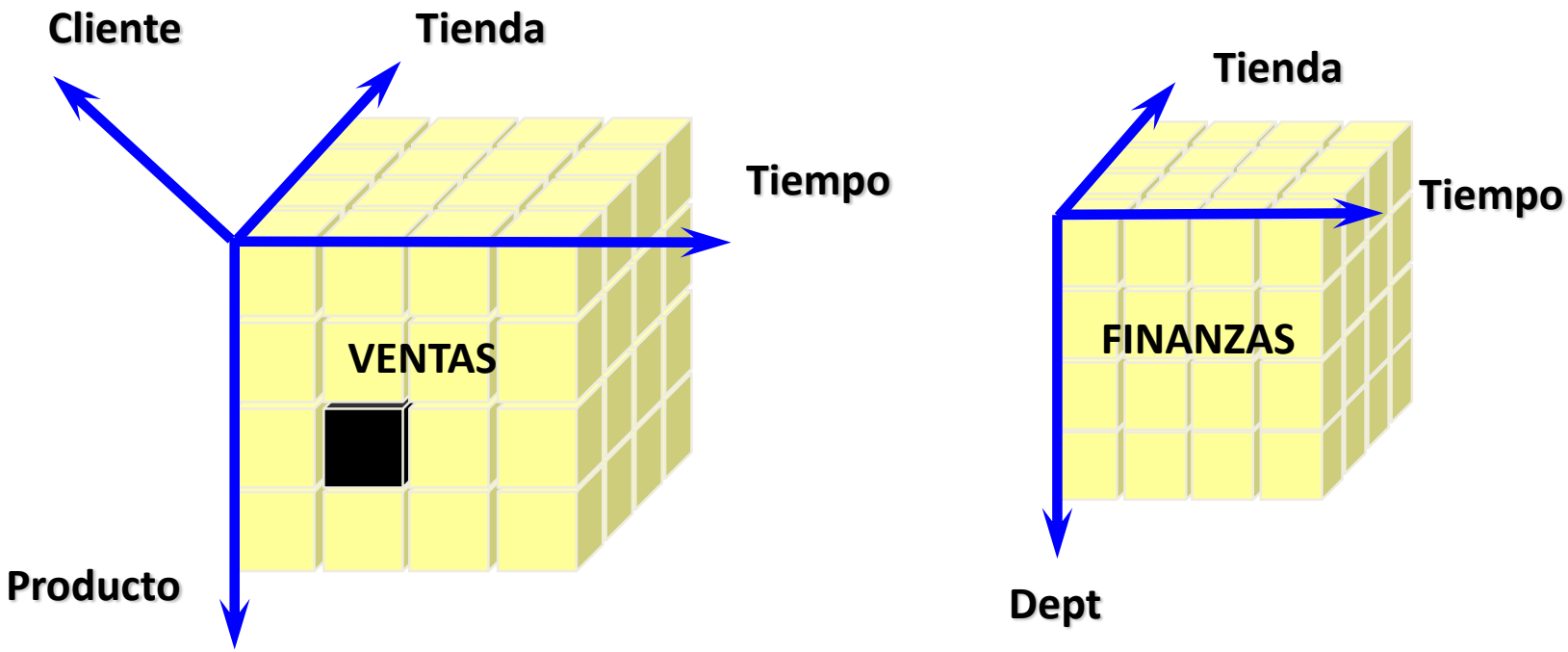
Un cubo n-D se llama un paralelepípedo.

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

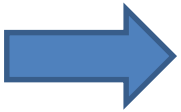
Modelo BD multidimensional



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
.....	.....	.....
.....	.....	.....



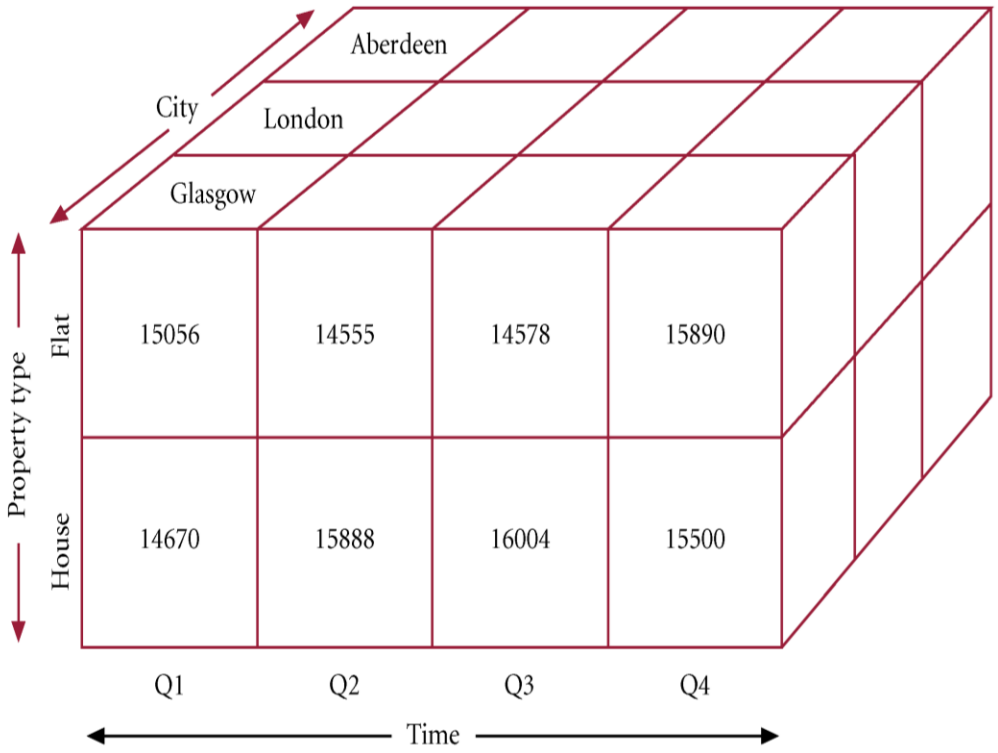
City

Time

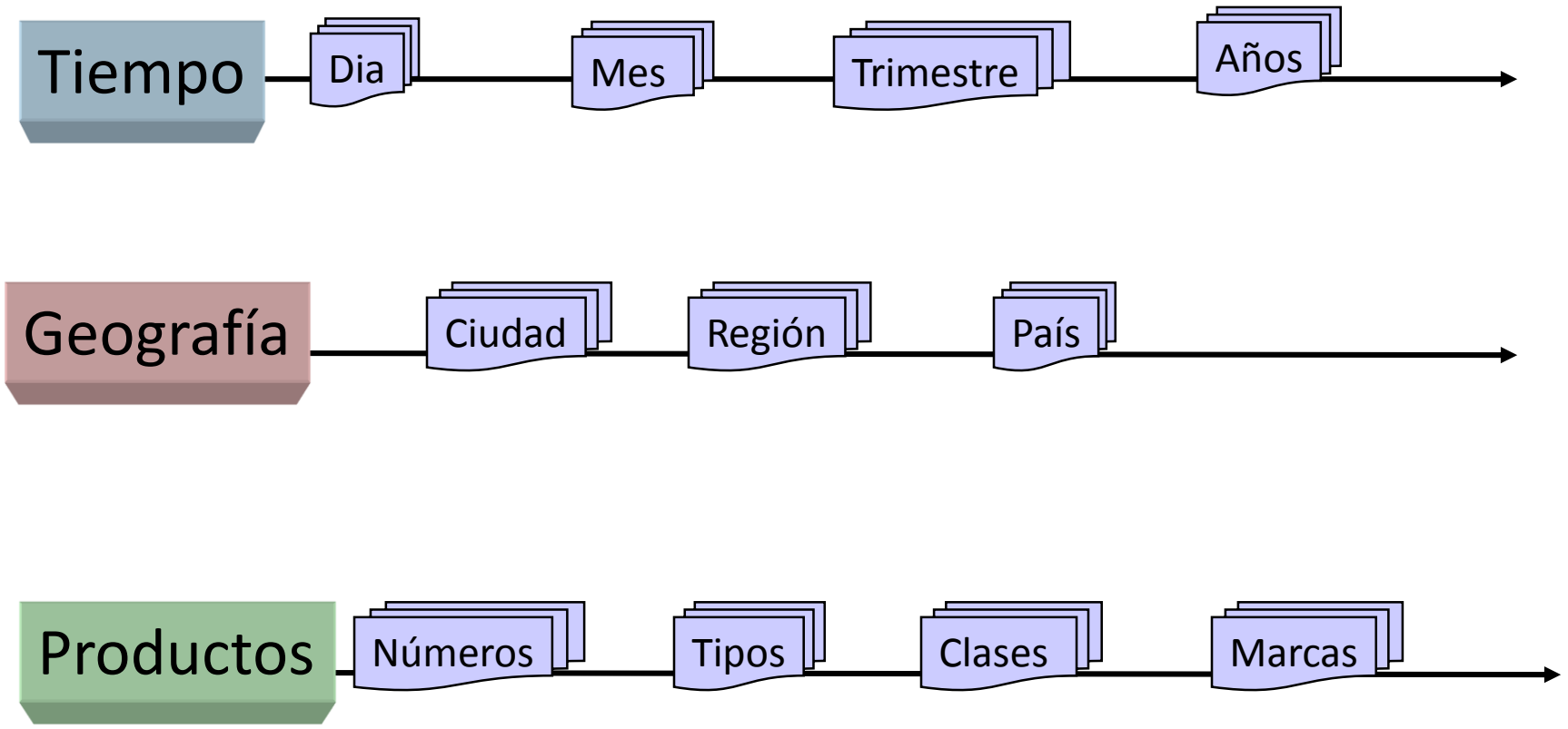
Quarter \ City	Glasgow	London	Aberdeen	.....
Q1	29726	43555	53210	.....
Q2	30443	48244	34567	.....
Q3	30582	56222	45677	.....
Q4	31390	45632	50056	.....

Tres dimensiones

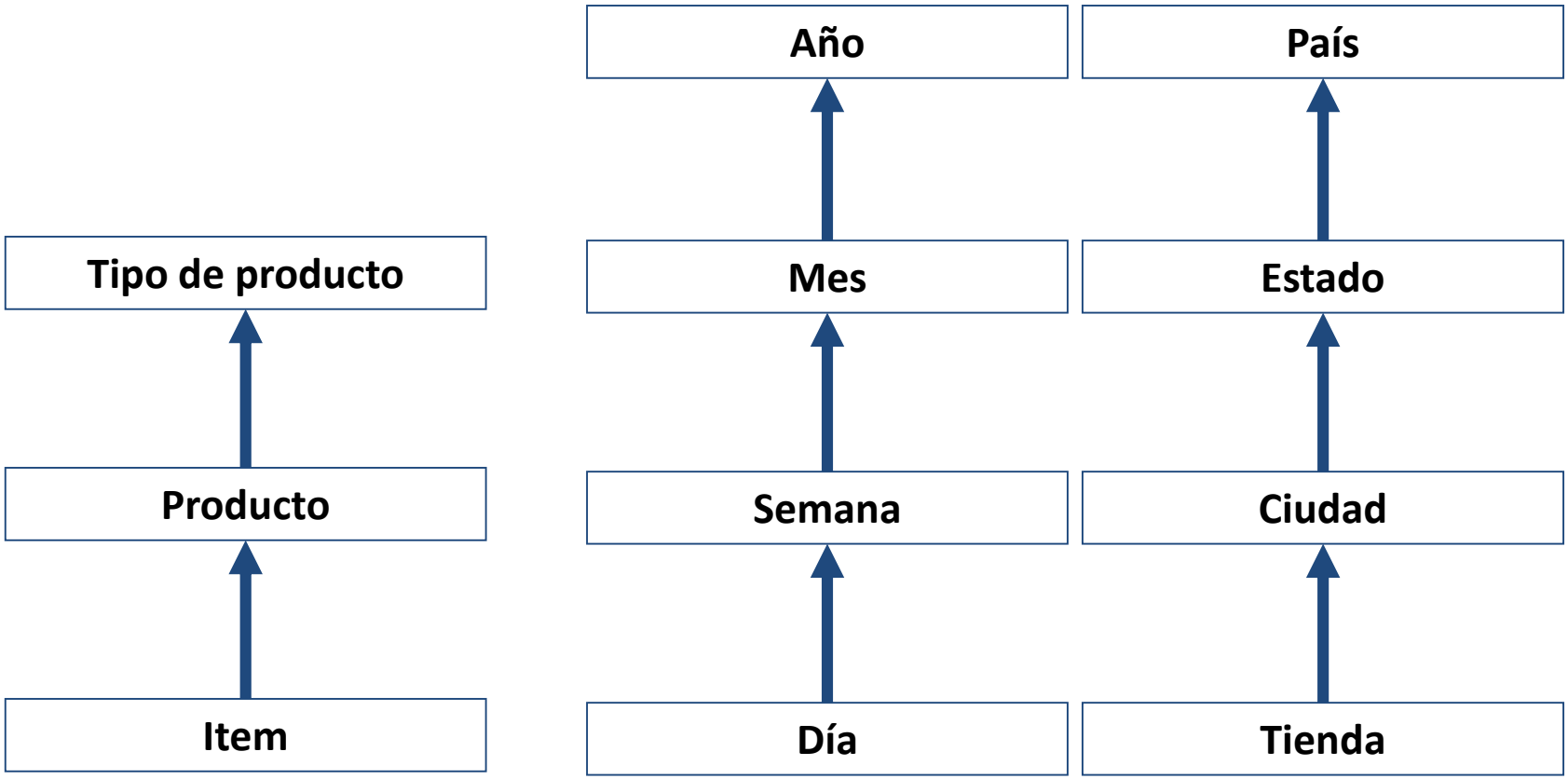
Property Type	City	Time	Total Revenue
Flat	Glasgow	Q1	15056
House	Glasgow	Q1	14670
Flat	Glasgow	Q2	14555
House	Glasgow	Q2	15888
Flat	Glasgow	Q3	14578
House	Glasgow	Q3	16004
Flat	Glasgow	Q4	15890
House	Glasgow	Q4	15500
Flat	London	Q1	19678
House	London	Q1	23877
Flat	London	Q2	19567
House	London	Q2	28677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



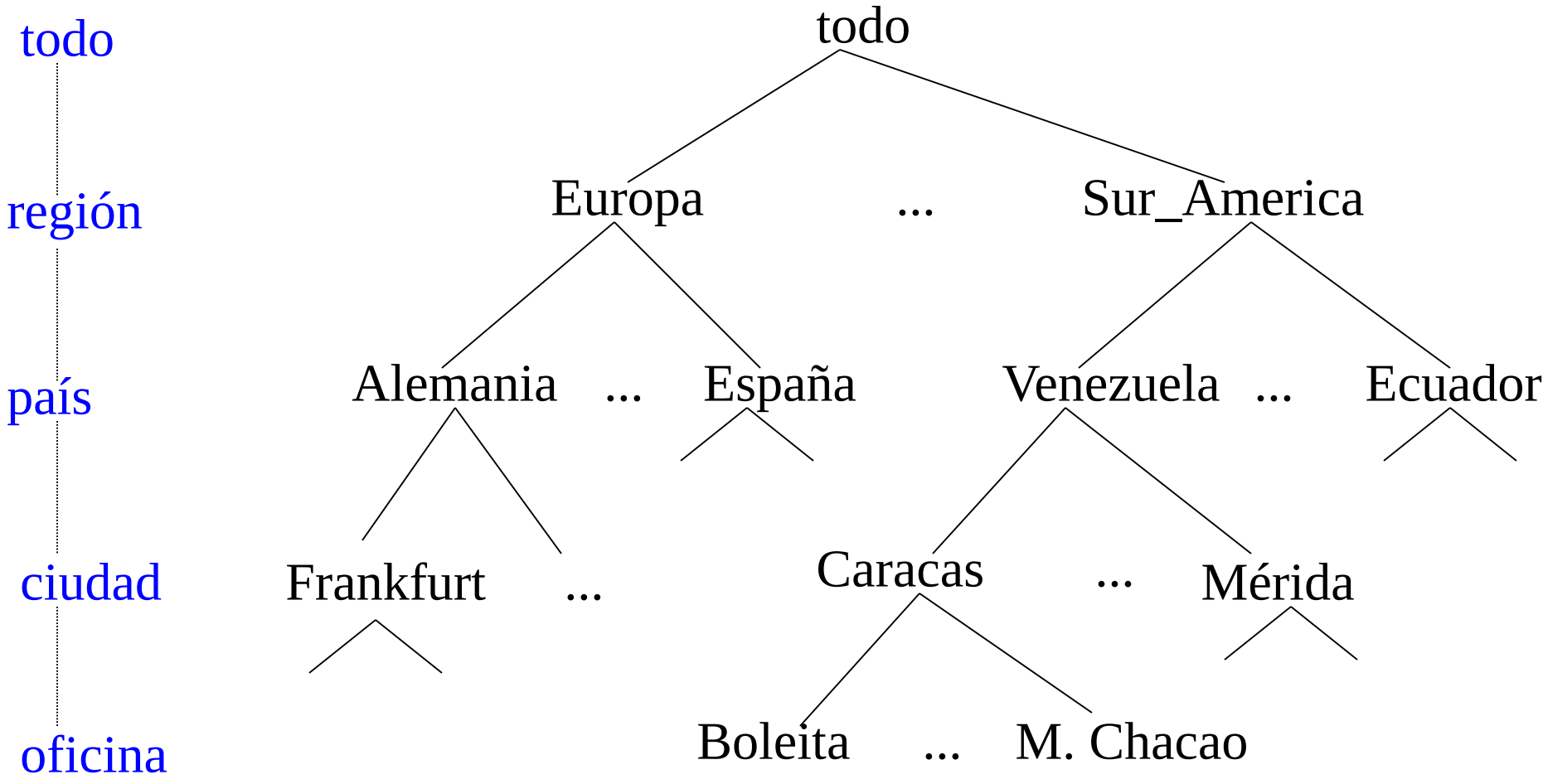
La granularidad de las dimensiones



Jerarquía Dimensional



Jerarquía Dimensional (localidad)



Jerarquía Dimensional

- jerarquía de esquema

  día < mes < cuatrimestre < año

- Se pueden agrupar las jerarquías

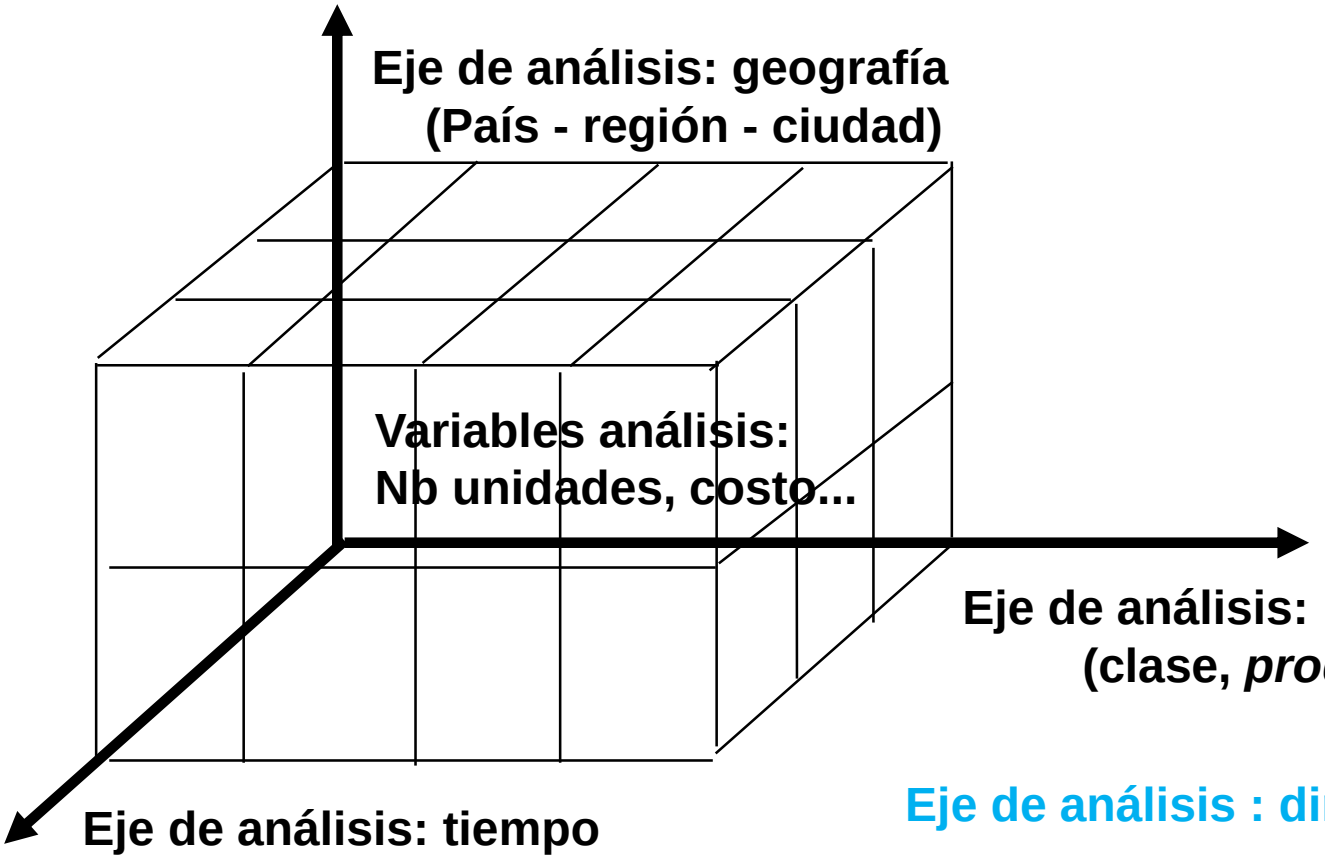
  {día 1 al 10}

  {días} < 30

Las multidimensiones

- **Dimensiones:**
  - Tiempo
  - Geografía
  - Productos
  - Clientes
  - Canales de ventas.....
- **Indicadores:**
  - Número de unidades vendidas
  - Costo

Cubo de dato, dimensiones e indicadores

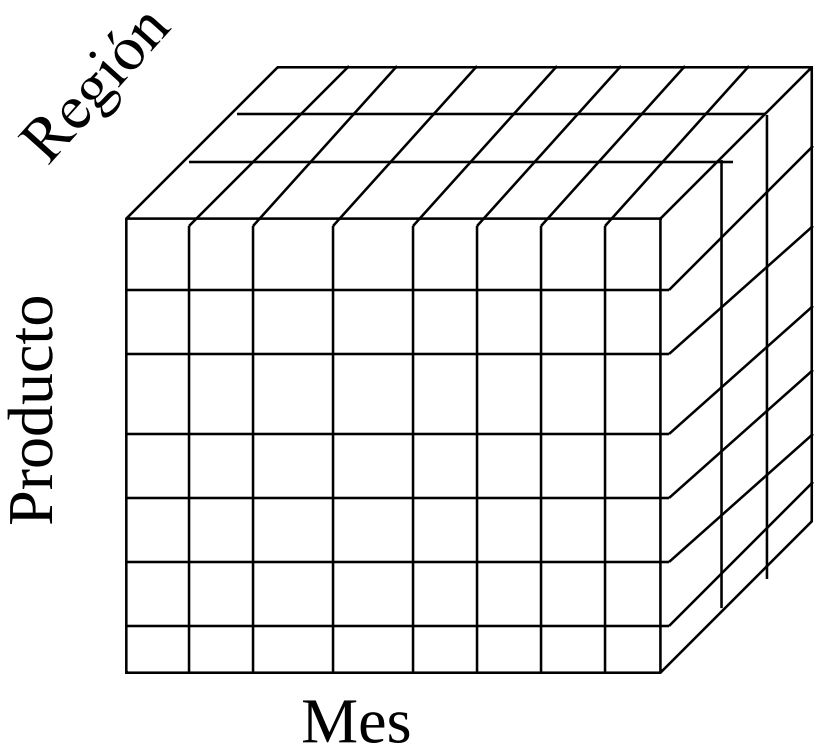


Eje de análisis : dimensiones

Variables análisis: indicadores

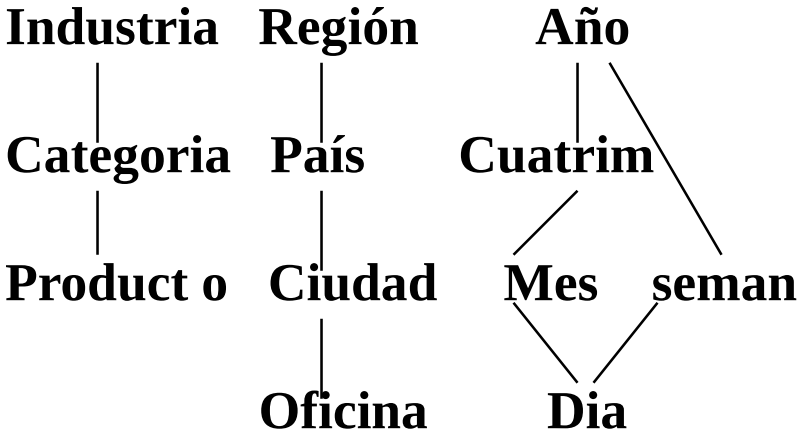
Datos Multidimensionales e e indicadores

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



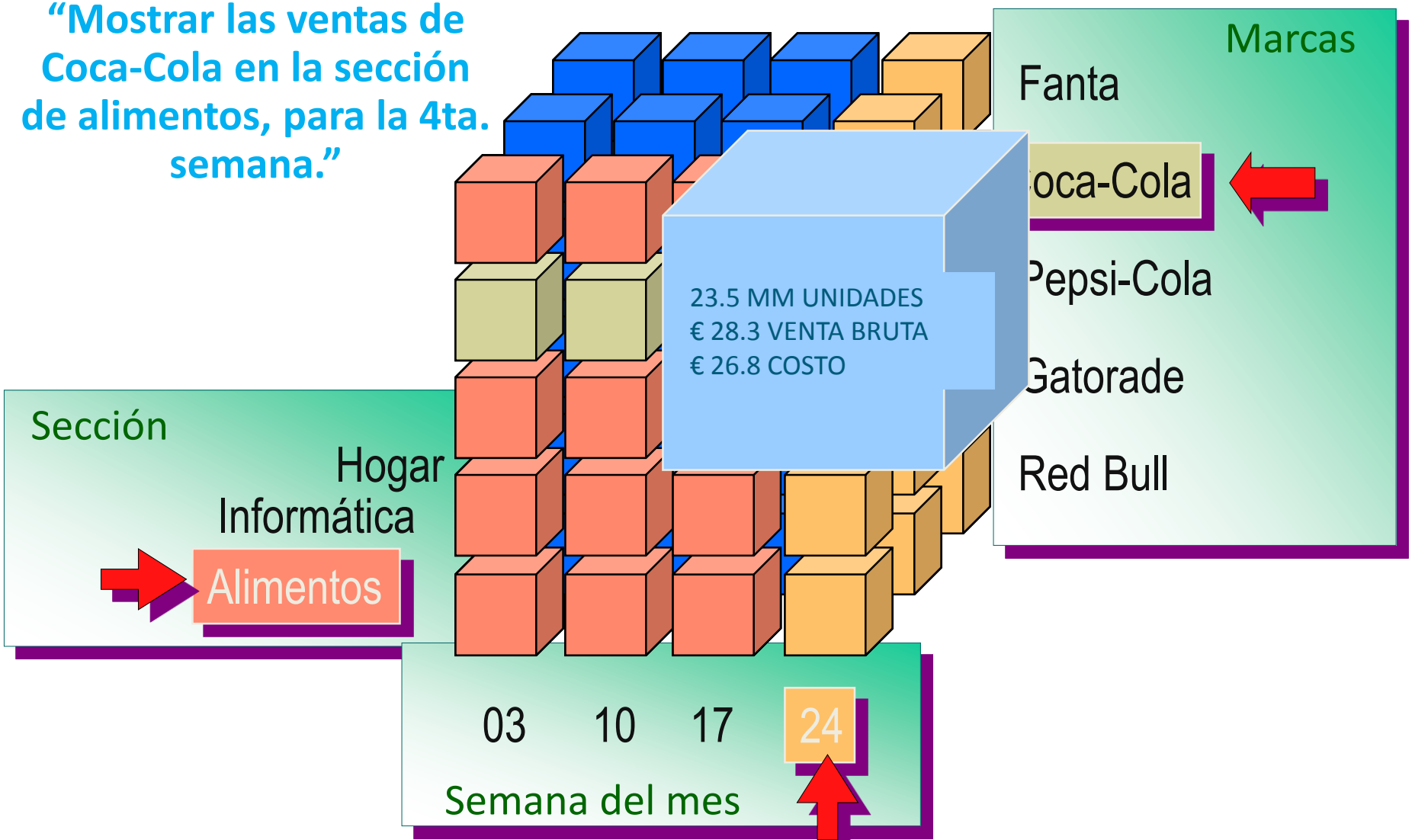
Dimensiones: Producto, Localidad, Tiempo

Caminos jerarquicos



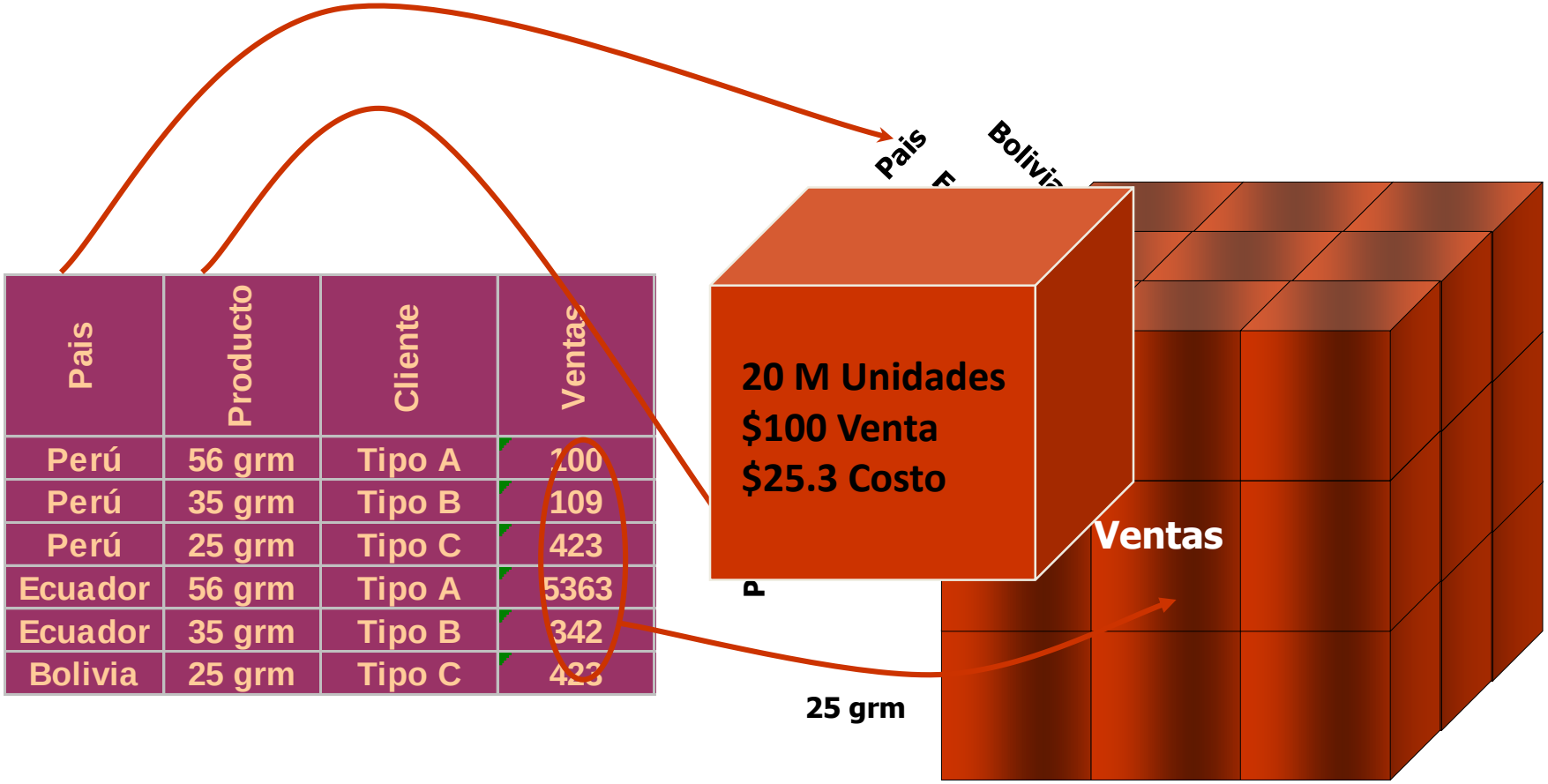
Cubo Multidimensional e e indicadores

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



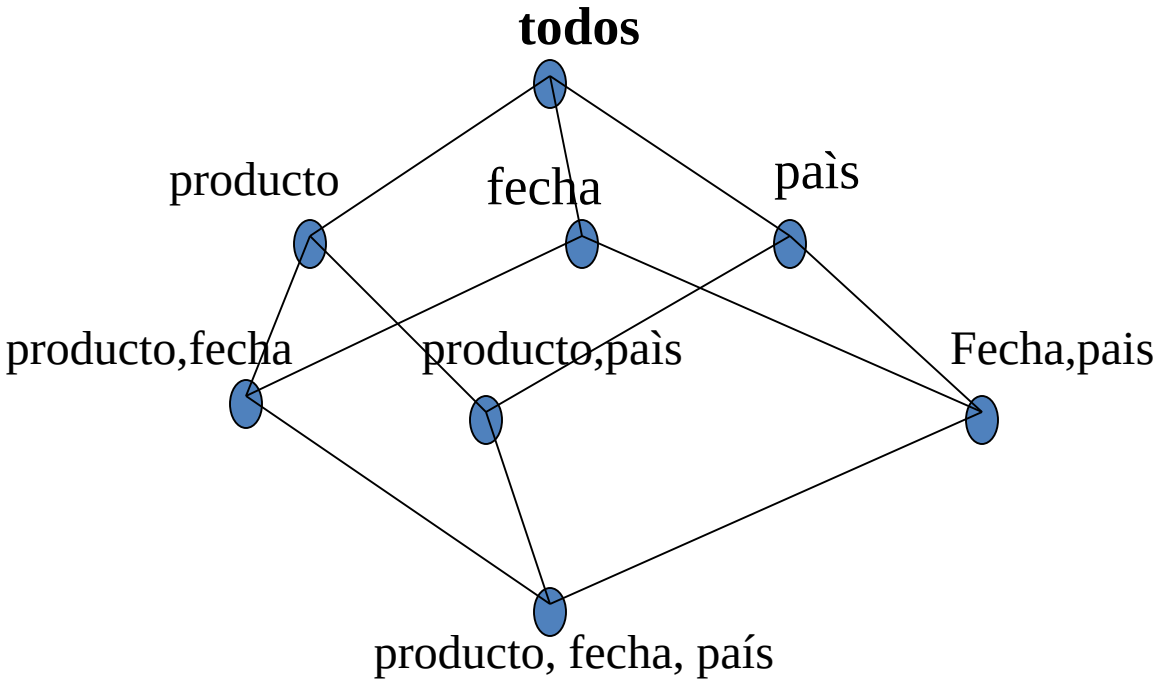
Cubo Multidimensional e indicadores

Pais	Producto	Cliente	Ventas
Perú	56 grm	Tipo A	100
Perú	35 grm	Tipo B	109
Perú	25 grm	Tipo C	423
Ecuador	56 grm	Tipo A	5363
Ecuador	35 grm	Tipo B	342
Bolivia	25 grm	Tipo C	423



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

Cuboides correspondientes al Cubo



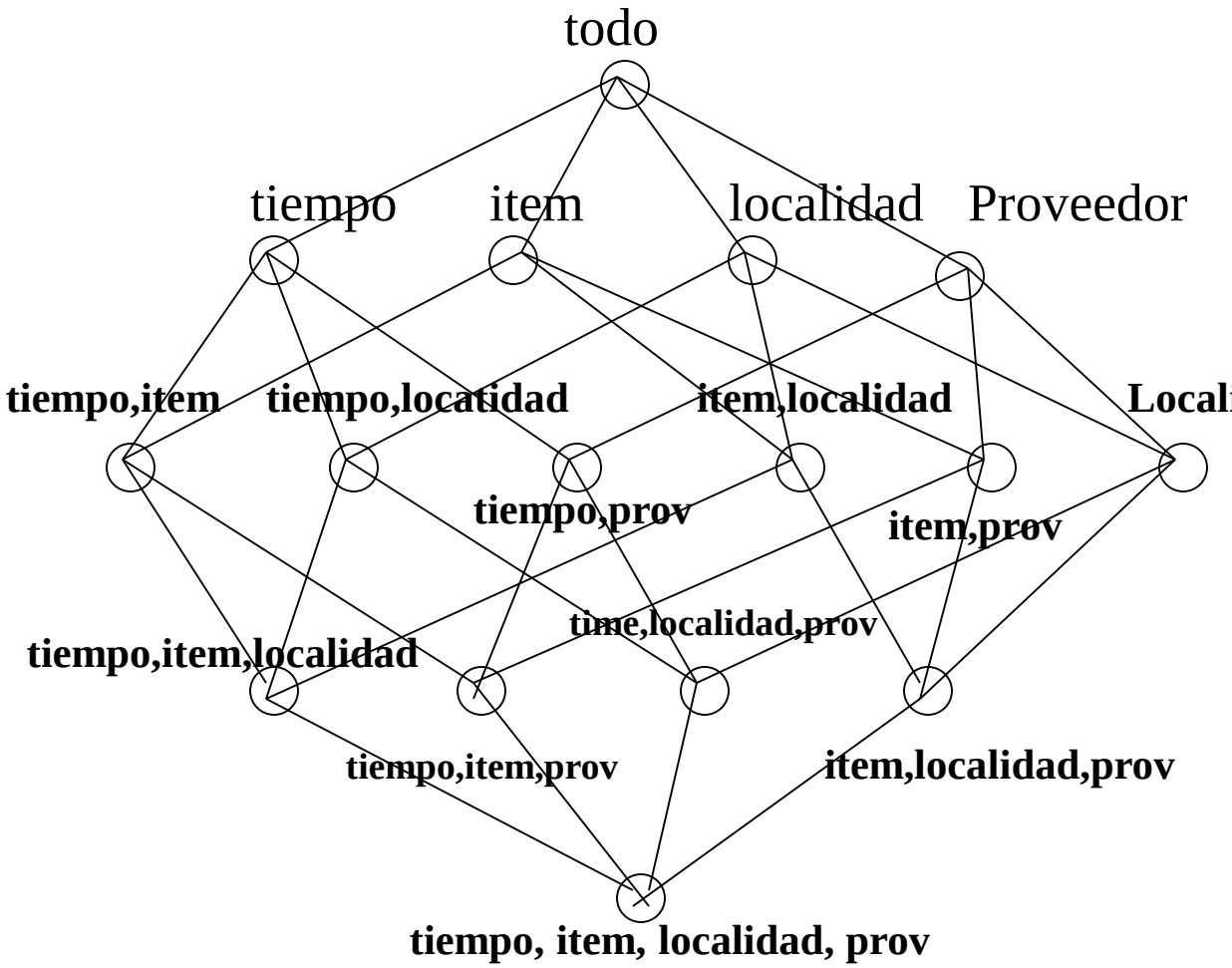
0-D cuboide

1-D cuboides

2-D cuboides

3-D cuboides

Cuboides correspondientes al Cubo



0-D cuboide

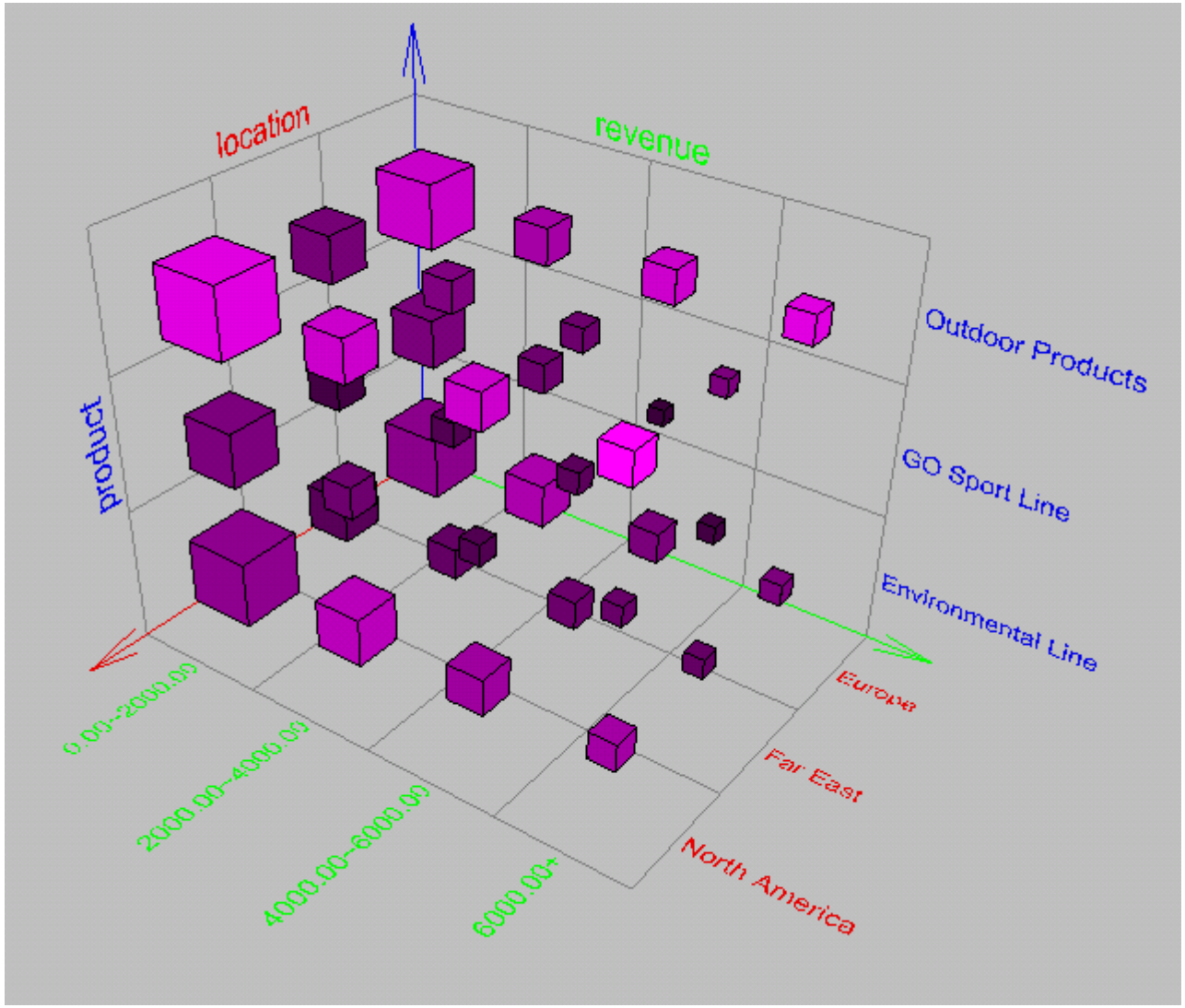
1-D cuboides

2-D cuboides

3-D cuboides

4-D cuboides

Navegar por un cubo de datos



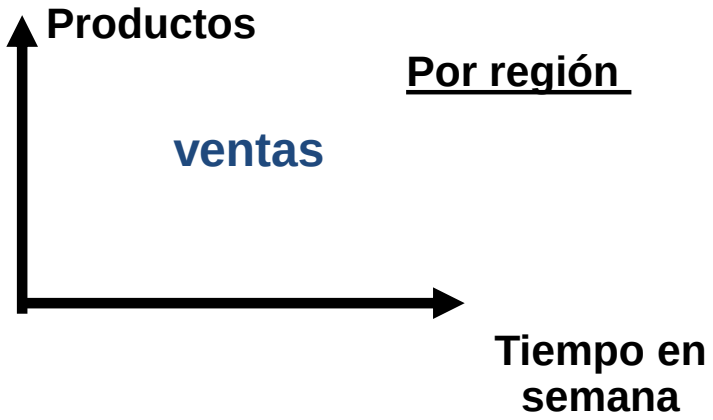
- Visualización
- OLAP

Navegar por un cubo de datos

2 dimensiones

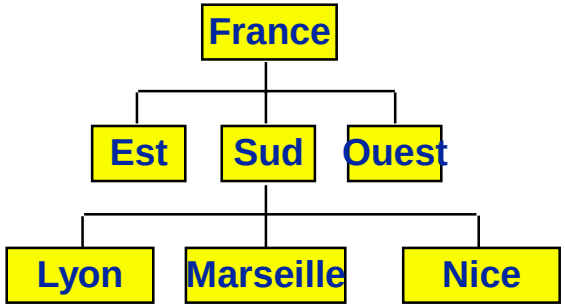
3 dimensiones

Copo de nieve

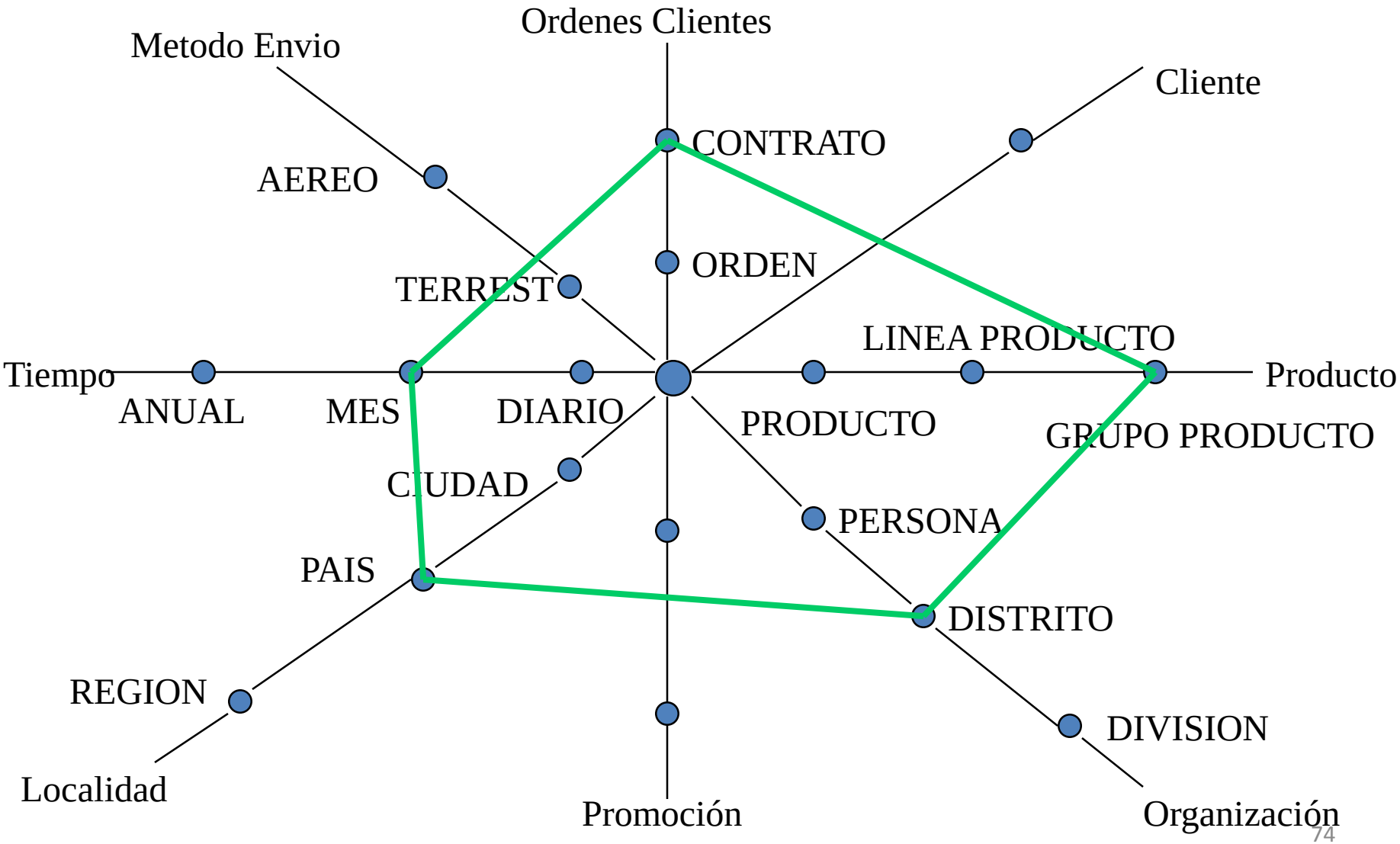


1 dimensión

Zoom a una dimensión



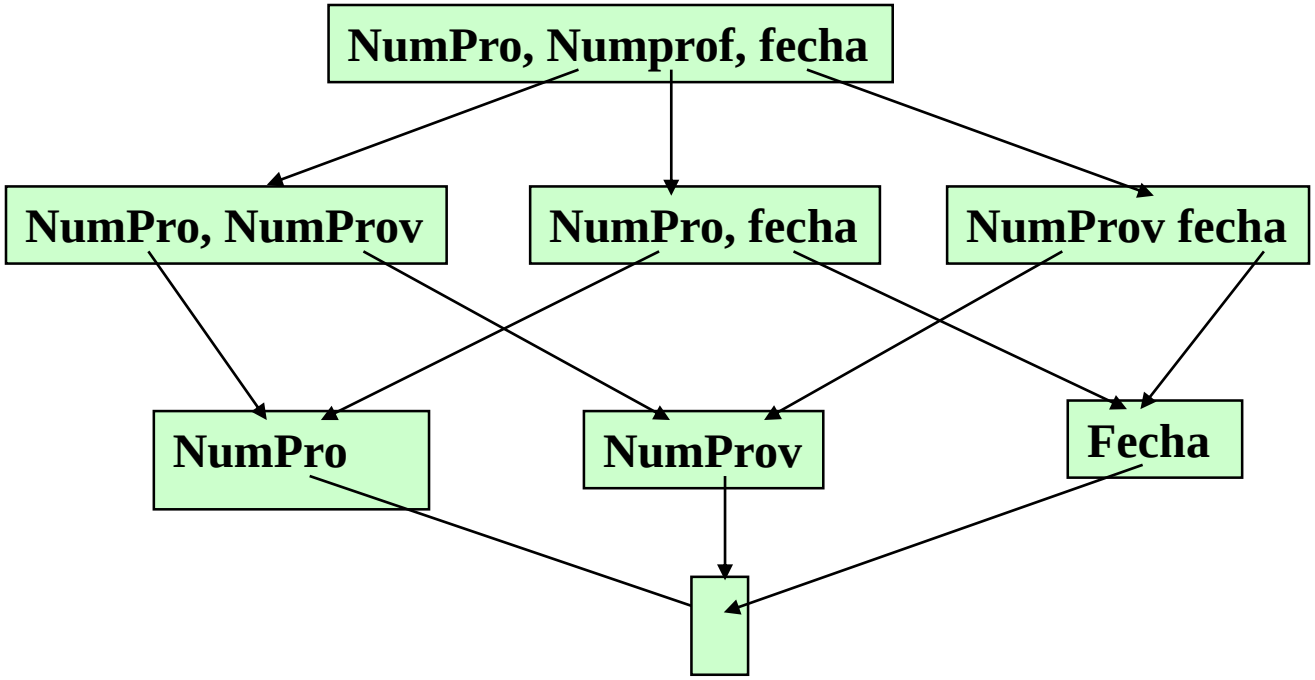
Modelo de consulta



Navegar por un cubo de datos

- Un cubo se puede rotar, agrupar, etc.

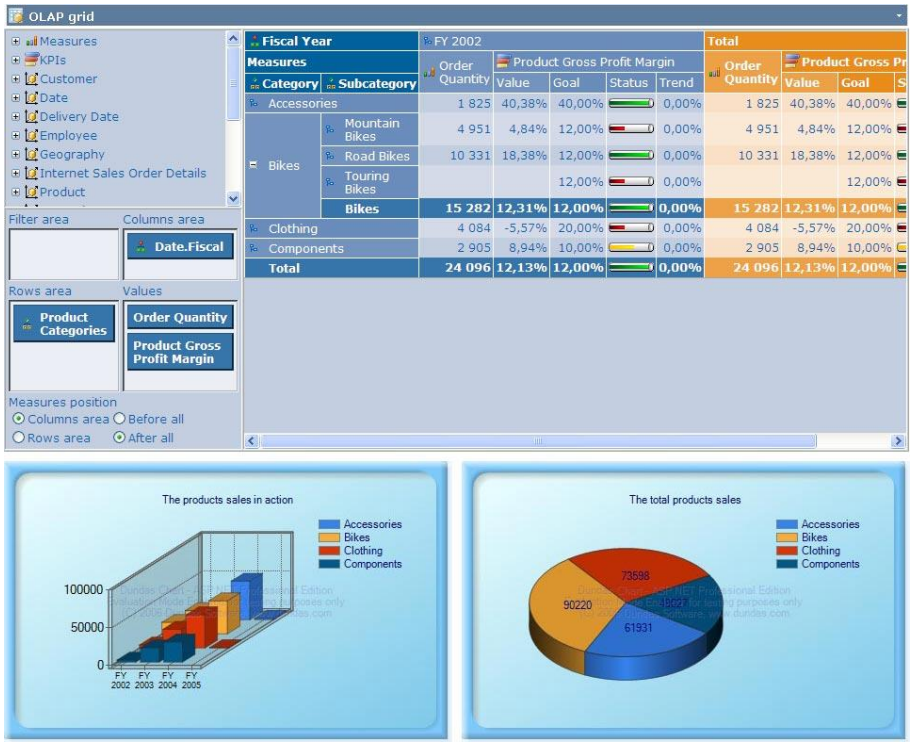
Se obtienen retículas de puntos de vista



Herramientas para explotación del Datawarehouse

Análisis multidimensional (OLAP online analytical processing)

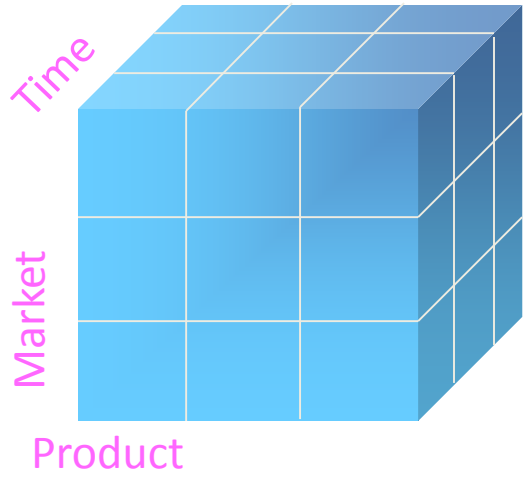
Facilitan el análisis de datos a través de dimensiones y jerarquías, uutilizando consultas rápidas predefinidas



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

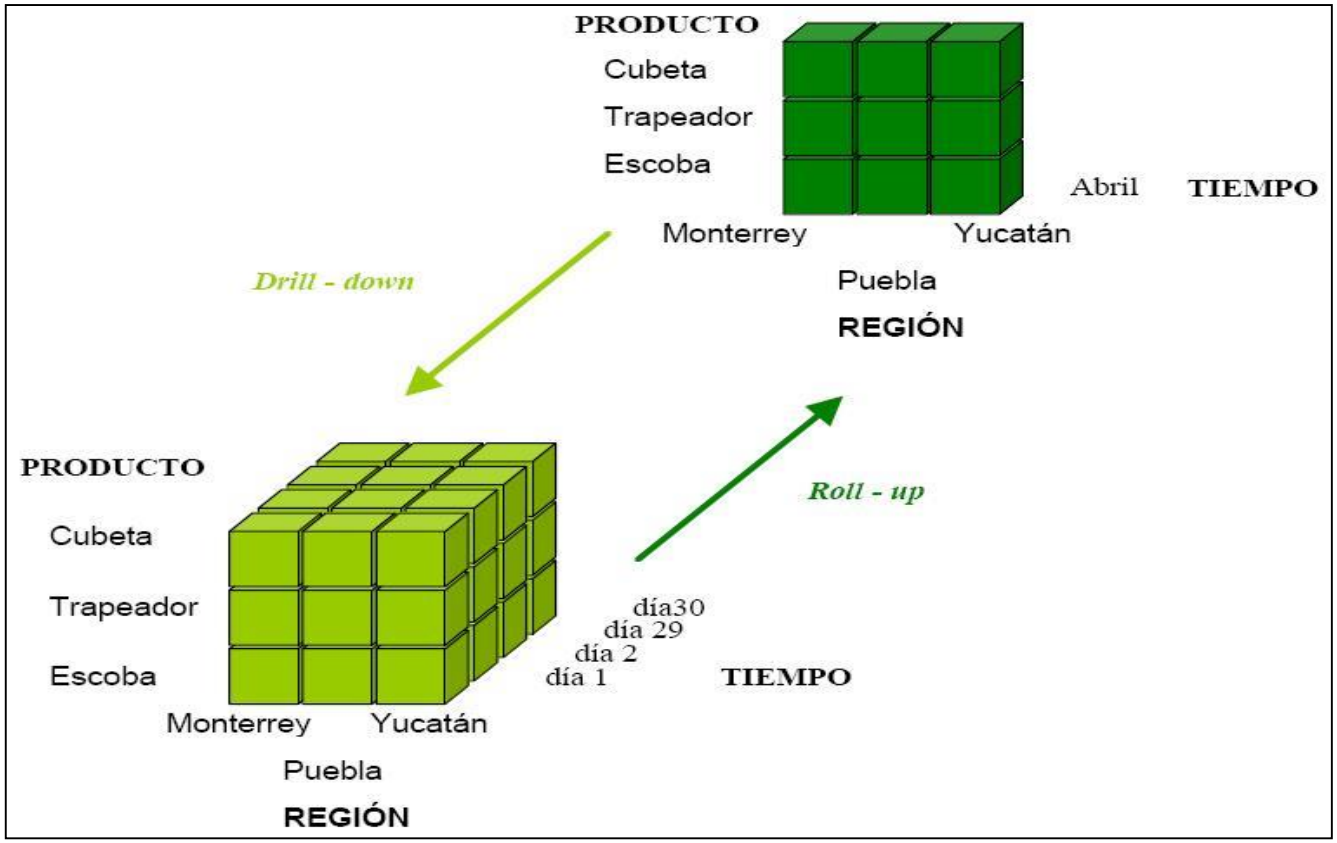
Arquitecturas OLAP

- **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

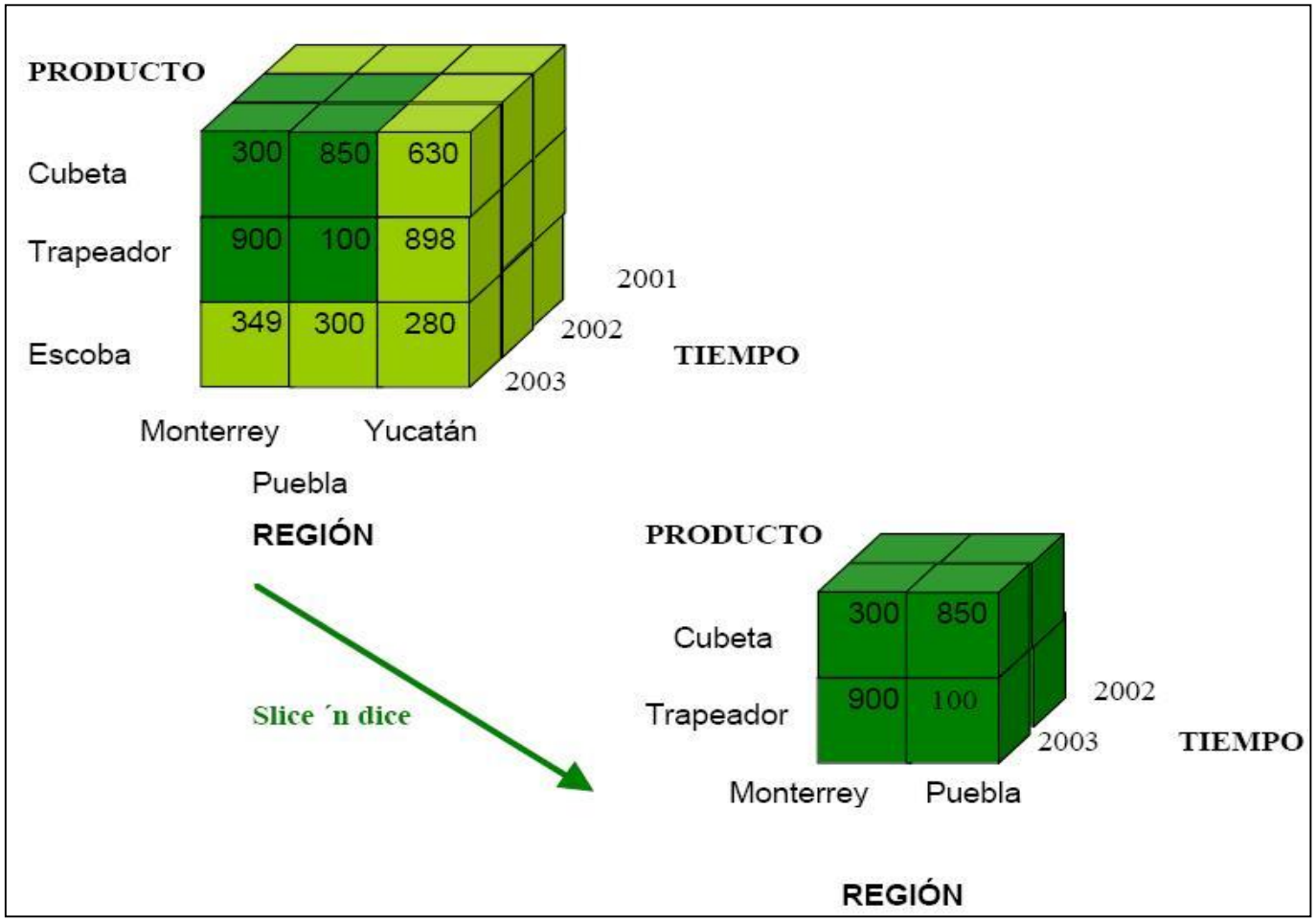
Roll up (drill-up): agrega medidas que van de un nivel Ni a un nivel mas general Nj de una dimensión.

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.



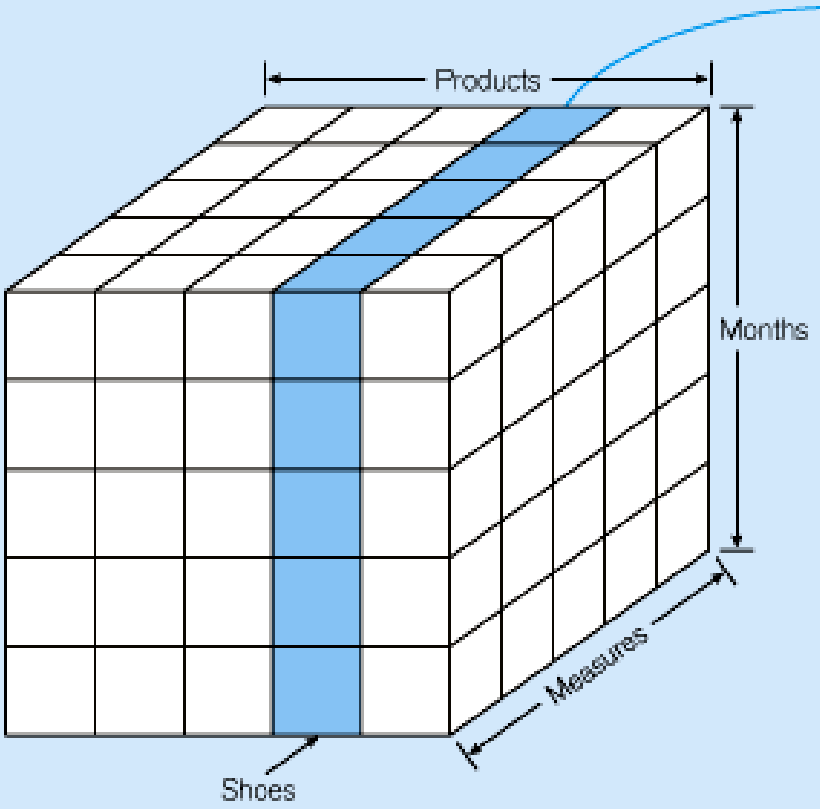
Otras operaciones

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

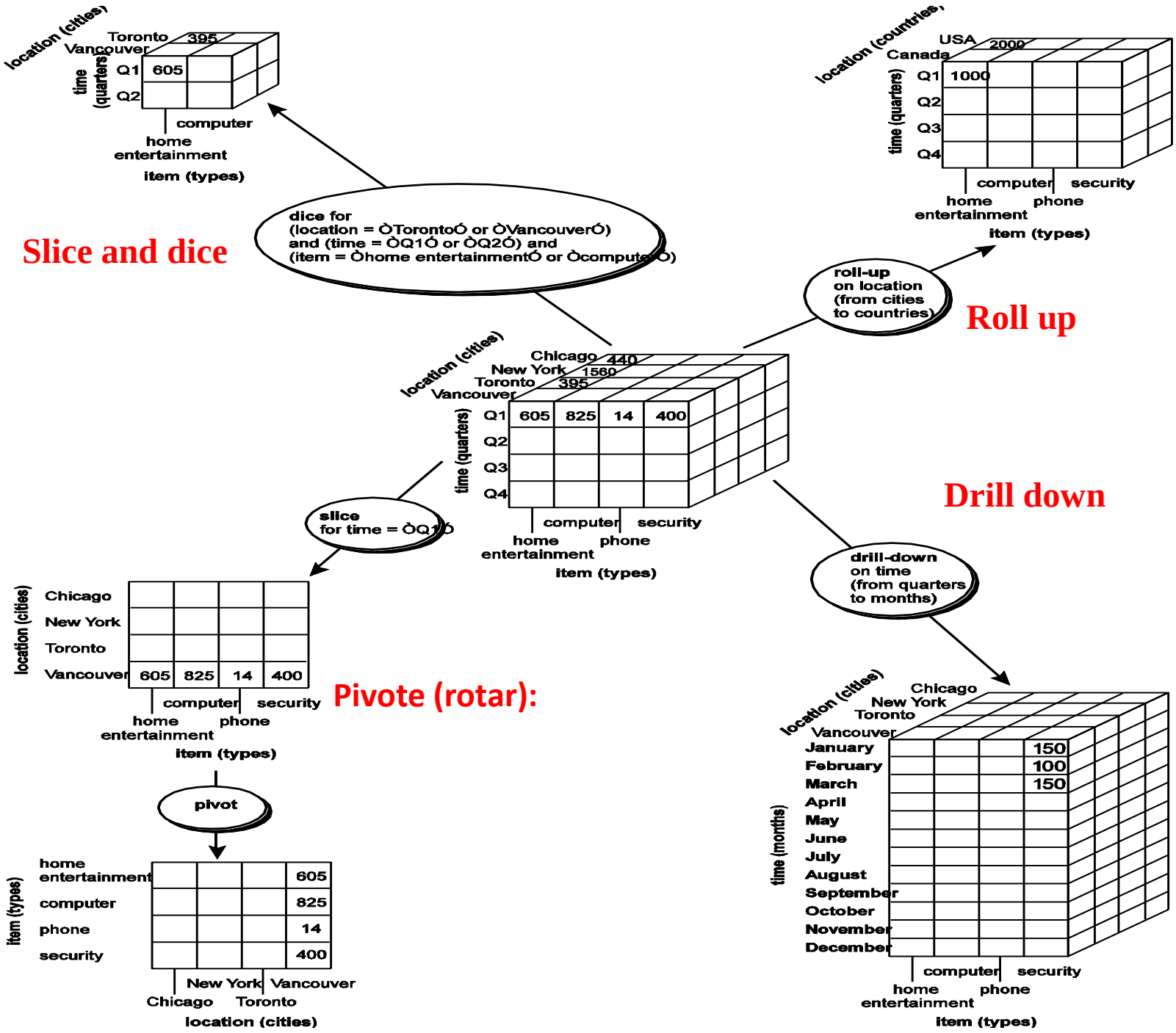
Pivote (rotar)
Rotar el cubo

Cortando/rebanando un cubo de datos



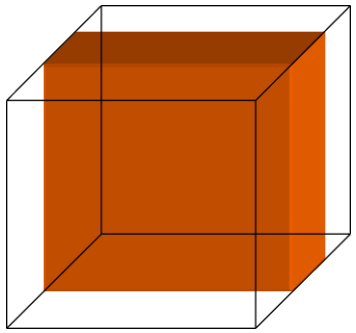
	Measure		
	Units	Revenue	Cost
January	250	1564	1020
February	200	1275	875
March	350	1800	1275
April	400	1935	1500
May	485	2000	1560

Product: Shoes

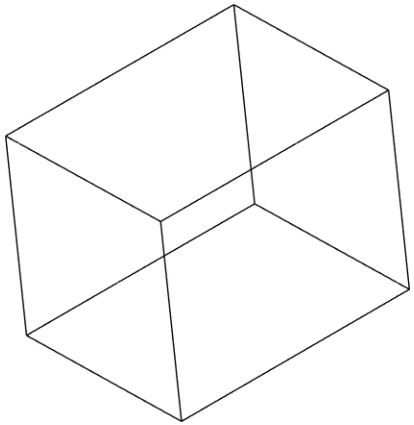

location (cities)
Toronto 395
Vancouver
time (quarters)
Q1 605
Q2
computer
home entertainment
item (types)
Slice and dice
dice for
(location = ÒTorontoÓ or ÒVancouverÓ)
and (time = ÒQ1Ó or ÒQ2Ó) and
(item = Òhome entertainmentÓ or Òcomputer
location (countries)
USA 2000
Canada
Q1 1000
Q2
Q3
Q4
computer security
home phone
entertainment
item (types)
roll-up
on location
(from cities
to countries)
Roll up
location (cities)
Chicago 440
New York 1560
Toronto 395
Vancouver
Q1 605 825 14 400
Q2
Q3
Q4
computer security
home phone
entertainment
item (types)
Drill down
slice
for time = ÒQ1
drill-down
on time
(from quarters
to months)
location (cities)
Chicago
New York
Toronto
Vancouver 605 825 14 400
computer security
home phone
entertainment
item (types)
Pivote (rotar):
pivot
item (types)
home entertainment 605
computer 825
phone 14
security 400
New York Vancouver
Chicago Toronto
location (cities)
location (cities)
Chicago
New York
Toronto
Vancouver
January 150
February 100
March 150
April
May
June
July
August
September
October
November
December
time (months)
computer security
home phone
entertainment
item (types)

Resumen Operaciones clásicas OLAP

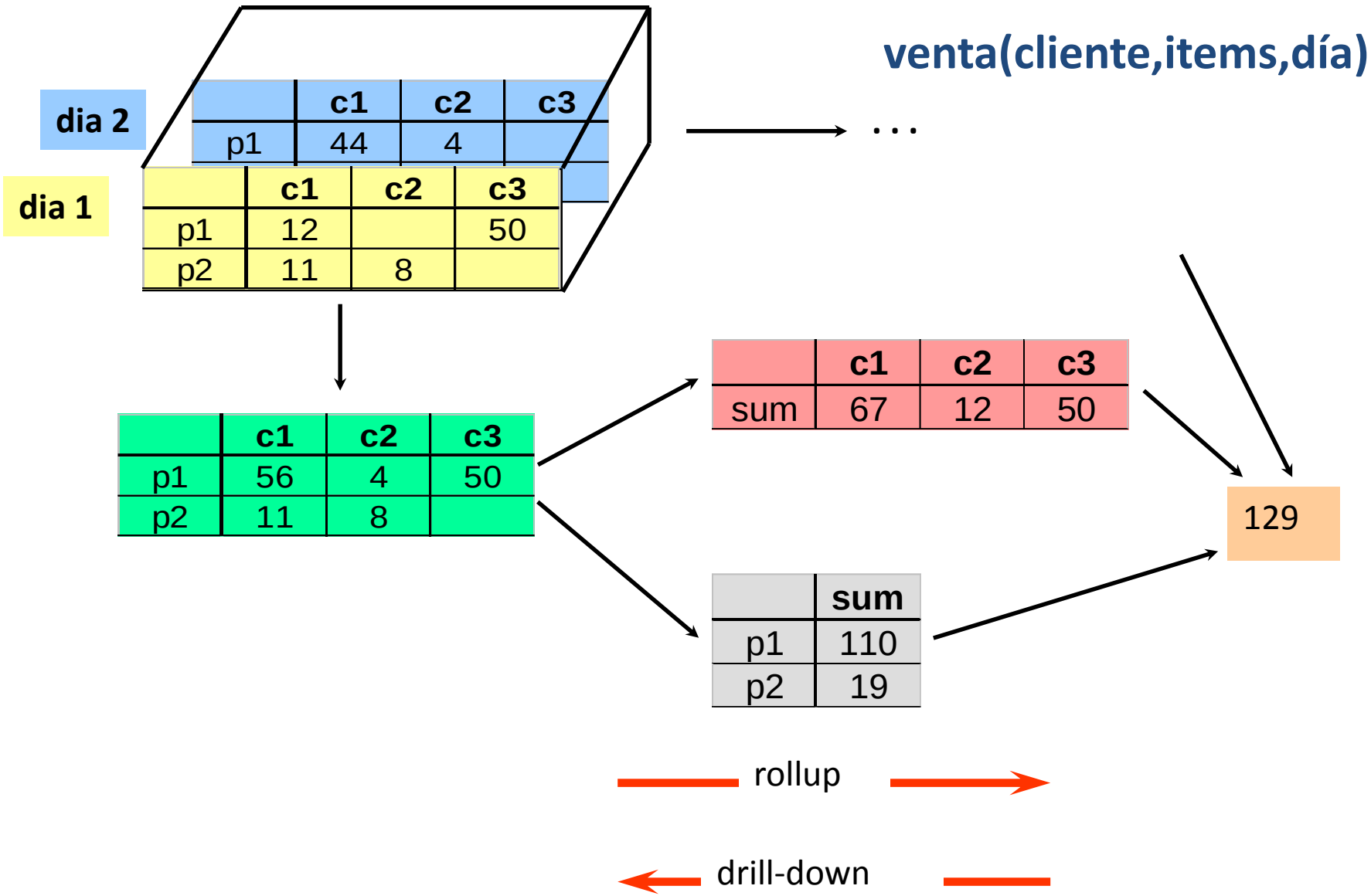
- *Rollup*: decrese nivel de detalle
- *Drill-down*: aumenta nivel de detalle
- *Slice-and-dice*: selección y proyección



- *Pivot*: re-orienta vista multidimensional

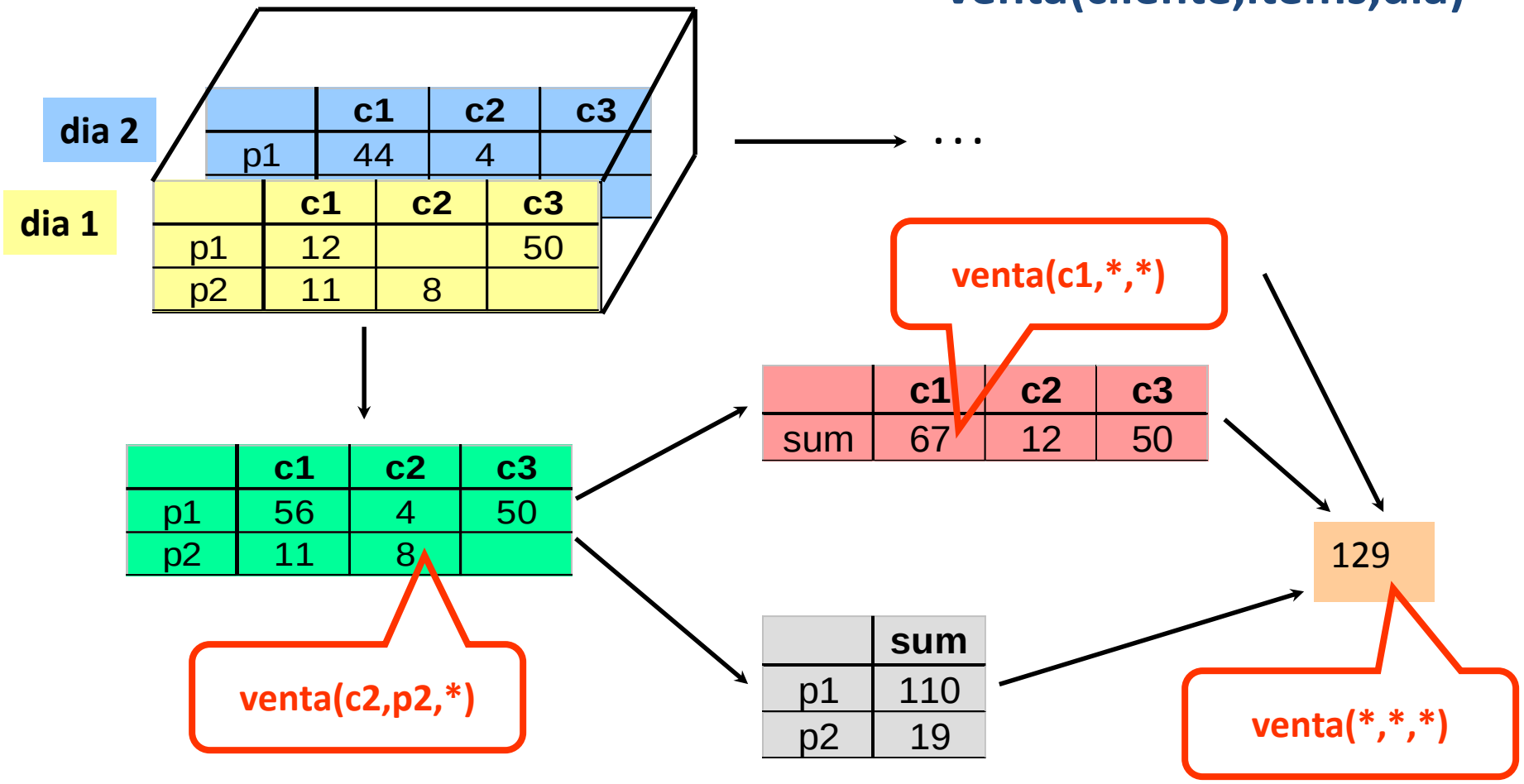


Agregación en Cubos



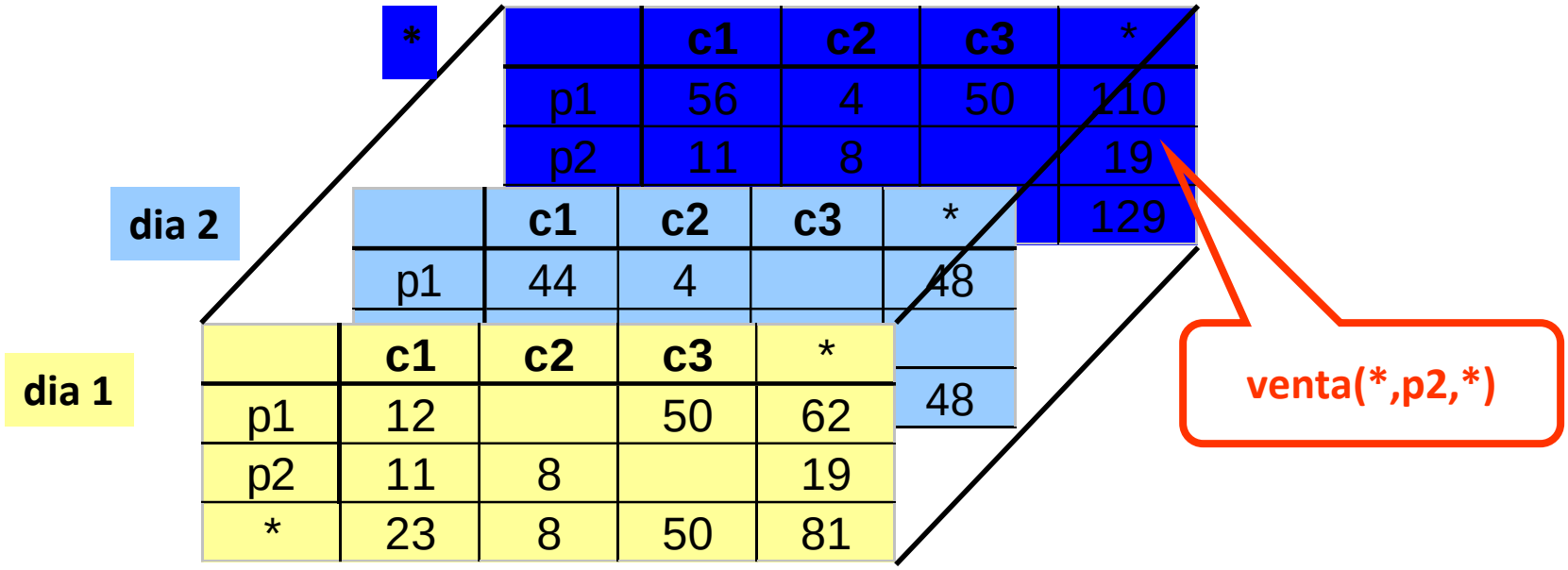
Agregación en Cubos

venta(cliente,items,día)



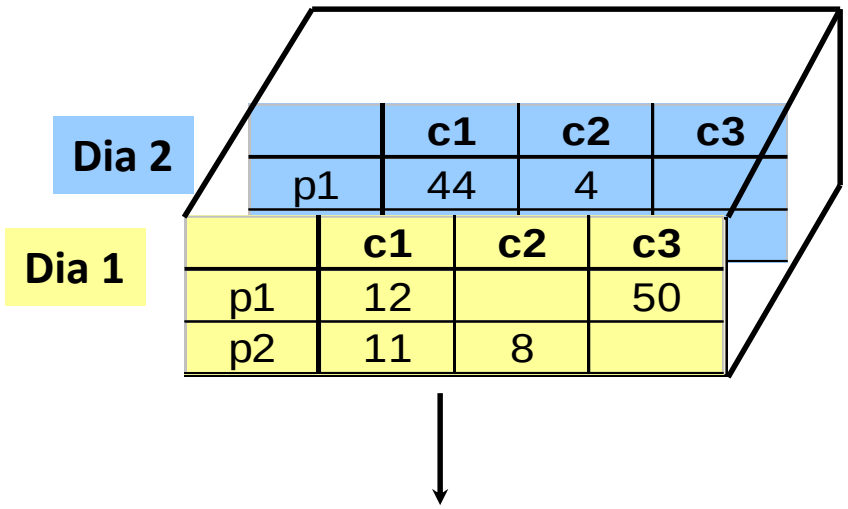
Cara de Cubos

venta(cliente,items,día)



Agregación usando jerarquía

venta(cliente,items,día)



	region A	region B
p1	56	54
p2	11	8

cliente

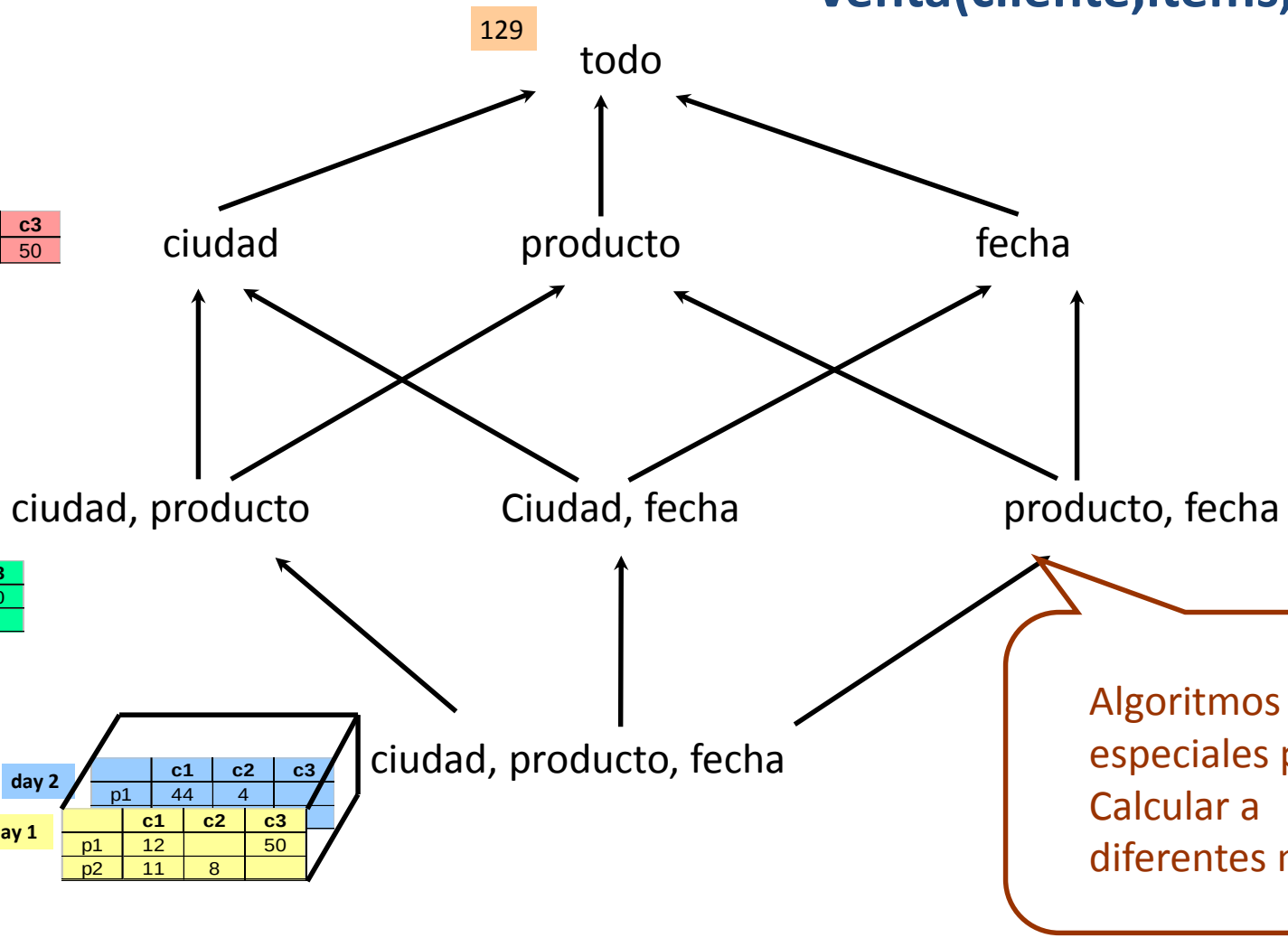
region

país

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

Agregación en Cubos

venta(cliente,items,día)



	c1	c2	c3
p1	67	12	50

	c1	c2	c3
p1	56	4	50
p2	11	8	

day 2

	c1	c2	c3
p1	44	4	

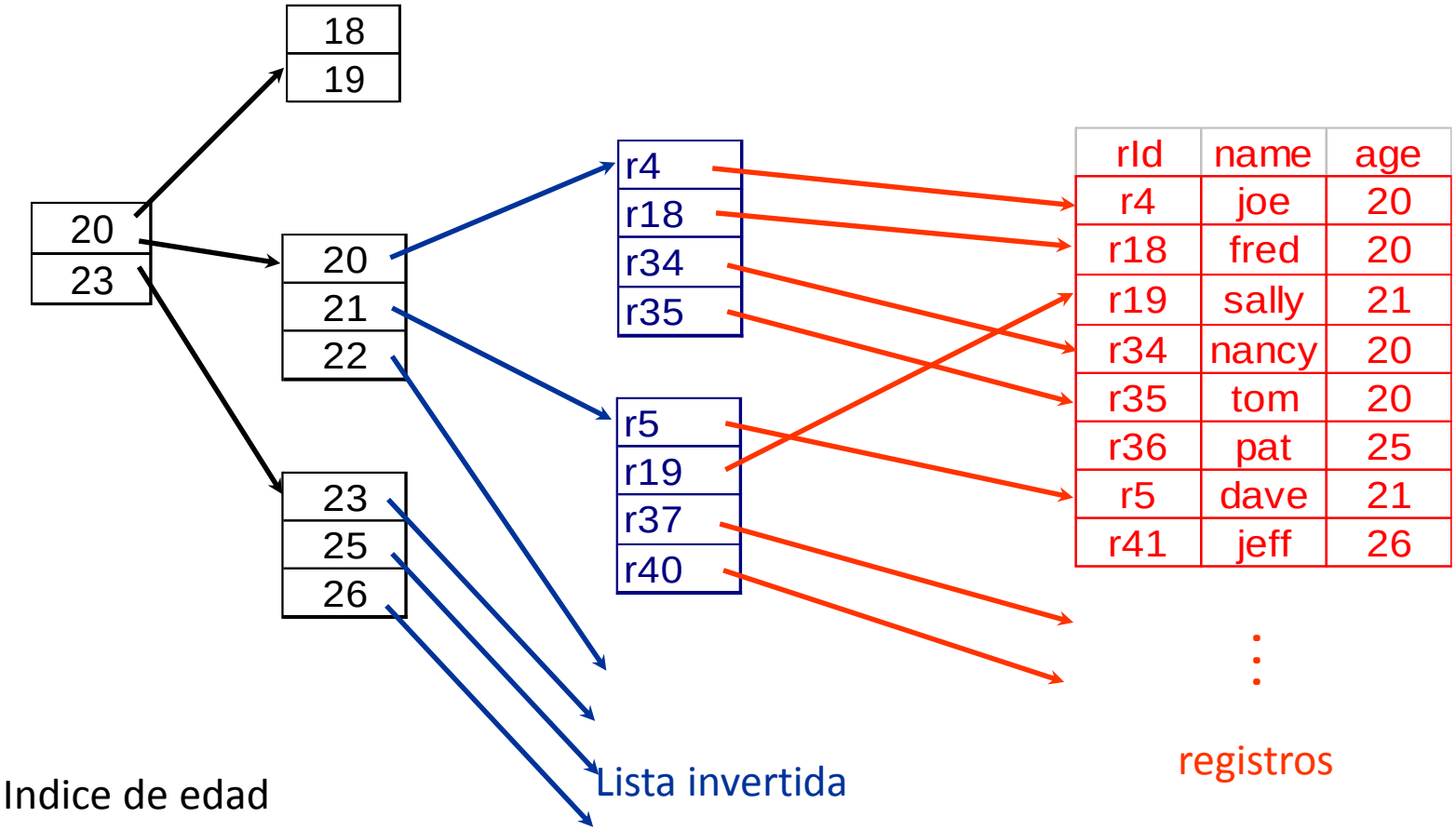
day 1

	c1	c2	c3
p1	12		50
p2	11	8	

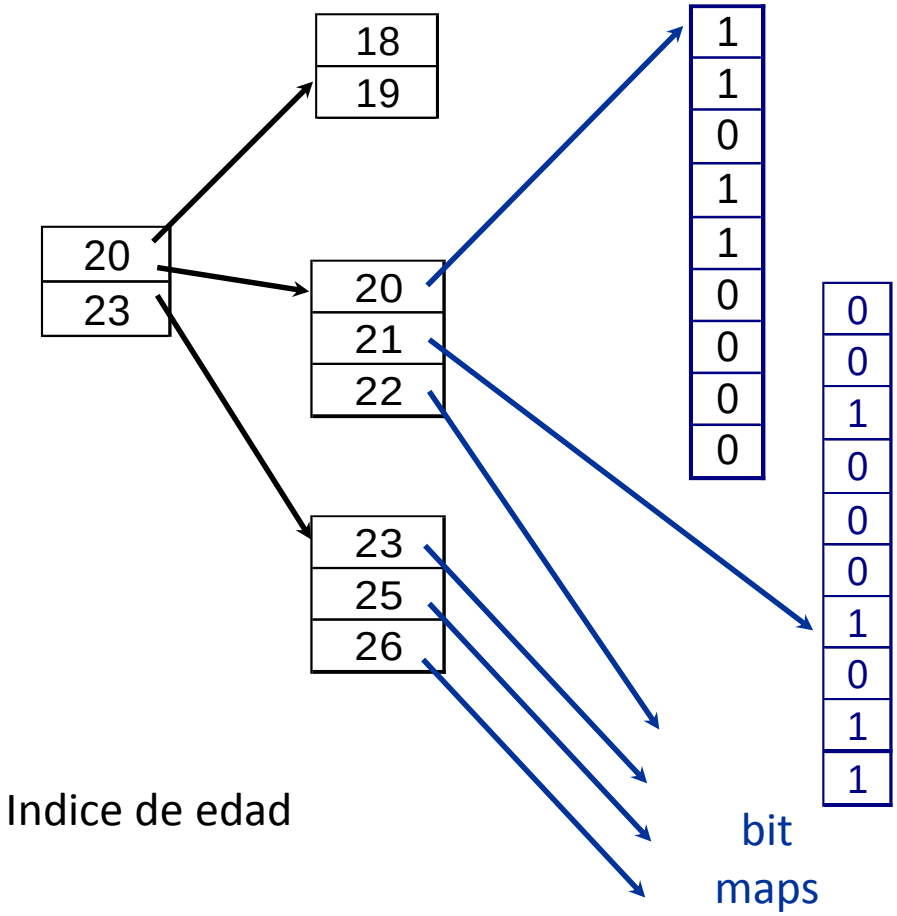
Estructuras Índices

- **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



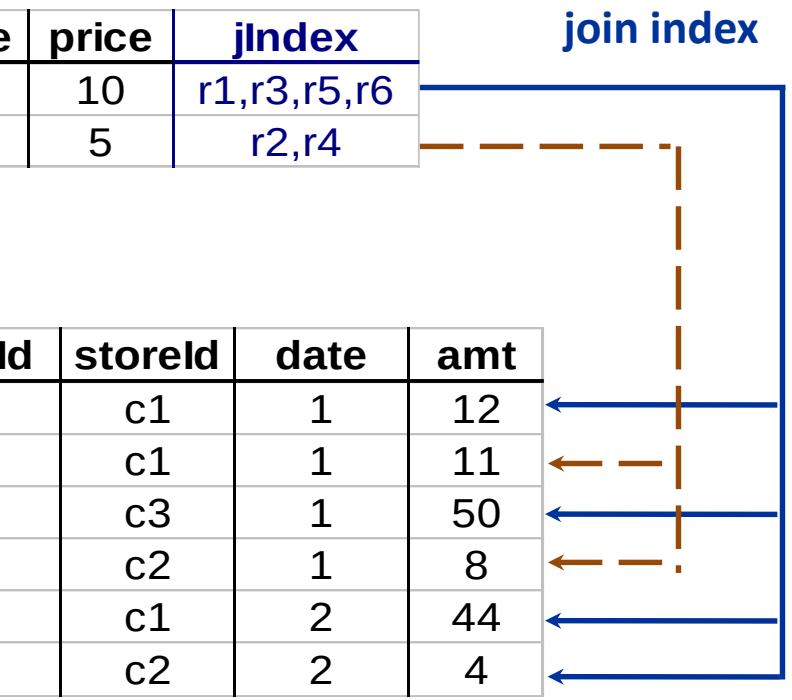
id	name	age
1	joe	20
2	fred	20
3	sally	21
4	nancy	20
5	tom	20
6	pat	25
7	dave	21
8	jeff	26

registros

Join Indexes

product	id	name	price	jIndex
	p1	bolt	10	r1,r3,r5,r6
	p2	nut	5	r2,r4

sale	rId	prodId	storeId	date	amt
	r1	p1	c1	1	12
	r2	p2	c1	1	11
	r3	p1	c3	1	50
	r4	p2	c2	1	8
	r5	p1	c1	2	44
	r6	p1	c2	2	4



- 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

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

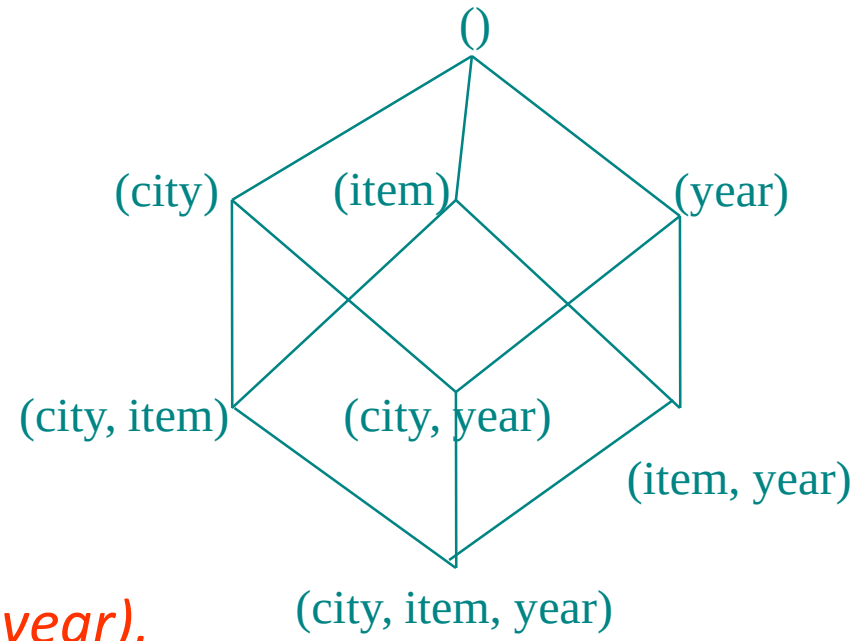
```
SELECT Animal, Lieu,
SUM(Quantite) as Quantite
FROM Animaux
GROUP BY ROLLUP Animal
```

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)

  ()

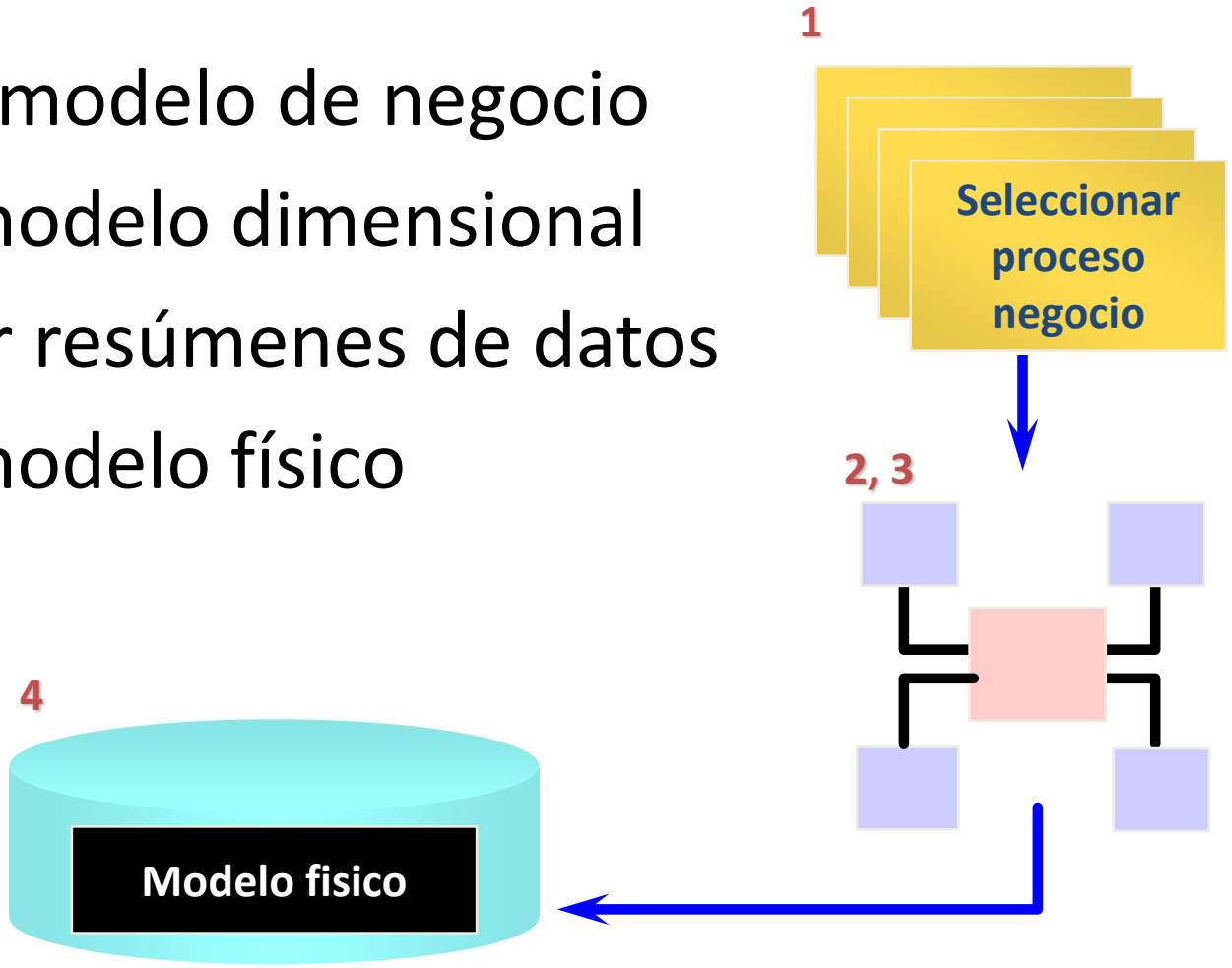
Extension de SQL

- 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 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)

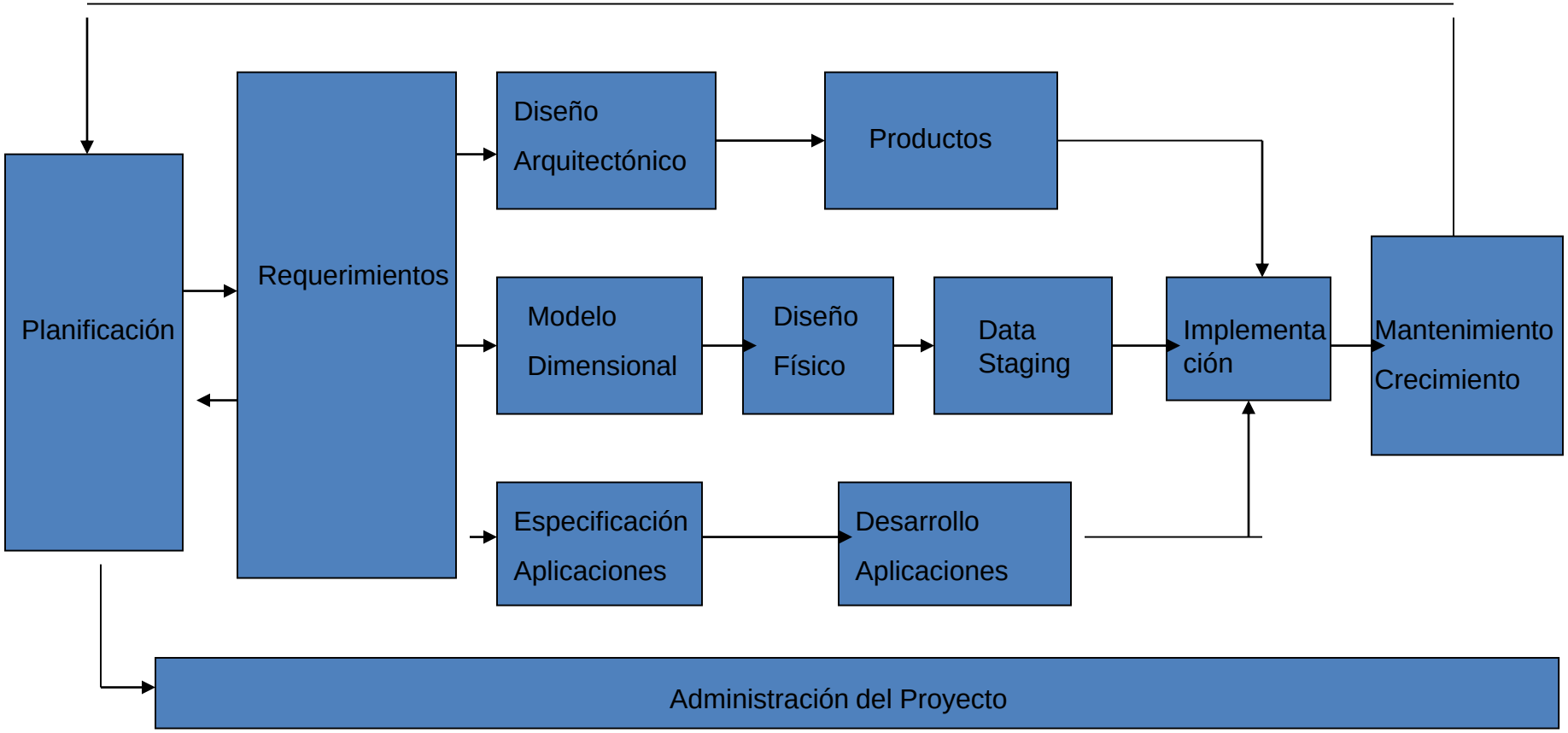
Identificar resumenes de datos

- 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

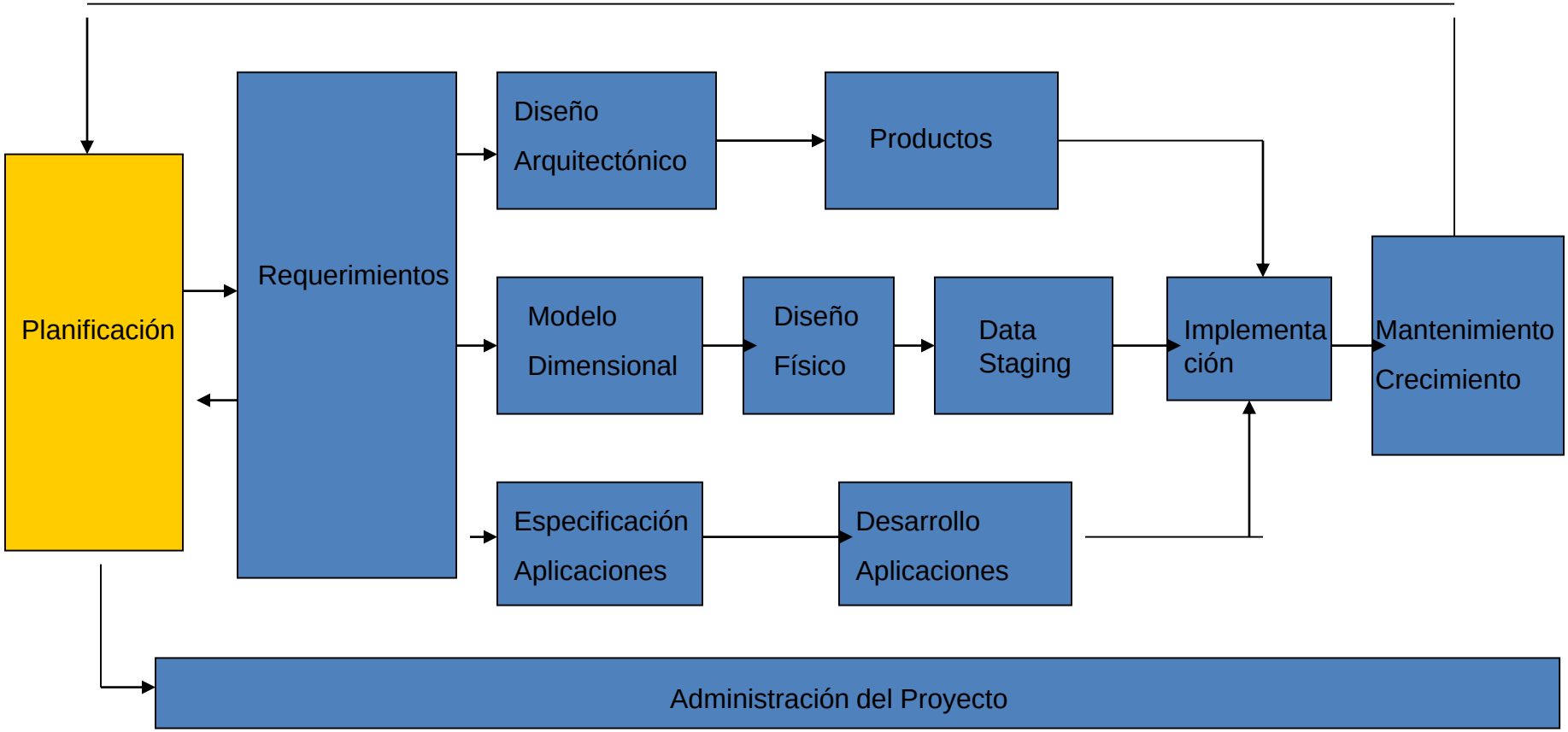
Identificar resumenes de datos

- Promedio
- Máximo
- Total
- Porcentaje

Ciclo de vida



Planificación



Planificación

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

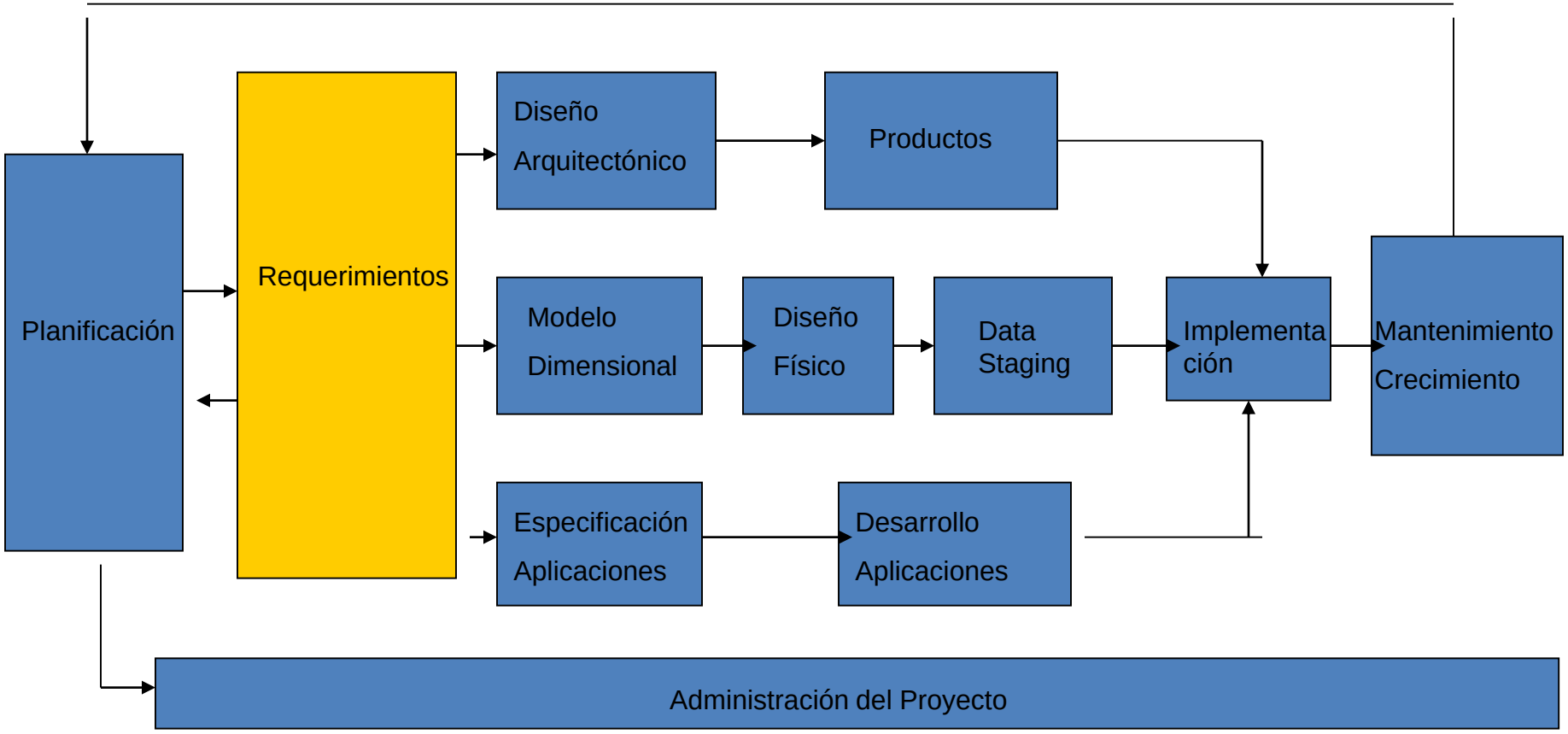
Predisposición de la organización

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

Puntos clave

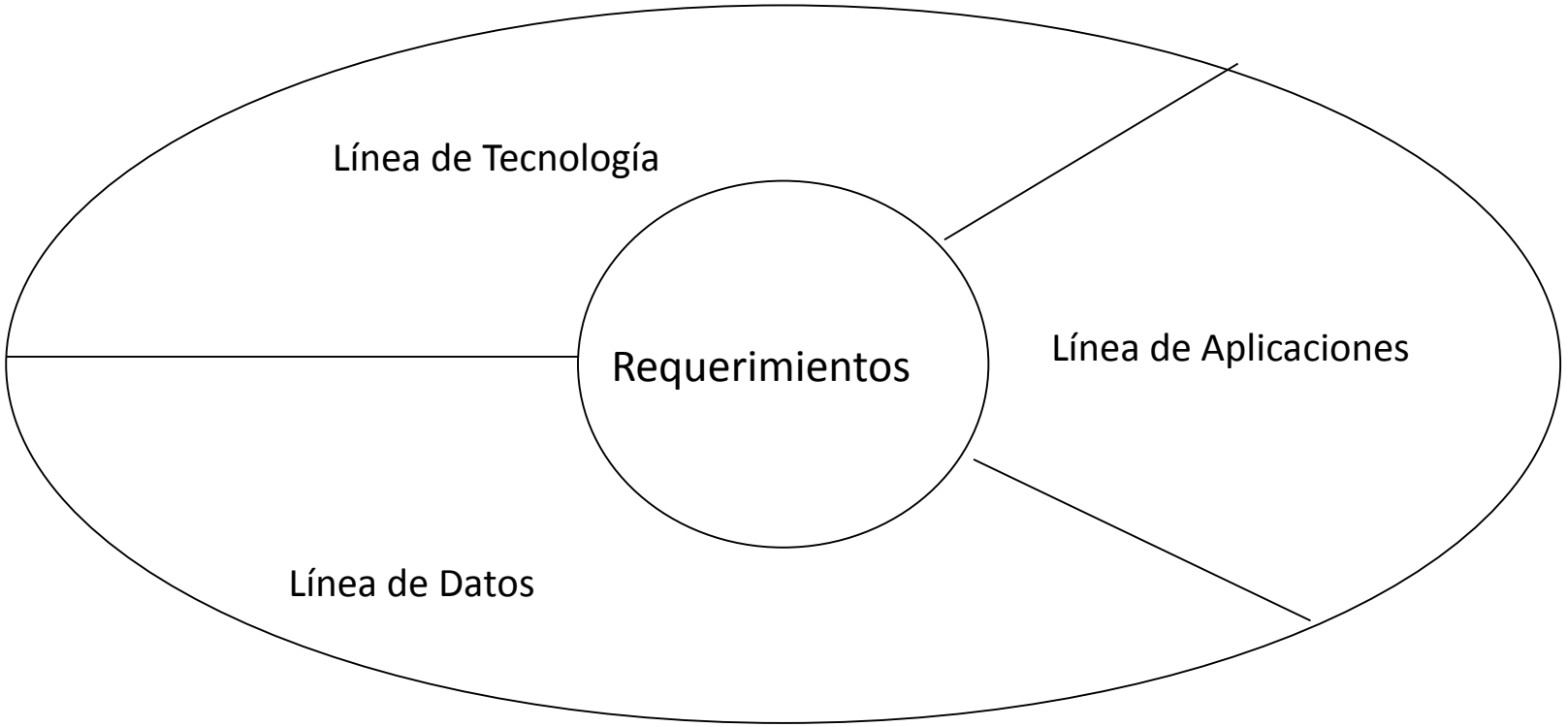
- Buscar un patrocinante bien ubicado
- Hacer un balance entre valor para la organización y manejabilidad
- Desarrollar cuidadosamente el plan del proyecto
- Ser un director de proyecto con capacidad de motivar, administrar y comunicar a todos los niveles

Requerimientos

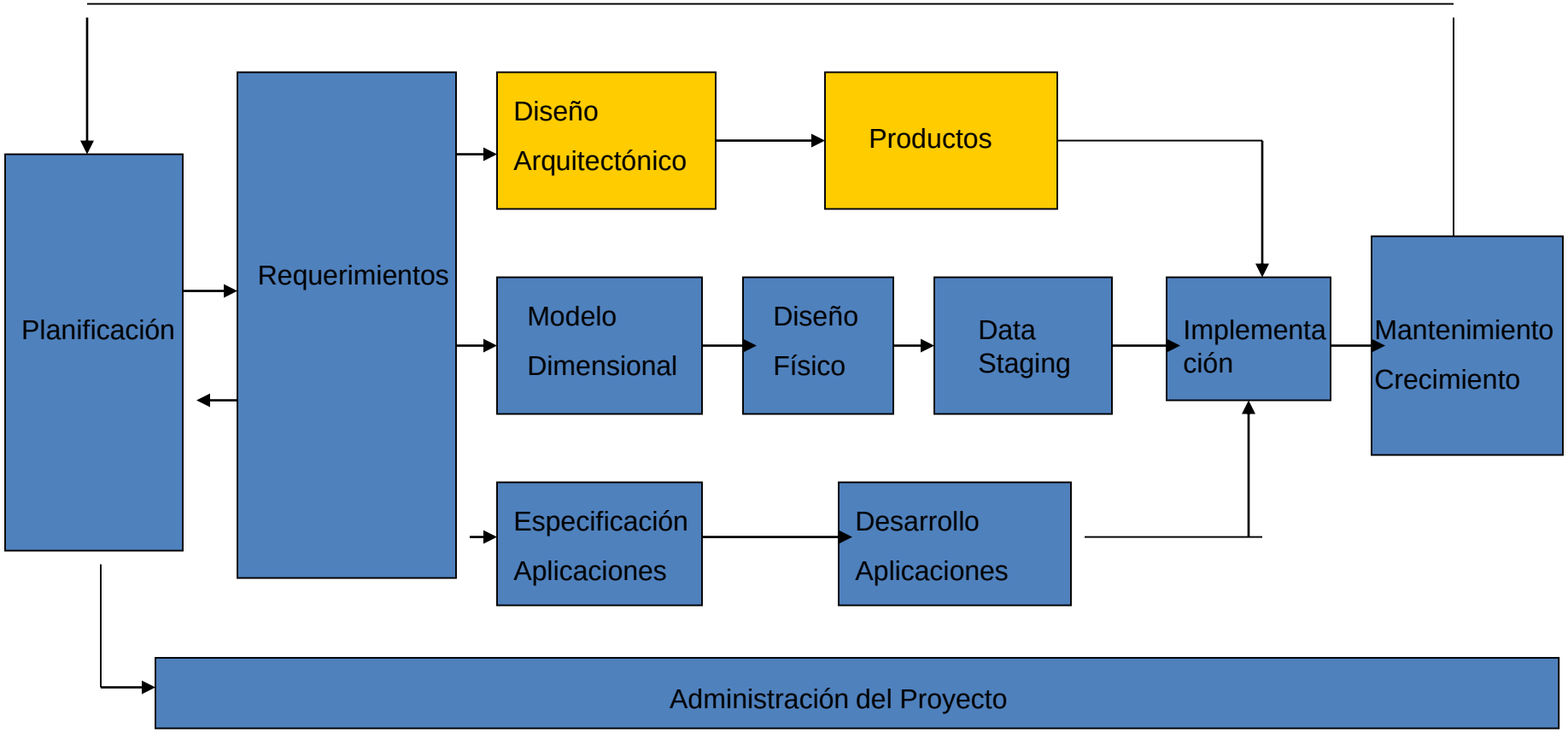


Desarrollo del proyecto

El desarrollo del proyecto se realiza en tres líneas



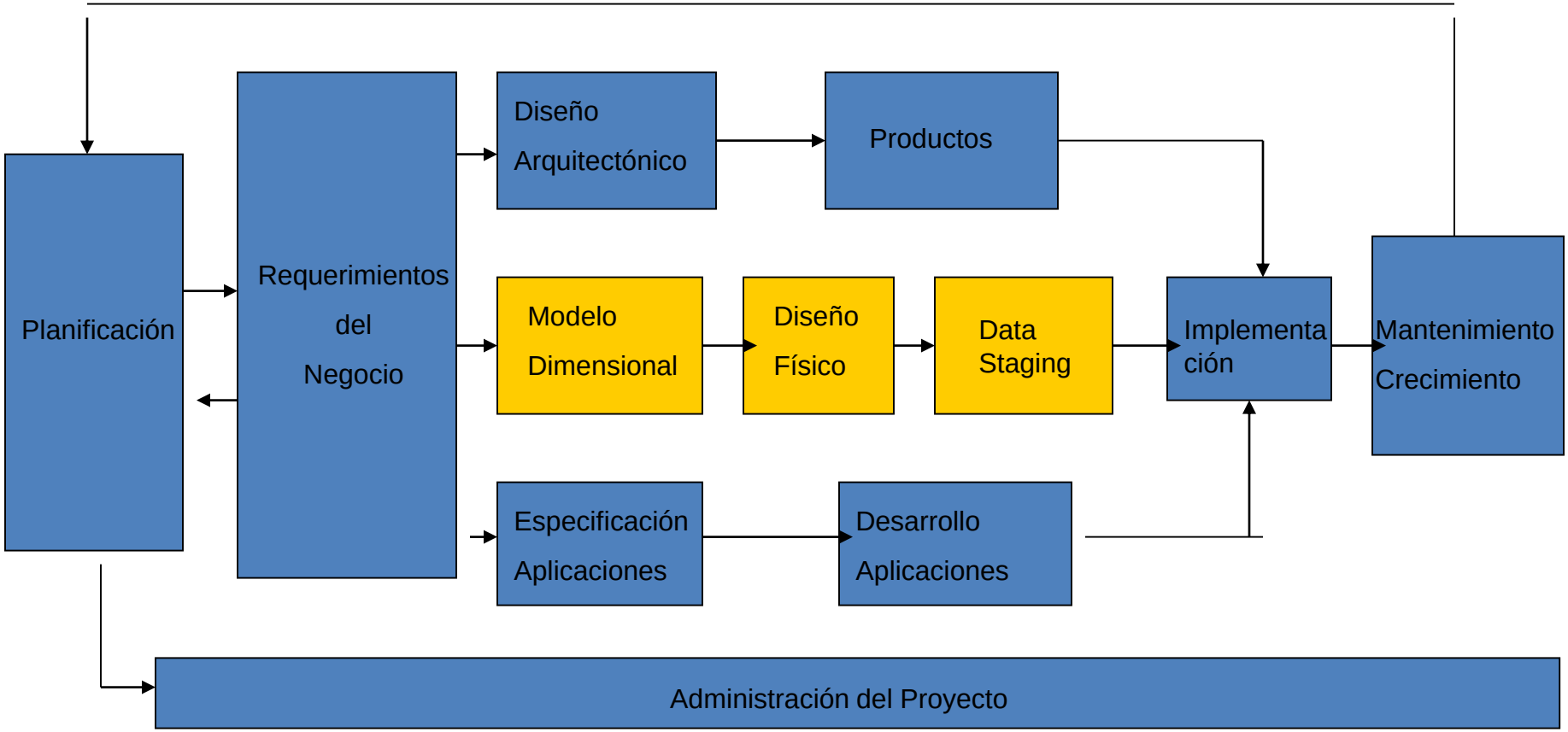
Línea de Tecnología



Selección de productos

- Interés Organizacional
- Requerimientos técnicos
- Criterios de selección
- Factores de ponderación
- Matriz de evaluación

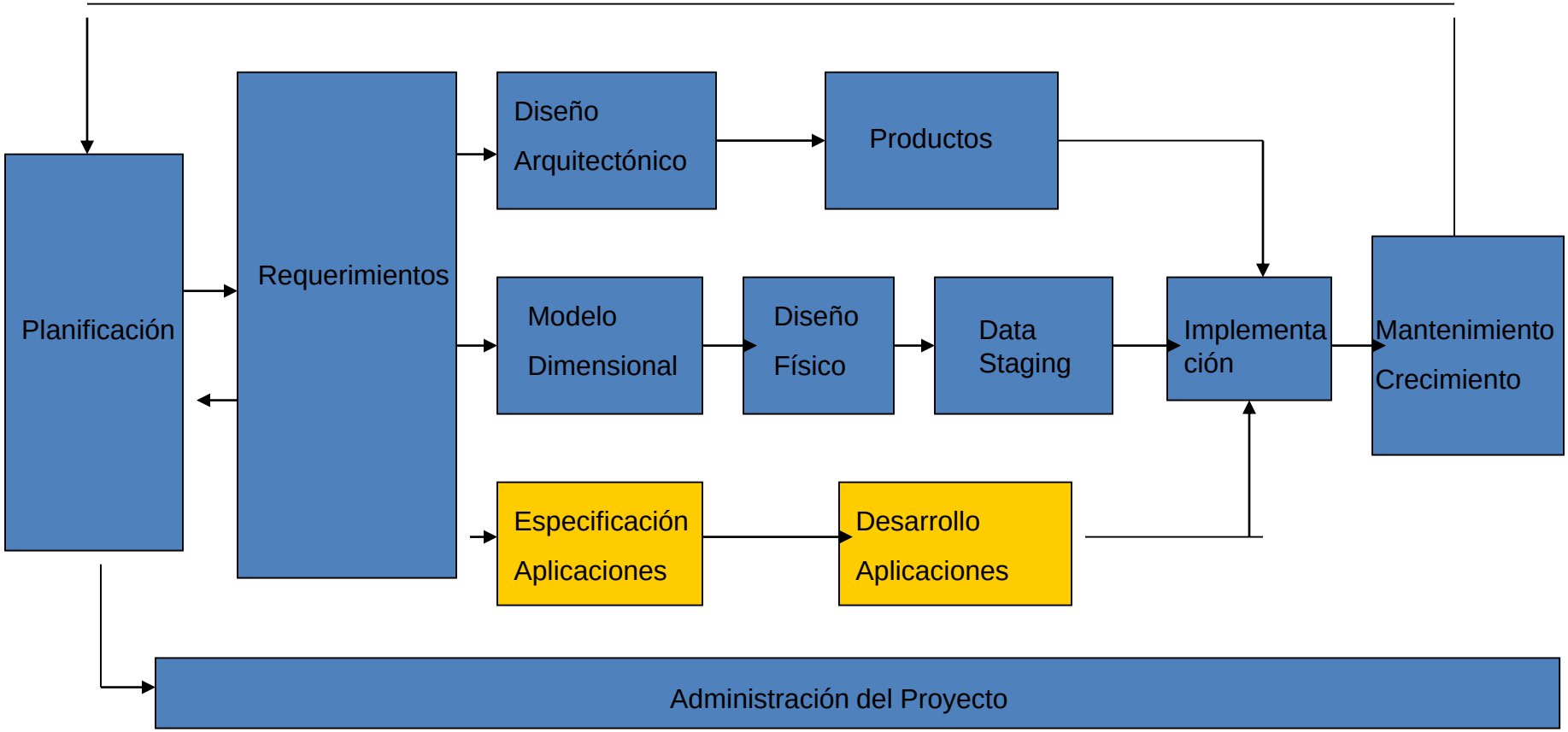
Línea de Datos



Línea de Datos

- Modelado dimensional
  - Tablas de hechos
  - Tablas de dimensión
  - Claves
- Diseño y desarrollo del ETL
  - Herramientas y técnicas
  - Organización de las tablas de dimensión
  - Organización de las tablas de hechos

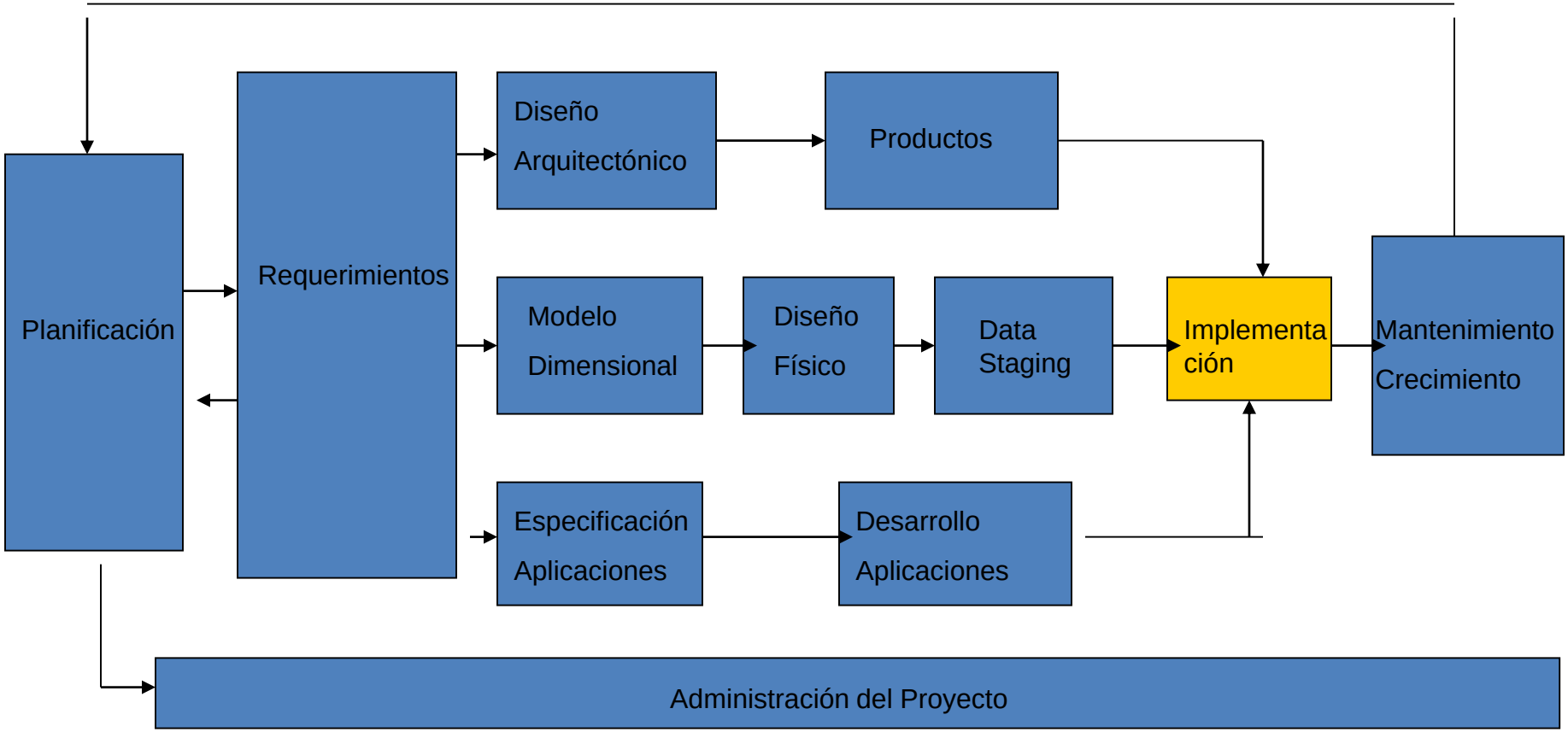
Línea de Aplicaciones



Línea de Aplicaciones

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

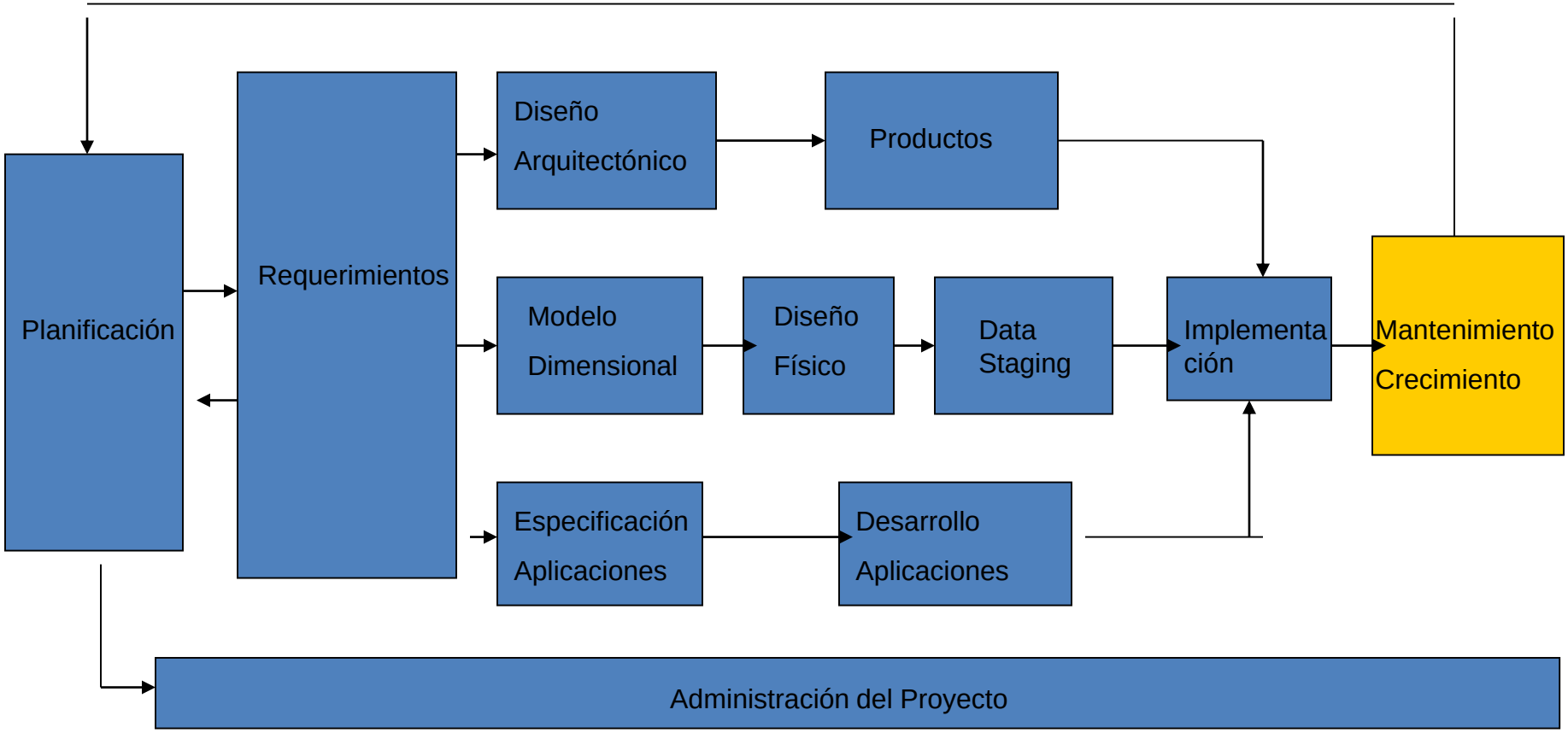
Implementación



Implementación

- Convergencia de las tres líneas
- La línea de datos es la que debería durar más tiempo
- 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**

UNIVERSIDAD
DE LOS ANDES
MERIDA VENEZUELA

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:

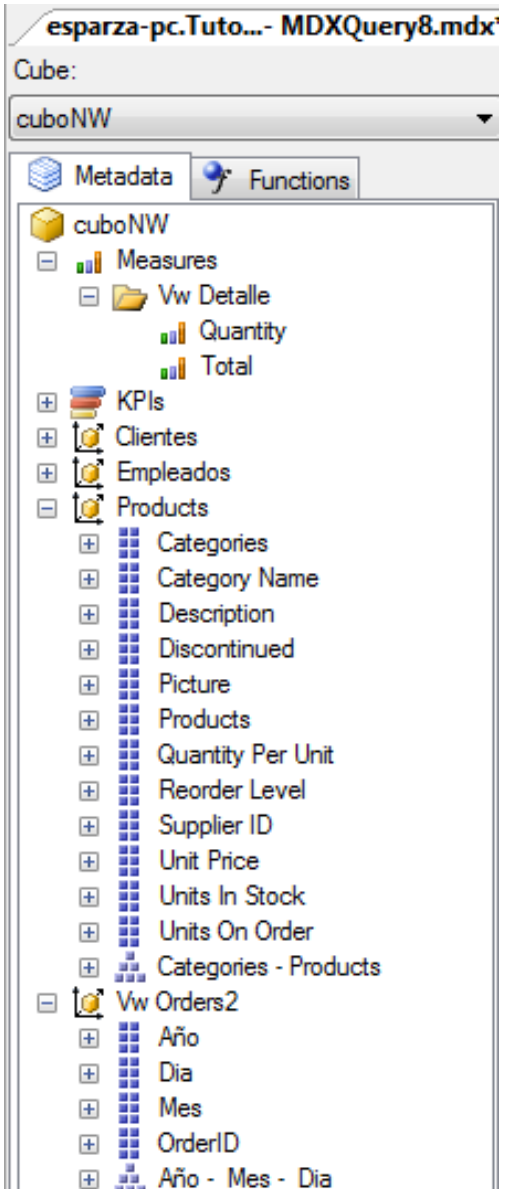
Jerarquias:Año-Mes-Dia

- 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 cantidad para todos los productos

Messages | Results

	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]
```

Messages | 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 uno 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)

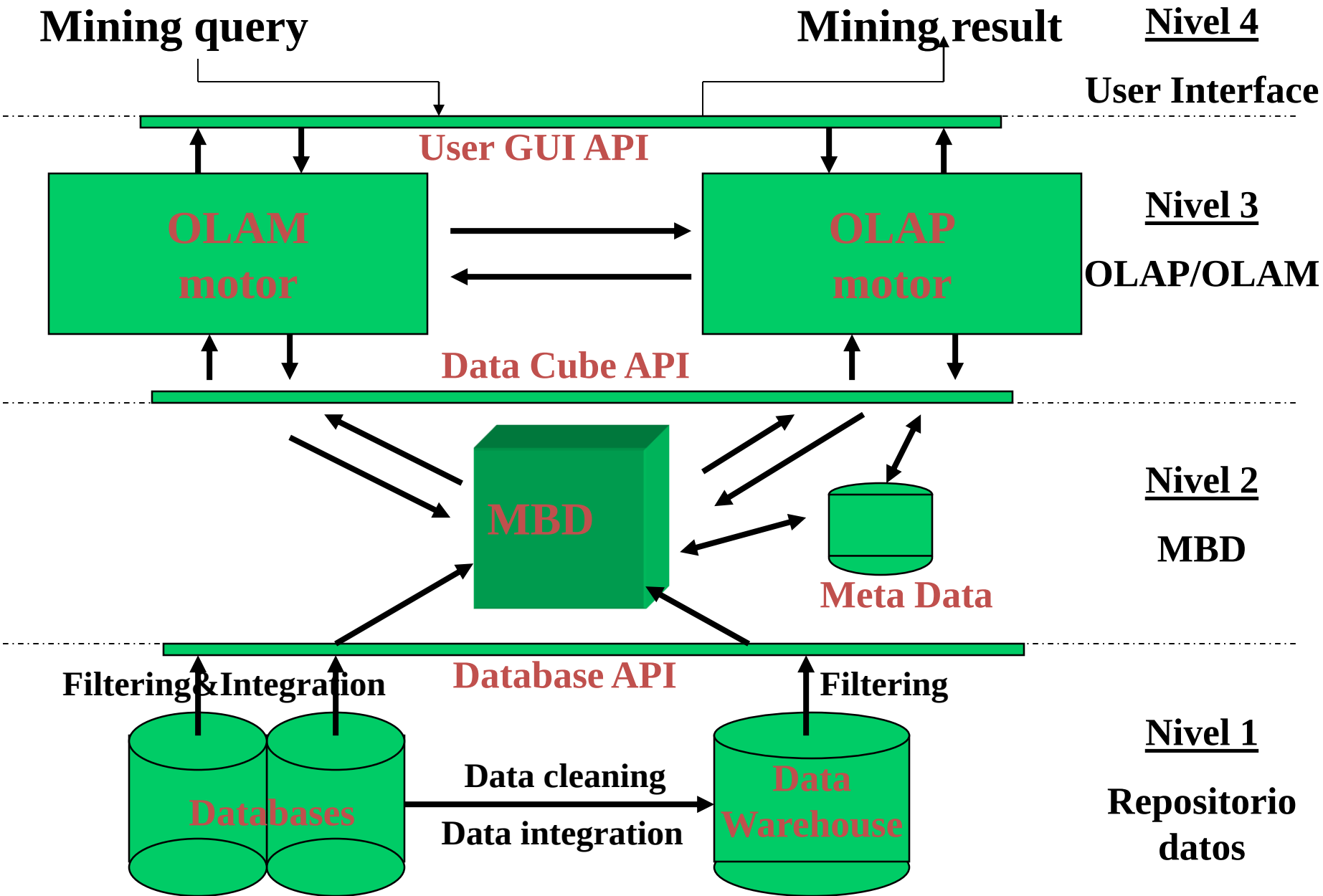
	All	1996	1997	1998
All	1354458.59	226298.5	658388.75	469771....
Beverages	286526.95	53879.2	110424	122223....
Condiments	113694.75	19458.3	59679	34557.45
Confections	177099.1	31511.6	87227.77	58359.73
Dairy Products	251330.5	44615.8	123910.8	82803.9
Grains/Cereals	100726.8	9817.6	60486.95	30422.25
Meat/Poultry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

De procesamiento analítico en línea a Minería análitica en línea (OLAM)

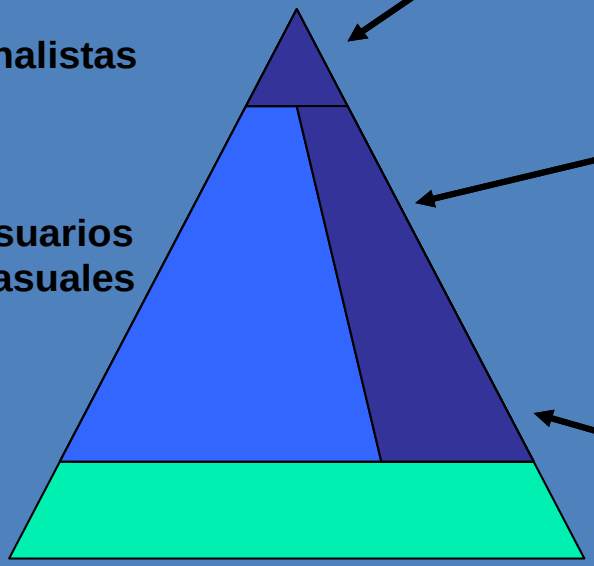
¿Por qué la minería analítico en línea?

- Alta calidad de los datos en data warehouses
  - DW contiene, datos integrados, coherentes, limpios
- Estructura de procesamiento de la información disponible
  - Herramientas presentación de informes y OLAP , etc.
- Análisis exploratorio de datos basados en OLAP
  - Minería con drilling, dicing, pivoting, etc.
- Selección on-line de las funciones de minería de datos
  - Integración e intercambio de múltiples funciones, algoritmos y tareas.

Arquitectura OLAM



Modelos Perfilados de Visualización



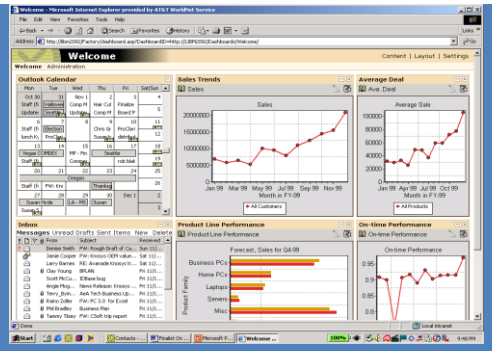
Analistas

Usuarios Casuales

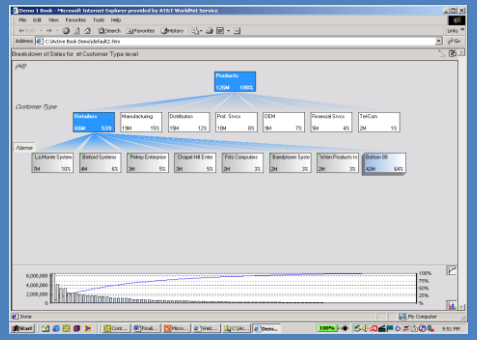
Análisis Estratégico

Consultas Analíticas

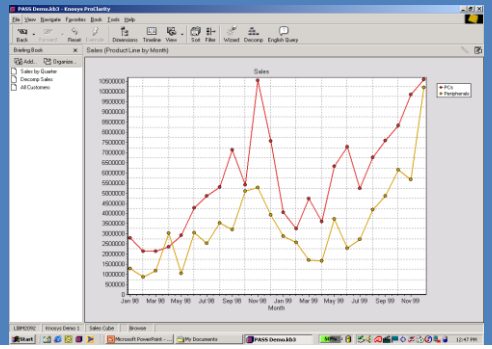
Reportes Interactivos



Tablero de Control



Herramienta de Consulta



Reportes Dinámicos

PENTAHO

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

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 Visualizacion
- 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

JetReports

ApeSoft

work meter
Buen trabajo

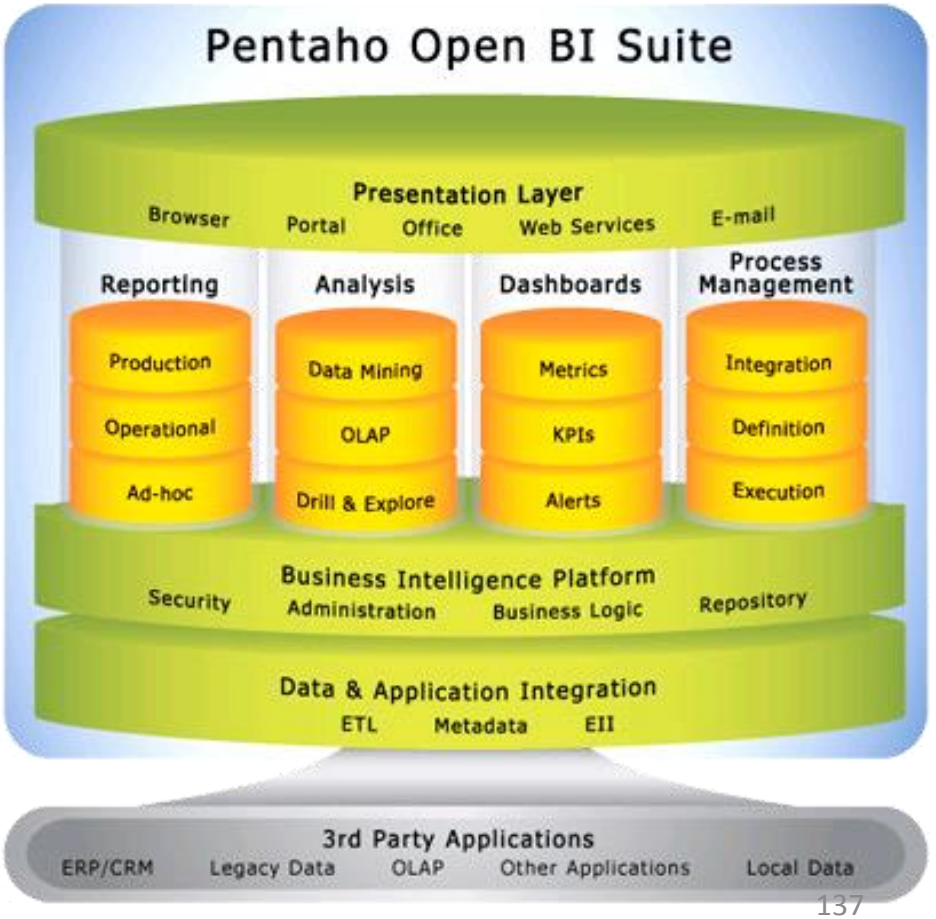
QlikView

Pentaho: La solucion Open Source Business Intelligence

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

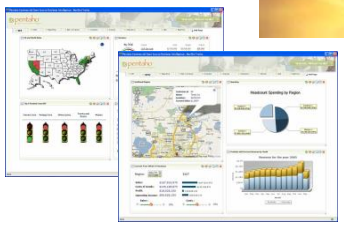
Pentaho incluye herramientas para hacer:

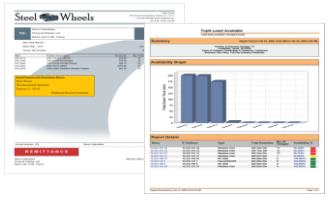
- Reportes
- Análisis
- Vizualización (Dahsboards)
- Manejo de Datos

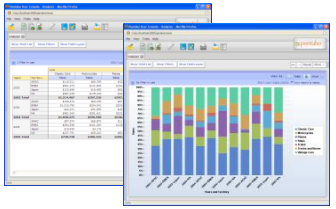


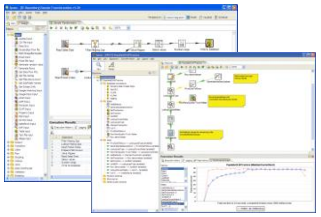
Componentes independientes

- **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







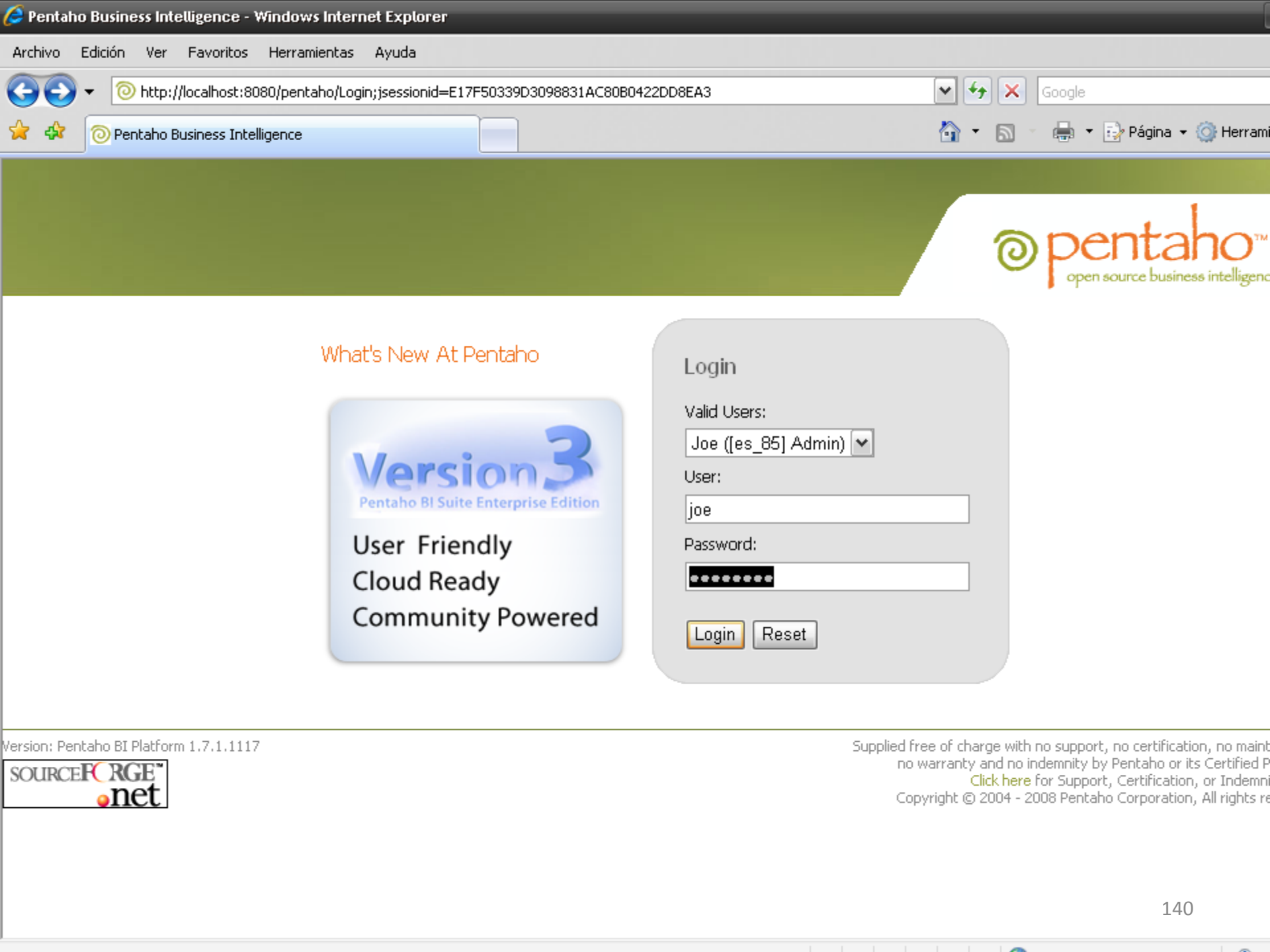


PENTAHO

- 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.

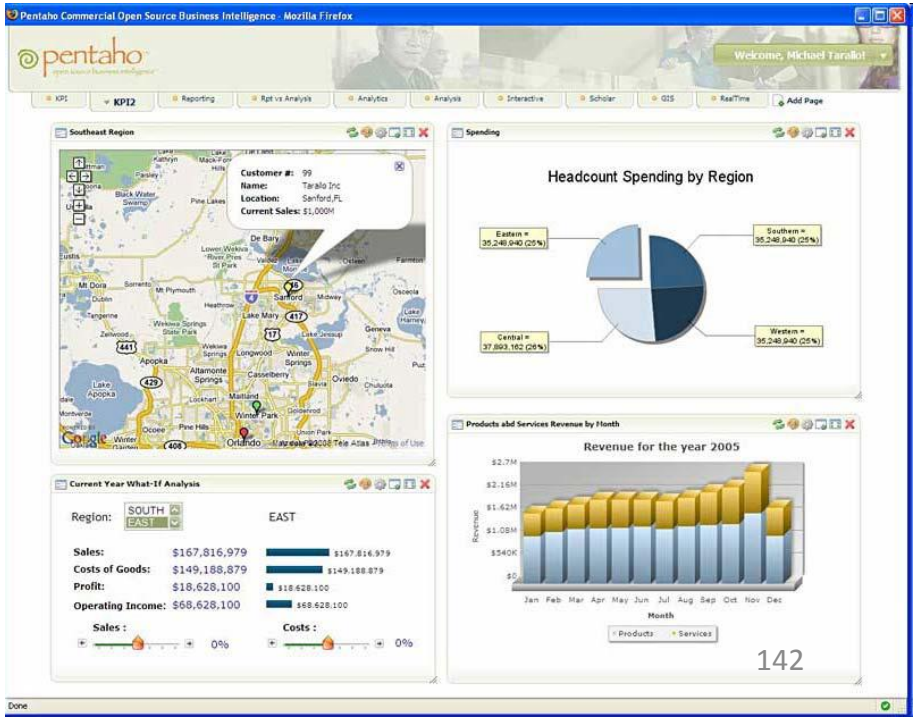

Pentaho Business Intelligence - Windows Internet Explorer
Archivo Edición Ver Favoritos Herramientas Ayuda
http://localhost:8080/pentaho/Login;jsessionid=E17F50339D3098831AC80B0422DD8EA3
Pentaho Business Intelligence
pentaho
open source business intelligenc
What's New At Pentaho
Version 3
Pentaho BI Suite Enterprise Edition
User Friendly
Cloud Ready
Community Powered
Login
Valid Users:
Joe ([es_85] Admin)
User:
joe
Password:
Login
Reset
Version: Pentaho BI Platform 1.7.1.1117
SOURCEFORGE.net
Supplied free of charge with no support, no certification, no maint
no warranty and no indemnity by Pentaho or its Certified P
Click here for Support, Certification, or Indemni
Copyright © 2004 - 2008 Pentaho Corporation, All rights re

Pentaho Dashboard

- Brinda a los usuarios la información crítica que necesitan para comprender y mejorar el desempeño organizacional.

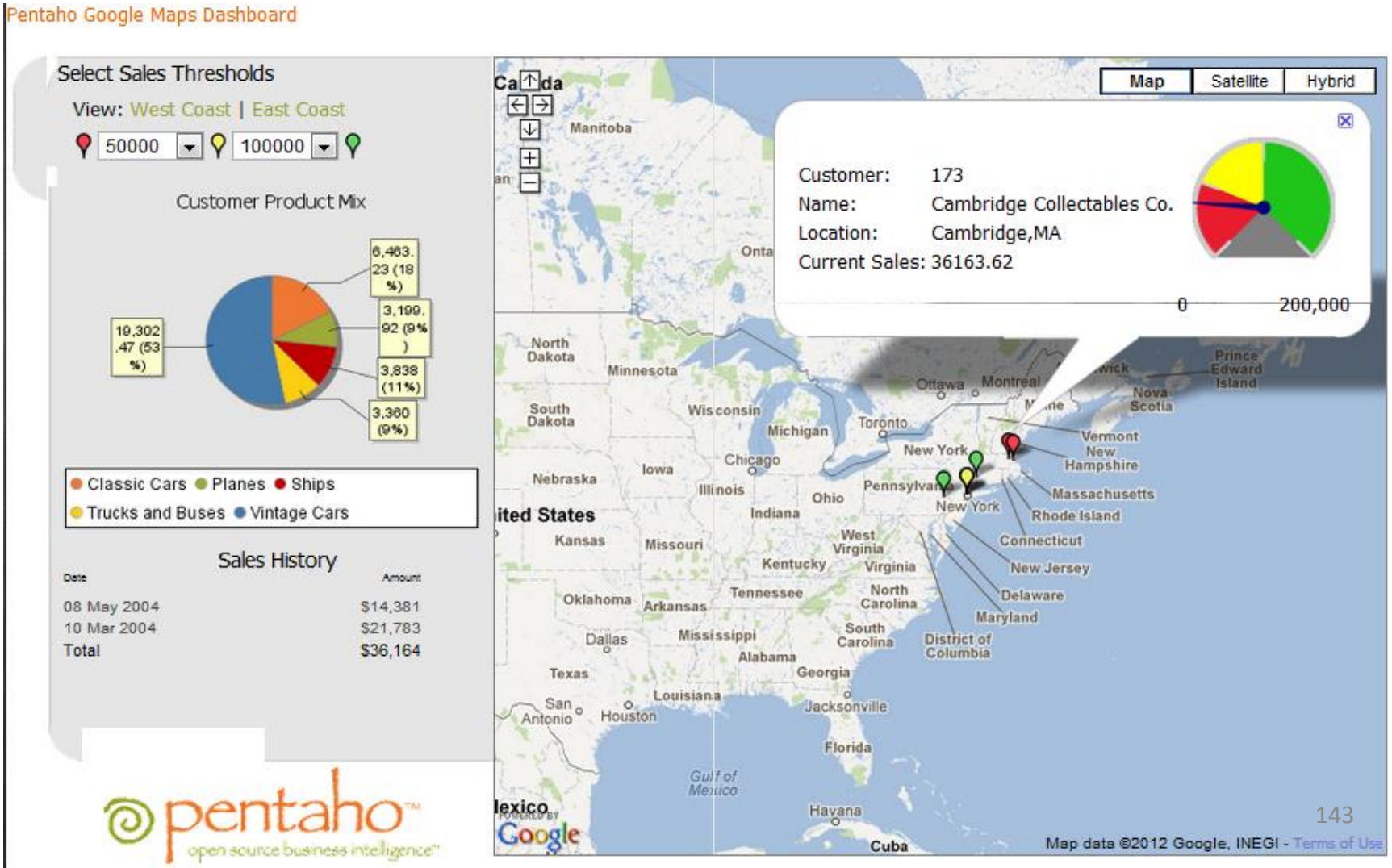
- Proporciona una visión inmediata del rendimiento individual, departamental, o de la organización, mediante la entrega de métricas claves a través de una interfaz visual atractiva e intuitiva.

Pentaho Dashboard



Interactivo

Pentaho Dashboard: Geo Location



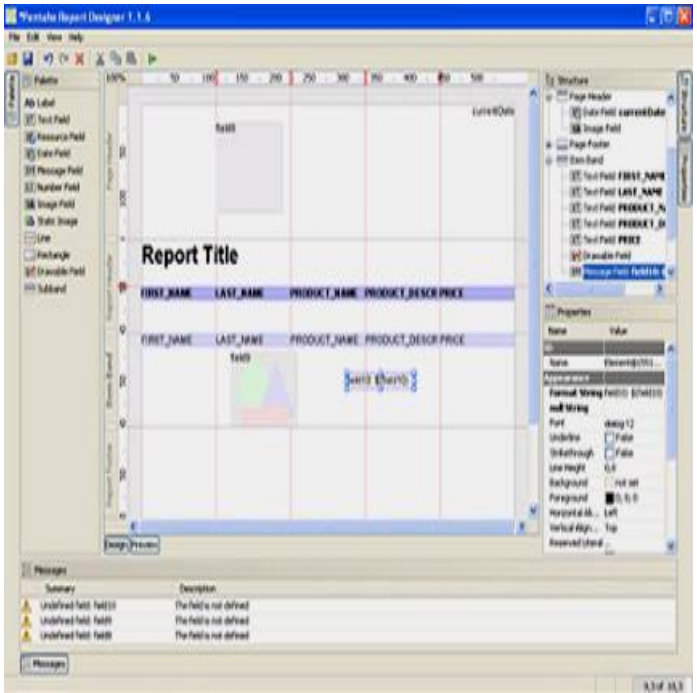
Reporting

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 Assisitant	$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	$574,895	$570,300	-$4,595
	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

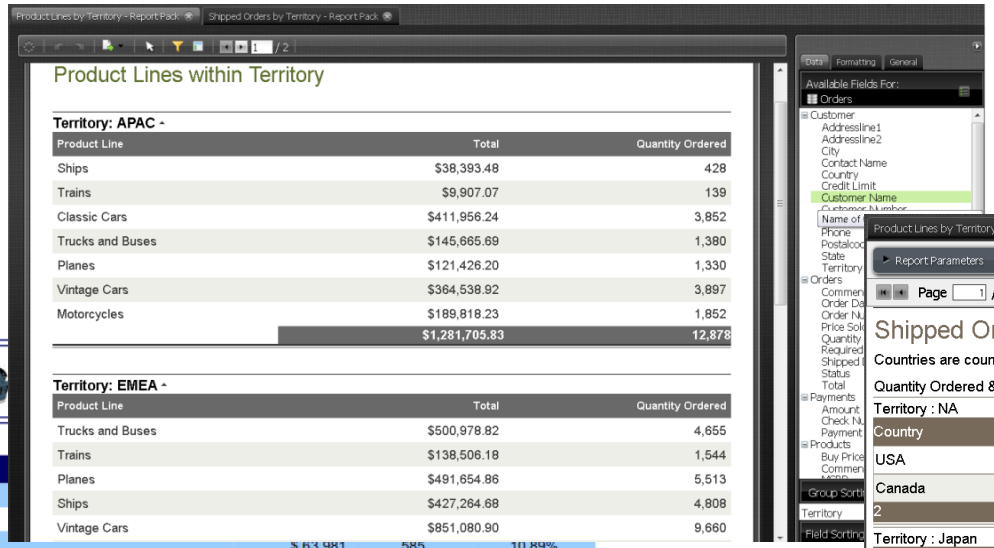
La solución de Reporting que plantea Pentaho, incluida dentro de su suite, es

Jfree Report.

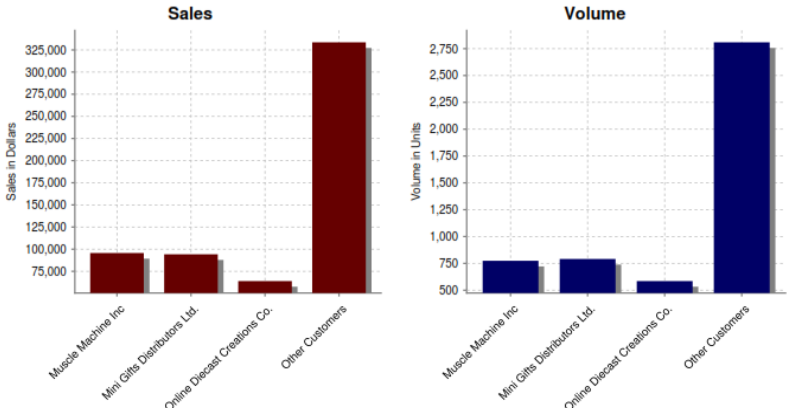


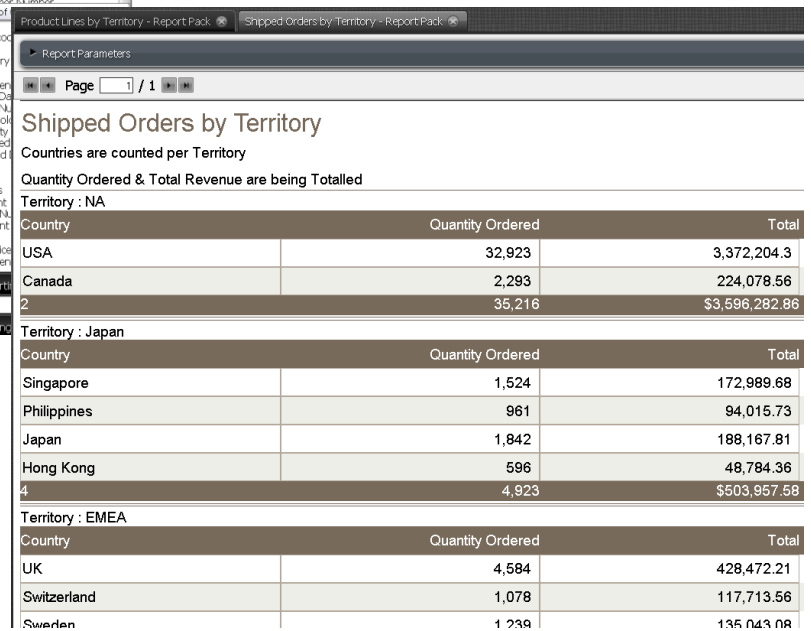
Reporting

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



Customer			
Muscle Machine Inc			
Mini Gifts Distributors Ltd.			
Online Diecast Creations Co.	$ 63,981	585	10.89%
Other Customers	$ 333,579	2,809	56.79%
Total	**$ 587,428**	**4,959**	





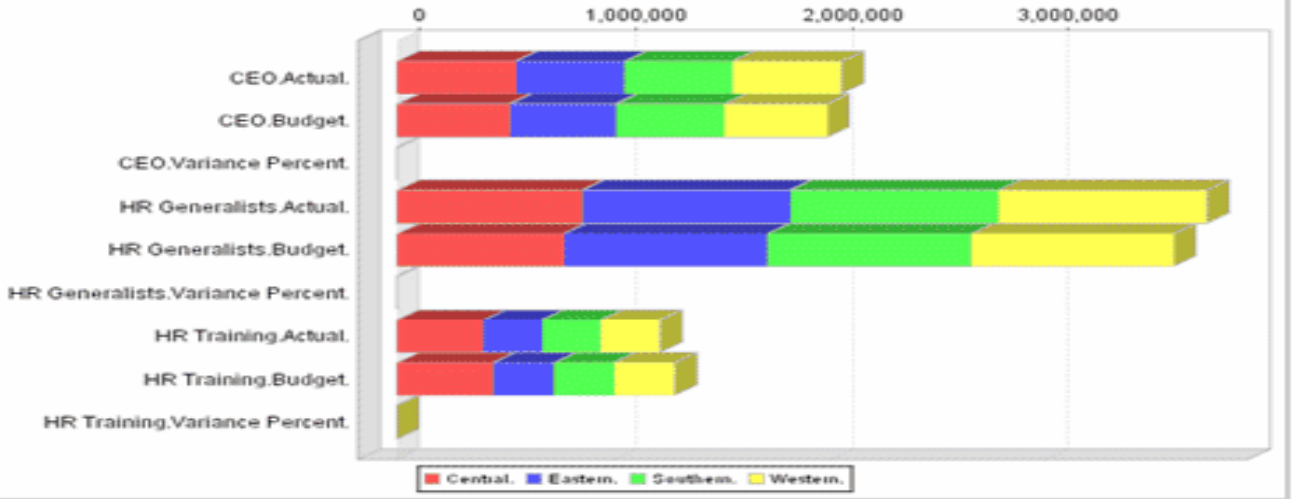
Mondrian permite

- 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]



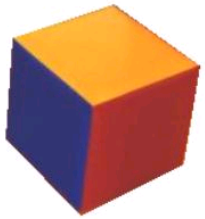
OLAP en Pentaho

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**

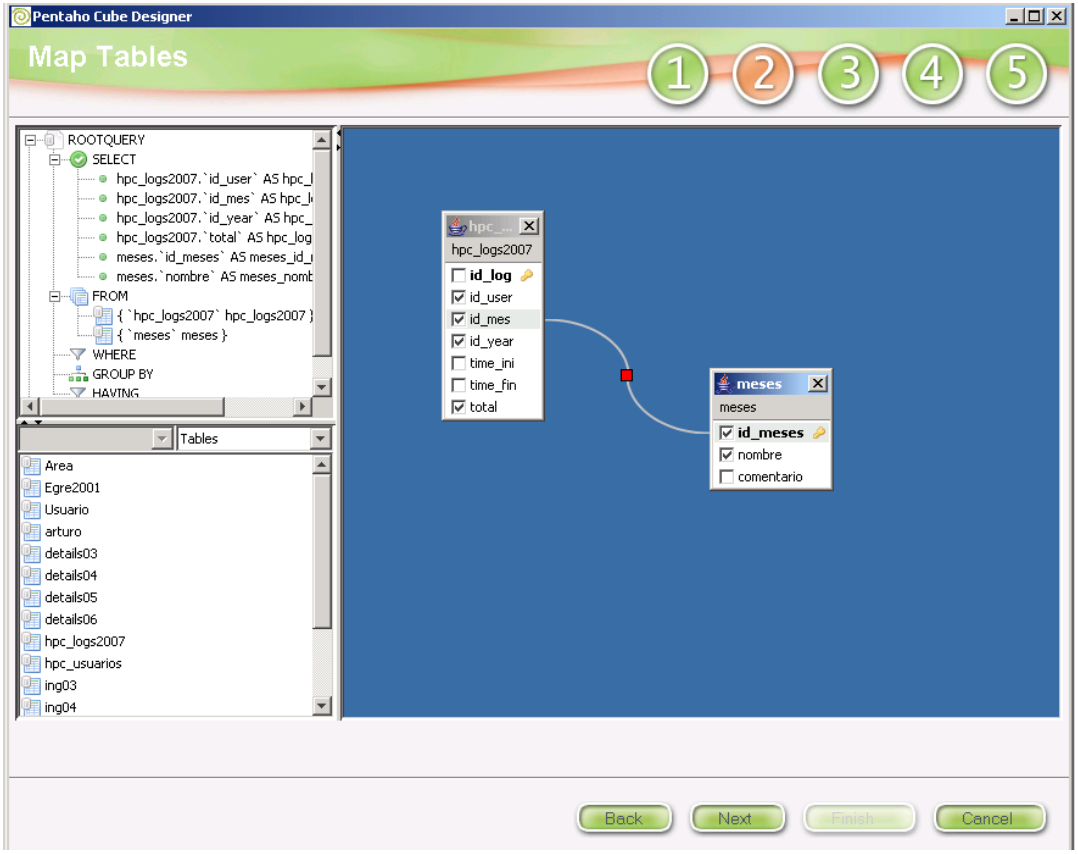
OLAP en Pentaho



Pentaho OLAP

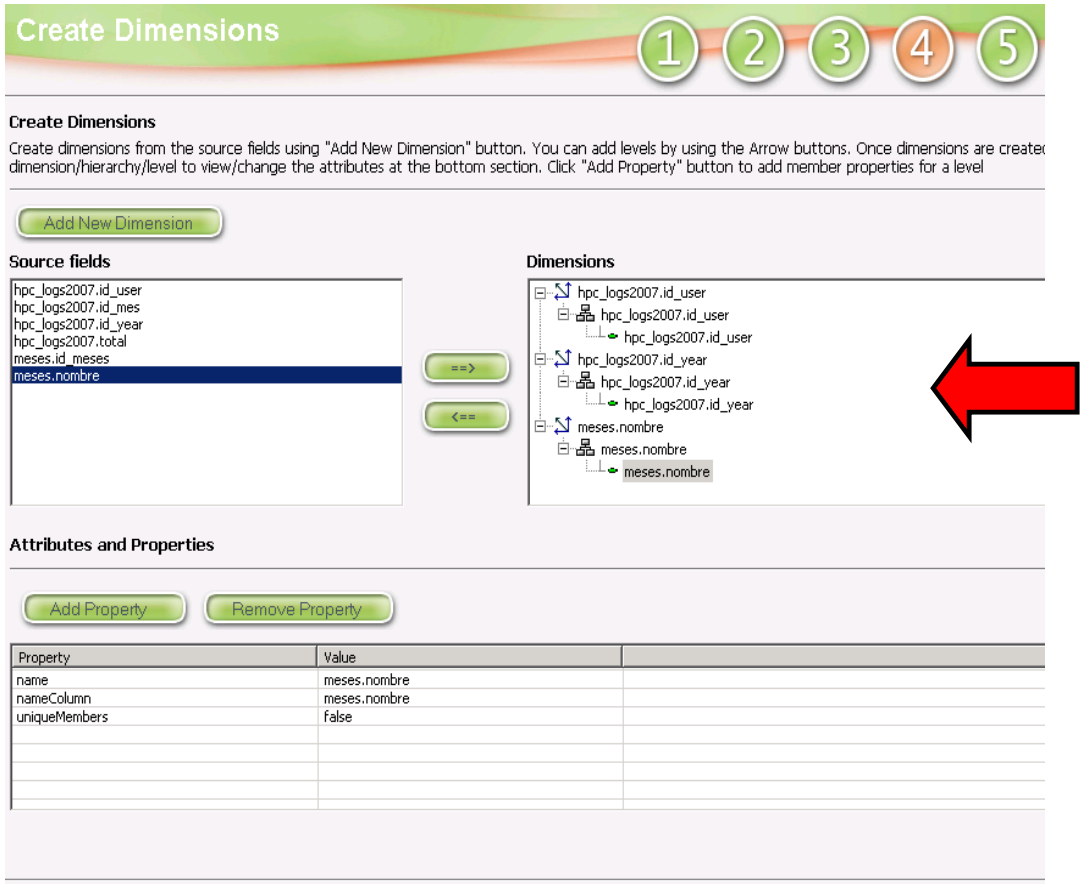
Permite diseñar cubos dinámicos, para un análisis de la información mucho más rápida y detallada, que apoya la Toma de Decisiones dentro de una organización

OLAP en Pentaho



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

OLAP en Pentaho



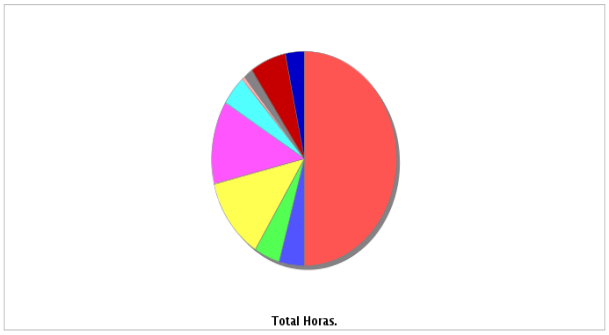
Crea las dimensiones del Cubo que luego seran mostradas en la Suite de Pentaho

Gráficas de Resultados

HPC

		Measures
meses.id_meses	hpc_logs2007.id_user	Total Horas
-All meses.id_meses	+All hpc_logs2007.id_user	1335
+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:



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

HPC

		Measures
meses.id_meses	hpc_logs2007.id_user	Total Horas
-All meses.id_meses	+All hpc_logs2007.id_user	1335
+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

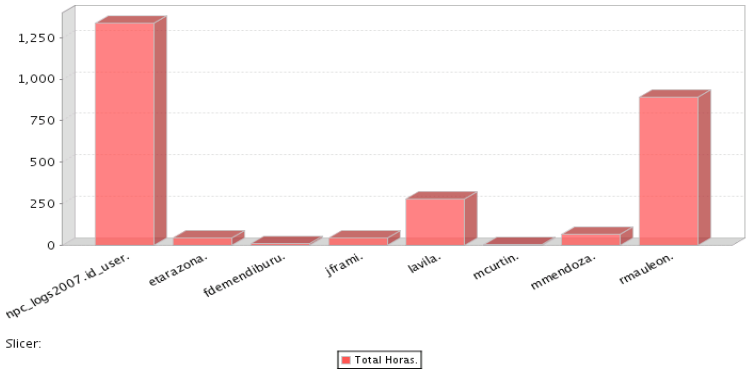
Slicer:

Gráficas de Resultados

HPC

	Measures
hpc_logs2007.id_user	Total Horas
−All hpc_logs2007.id_user	1335
etarazona	45
fdemendiburu	8
jframi	43
lavila	281
mcurtin	2
mmendoza	64
rmauleon	893

Slicer:



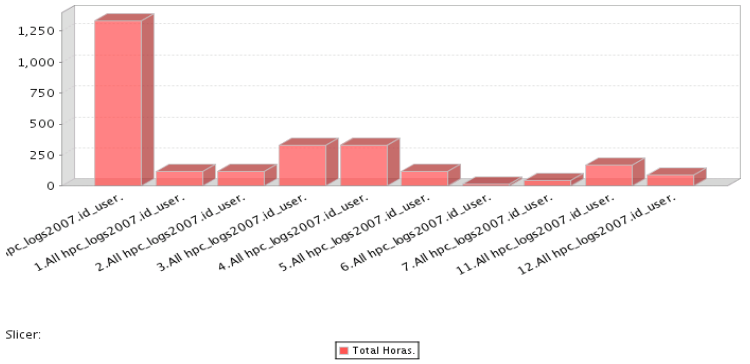
Slicer:

Estadistica por Usuario

HPC

		Measures
meses.id_meses	hpc_logs2007.id_user	Total Horas
−All meses.id_meses	+All hpc_logs2007.id_user	1335
+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:

Estadistica por Mes

pentaho™
open source business intelligence™

Kettle

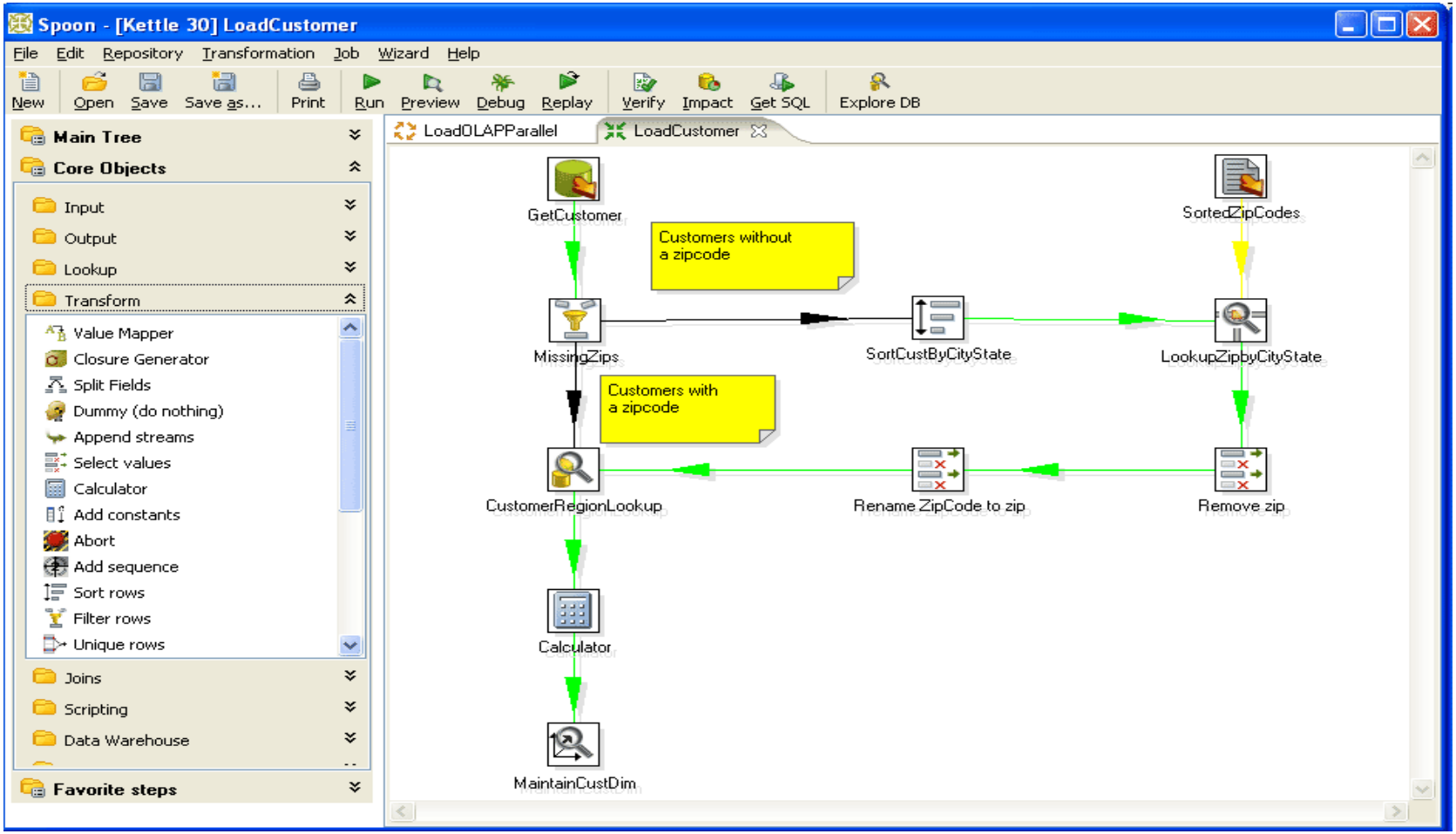
- tiene capacidades de accesar, limpiar e integrar datos desde cualquier lugar de la organización.

- ofrece una poderosa mecanismo de extracción, transformación y carga (ETL), con un diseño intuitivo y una arquitectura escalable basada en estándares probados.

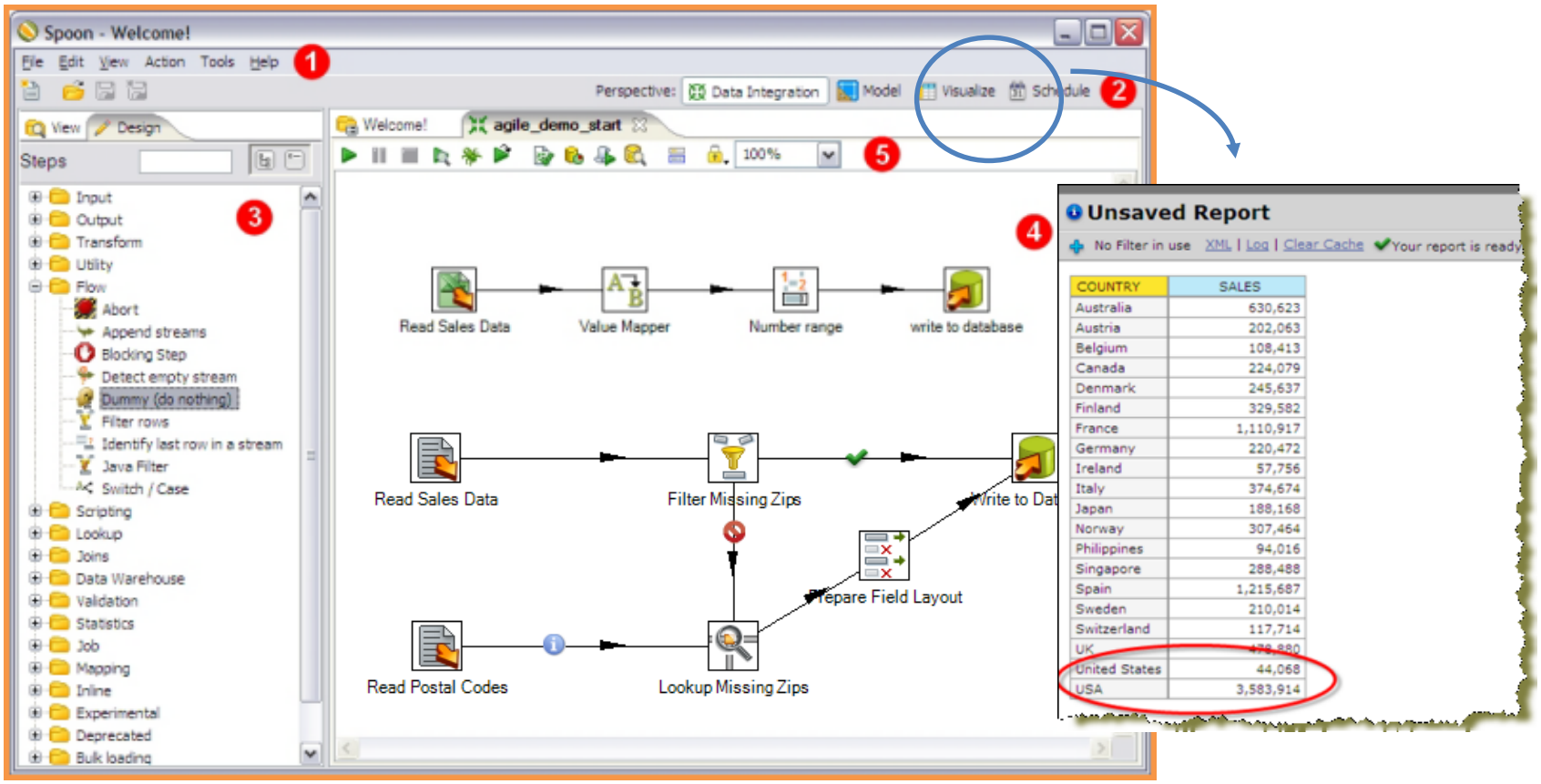
Suite

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

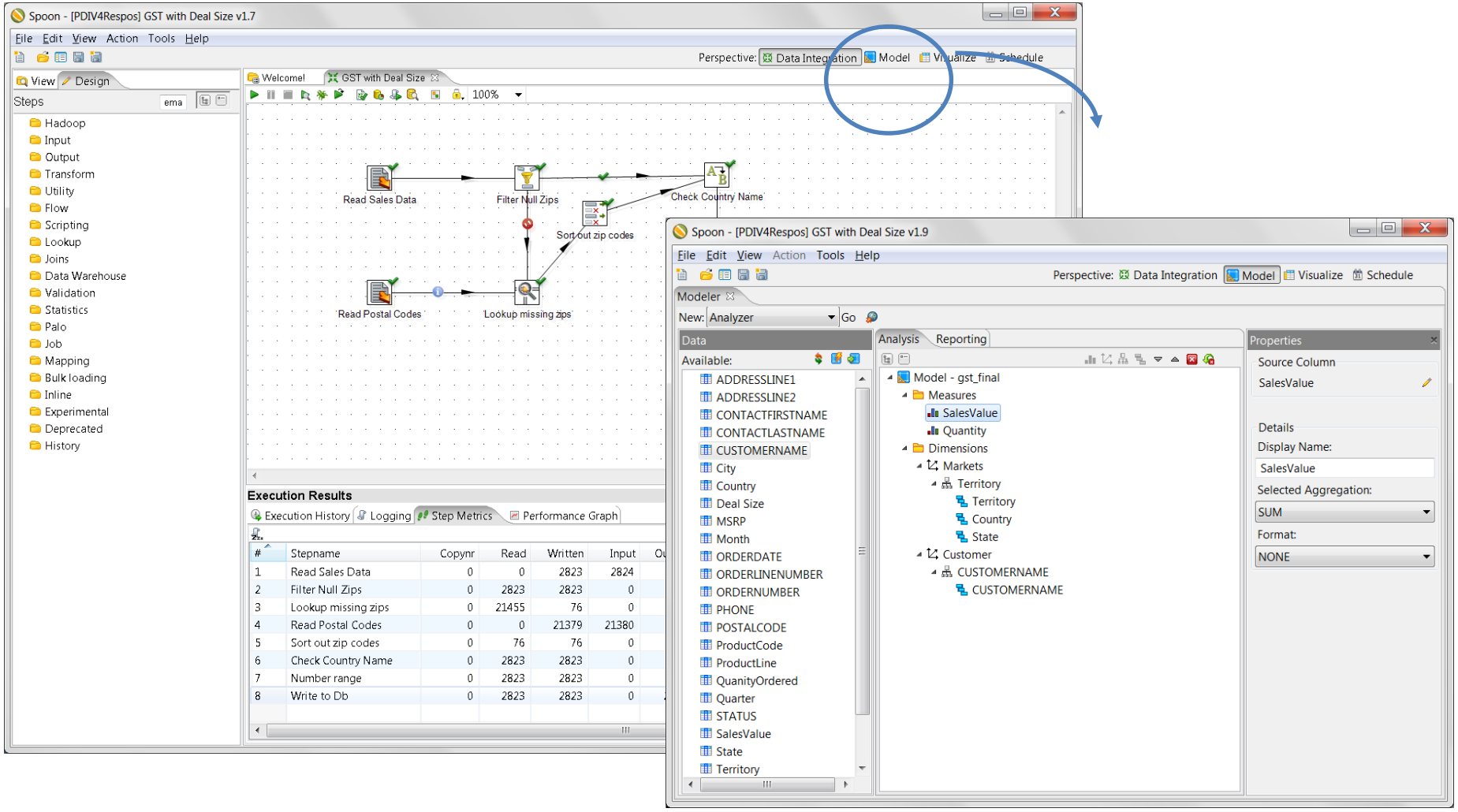
Kettle


Spoon - [Kettle 30] LoadCustomer
File Edit Repository Transformation Job Wizard Help
New Open Save Save as... Print Run Preview Debug Replay Verify Impact Get SQL Explore DB
Main Tree
Core Objects
Input
Output
Lookup
Transform
Value Mapper
Closure Generator
Split Fields
Dummy (do nothing)
Append streams
Select values
Calculator
Add constants
Abort
Add sequence
Sort rows
Filter rows
Unique rows
Joins
Scripting
Data Warehouse
Favorite steps
LoadOLAPParallel
LoadCustomer
GetCustomer
Customers without a zipcode
SortedZipCodes
MissingZips
SortCustByCityState
LookupZipbyCityState
Customers with a zipcode
CustomerRegionLookup
Rename ZipCode to zip
Remove zip
Calculator
MaintainCustDim

Integración de datos



Integración de datos: crear metadatos



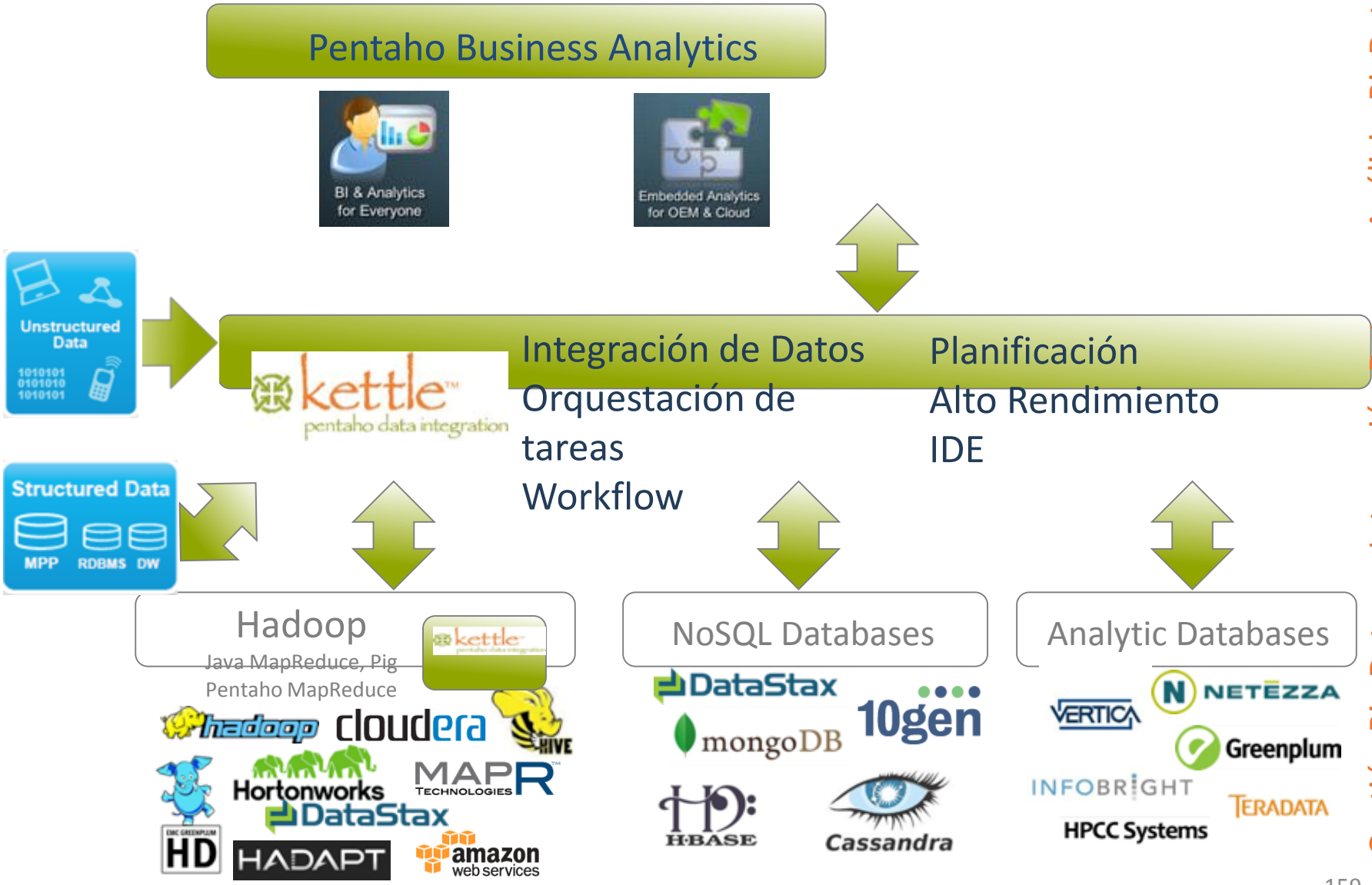
pentaho™
open source business intelligence™

Weka

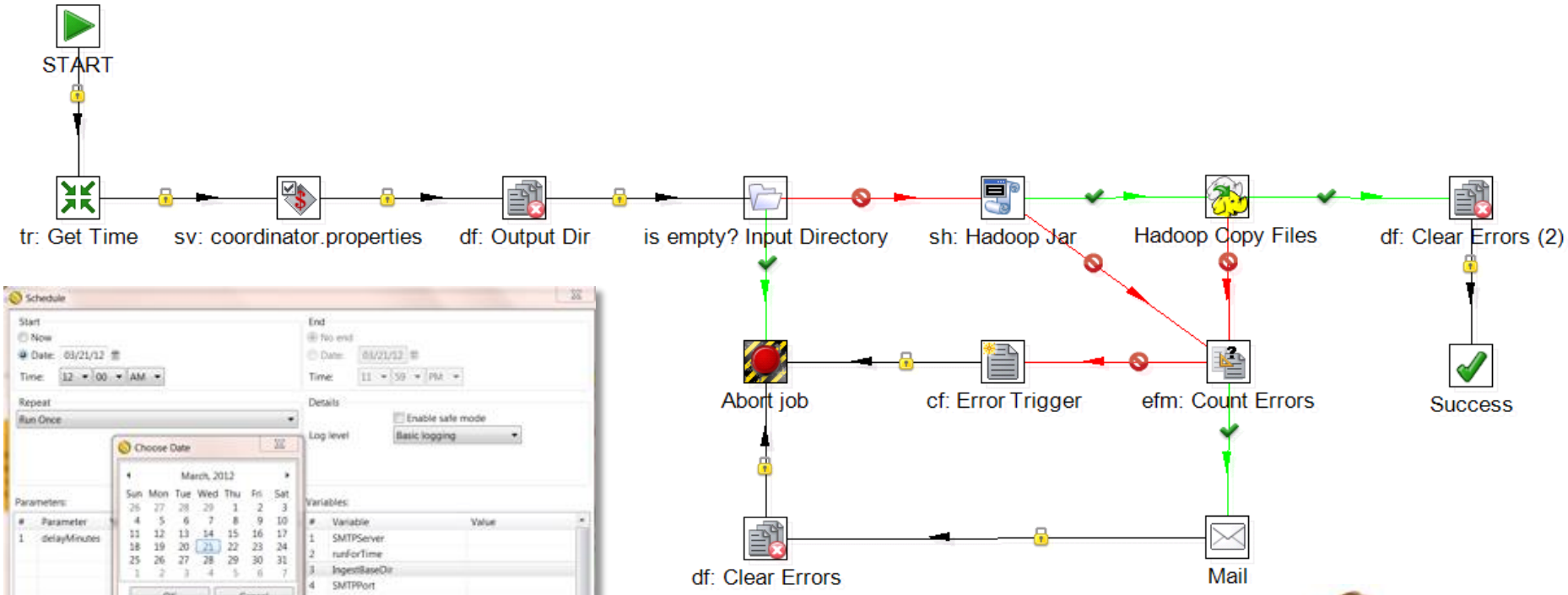
- Descubrir relaciones ocultas en los datos y descubrir indicadores de desempeño futuro.

- Análisis predictivo y exploración de correlaciones en los datos para mejorar el desempeño organizacional

Pentaho y Big Data



Orquestación virtual de tareas con diferentes fuentes de datos



Planificación