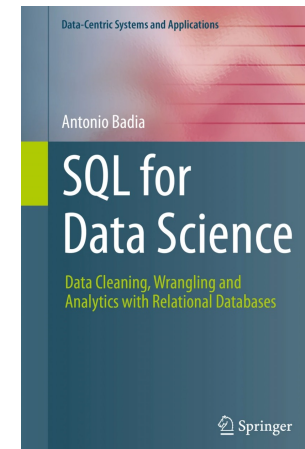
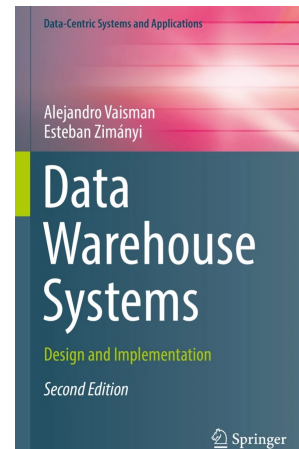
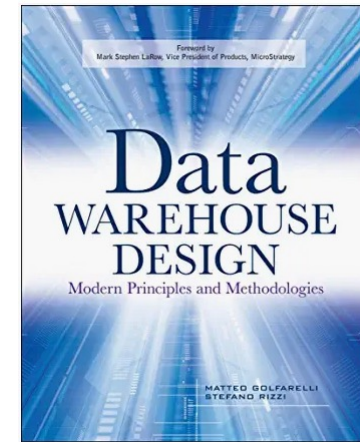
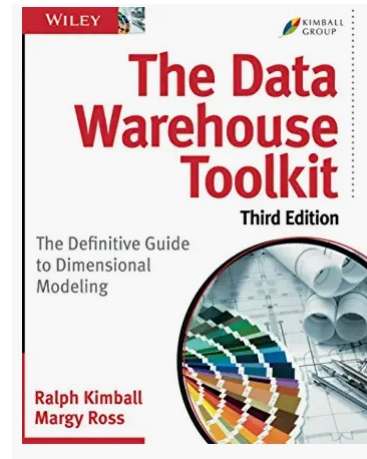
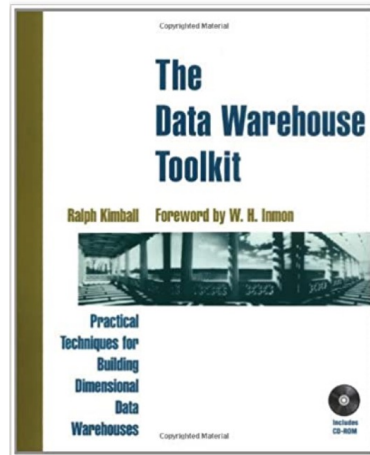


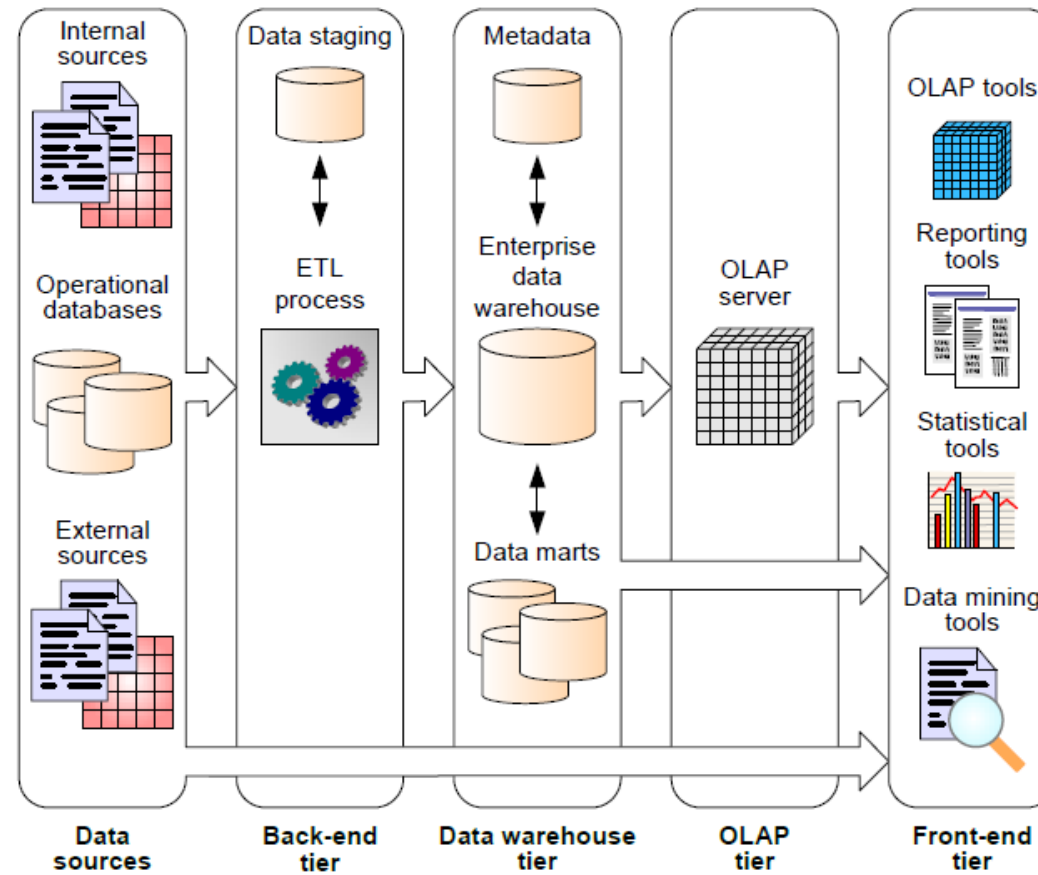
DW y el Modelo Multidimensional

Alejandro Vaisman
Instituto Tecnológico de Buenos Aires, Argentina
avaisman@itba.edu.ar

Bibliografía



Arquitectura de un DW



Backend

Extracción, Transformación y Carga (Loading)

- **Extracción:** toma datos de fuentes heterogéneas, internas y/o externas
- **Transformación:** modifica los datos desde el formato de las fuentes al formato del DW
 - Limpieza: elimina errores e inconsistencias, lleva los datos a un formato standard
 - Integración: Reconcilia datos de distintas fuentes, a nivel de esquema e instancia
 - Agregación: Sumariza los datos de acuerdo a la granularidad del
- **Carga:** Alimenta el DW con los datos transformados. Incluye el refresco periódico del DW
- Area de Data staging: BD donde se transforman los datos.

Capa de DW

- **Componentes**
 - DW organizacional
 - Data Marts
- **Metadatos**
 - De negocios: describen la semántica de los datos, reglas organizacionales, restricciones
 - Técnicos: Describen la estructura de los datosMetadata
- **Otros**
 - Datos de estructura del DW: facts, dimensiones, métricas, funciones de agregación
 - Datos físicos: índices, vistas materializadas
 - Estadísticas
 - Seguridad y acceso, etc.

Capa OLAP

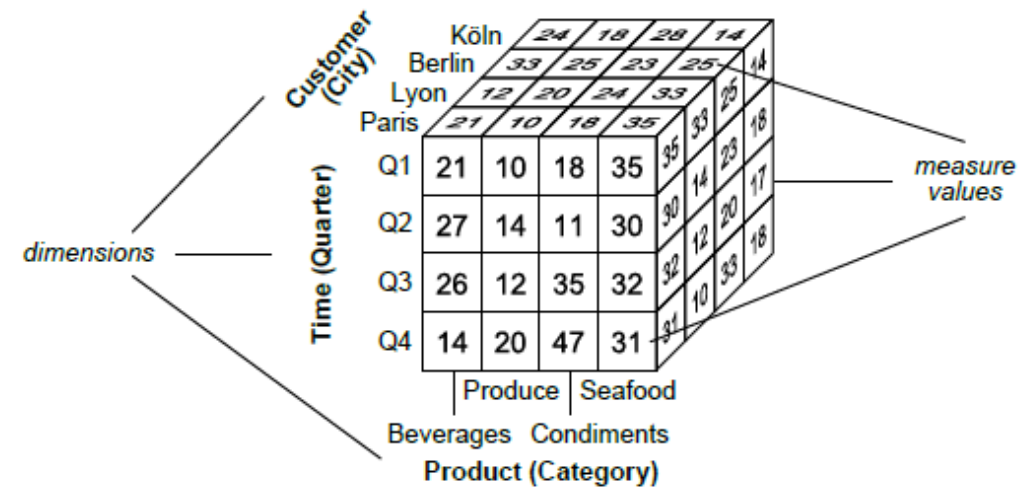
- Un servidor OLAP que expone los datos de los DWs y DMs en forma multidimensional
- Las aplicaciones perciben los datos como un cubo MOLAP
- Son consumidos por herramientas OLAP, y consultados por ejemplo, en el lenguaje MDX (MultiDimensional eXpressions)
- EL DW y los DMs pueden consultarse directamente sin pasar por la capa OLAP, en SQL y SQL extendido con funciones de ventana (window functions)

Capa Front-end

- Aplicaciones que consumen los datos
 - Clientes OLAP
 - Herramientas de data mining
 - Dashboards
 - Herramientas Estadísticas

El Modelo Multidimensional

- Percibe los datos como un espacio multidimensional (MD): el **data cube**
- Data cube: compuesto por **dimensions** y hechos (facts)
- **Dimensiones**: Perspectives utilizadas para analizar los datos
- **Atributos**: describen a las dimensiones
 - Por ejemplo, la **dimension Producto** puede tener atributos **ProductNumber** y **UnitPrice** (not se muestran en la figura)
- The **celdas** or **facts** tienen asociados valores numéricos llamados **métricas (measures)**
 - Ejemplo: **cantidad de unidades vendidas por categoría y ciudad del cliente**
- **Dimensiones: Product, Time, and Customer,**
- Métrica: **Quantity**



Introducción al modelo multidimensional

EJEMPLO

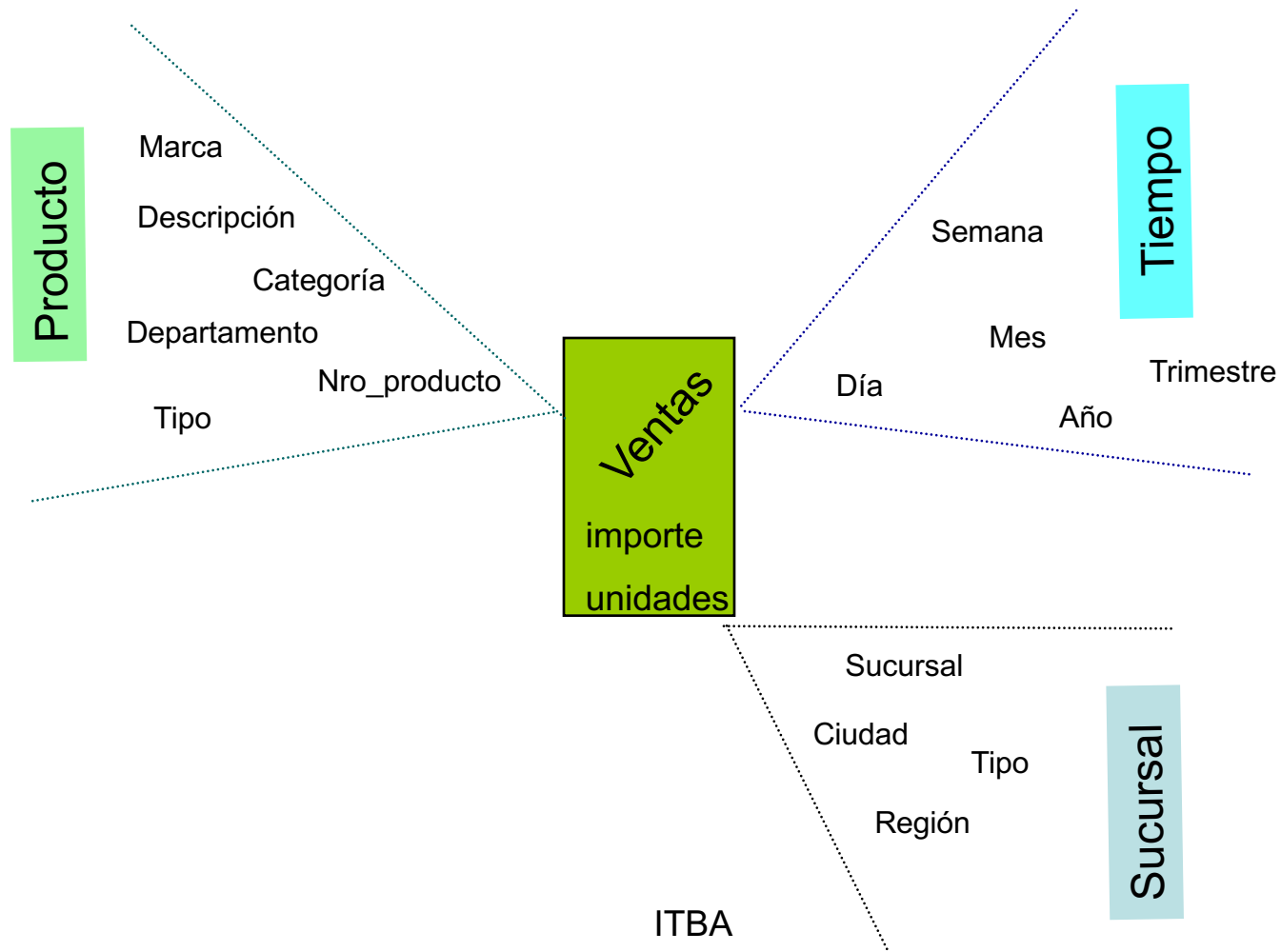
Organización: Cadena de supermercados.

Actividad objeto de análisis: ventas de productos.

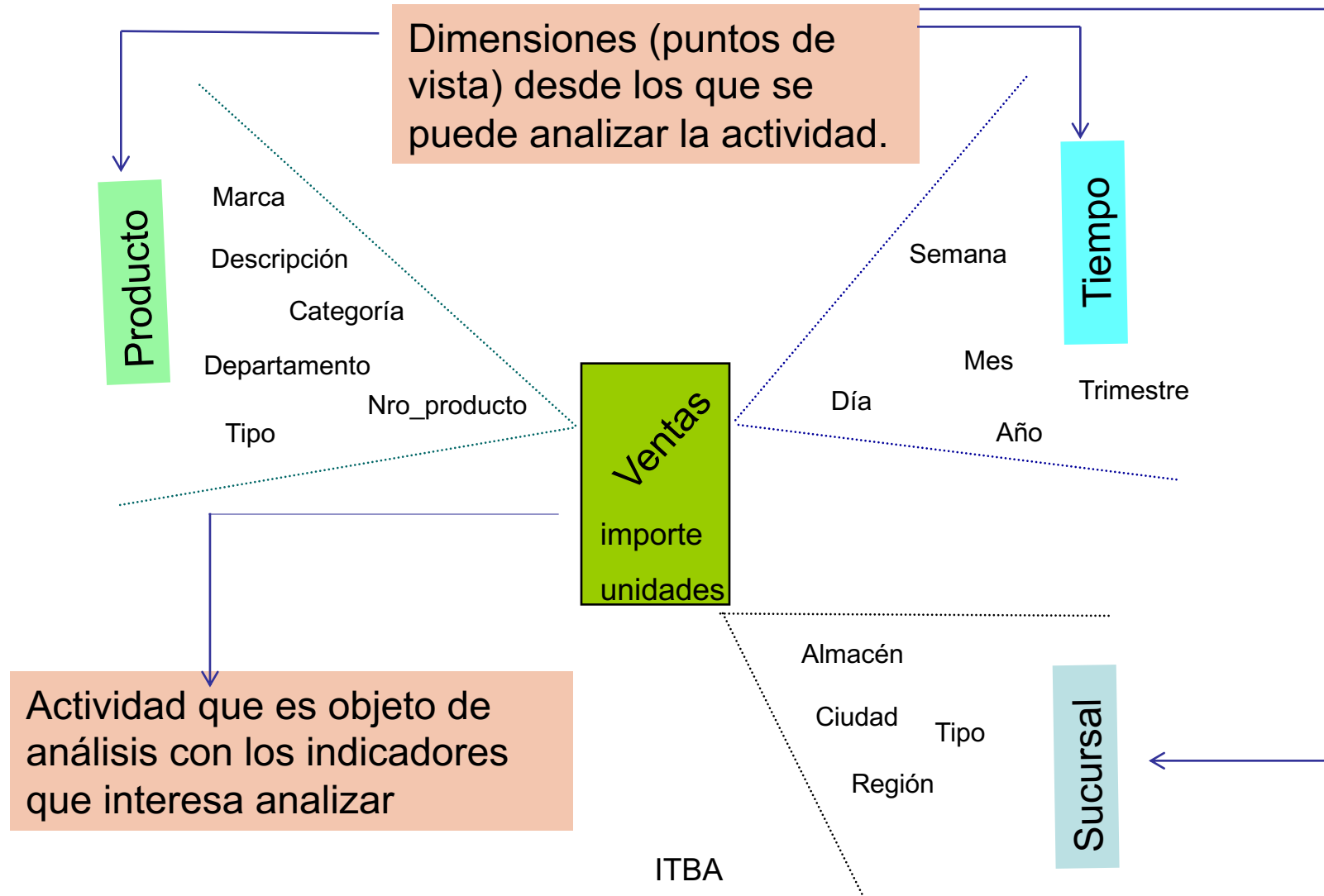
Fact de venta: “del **producto** “p1” se vendieron en la **sucursal** “Sucursal nro.1” el **día** 17/7/2003, 5 **unidades**, por un **importe** de 105 pesos”

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

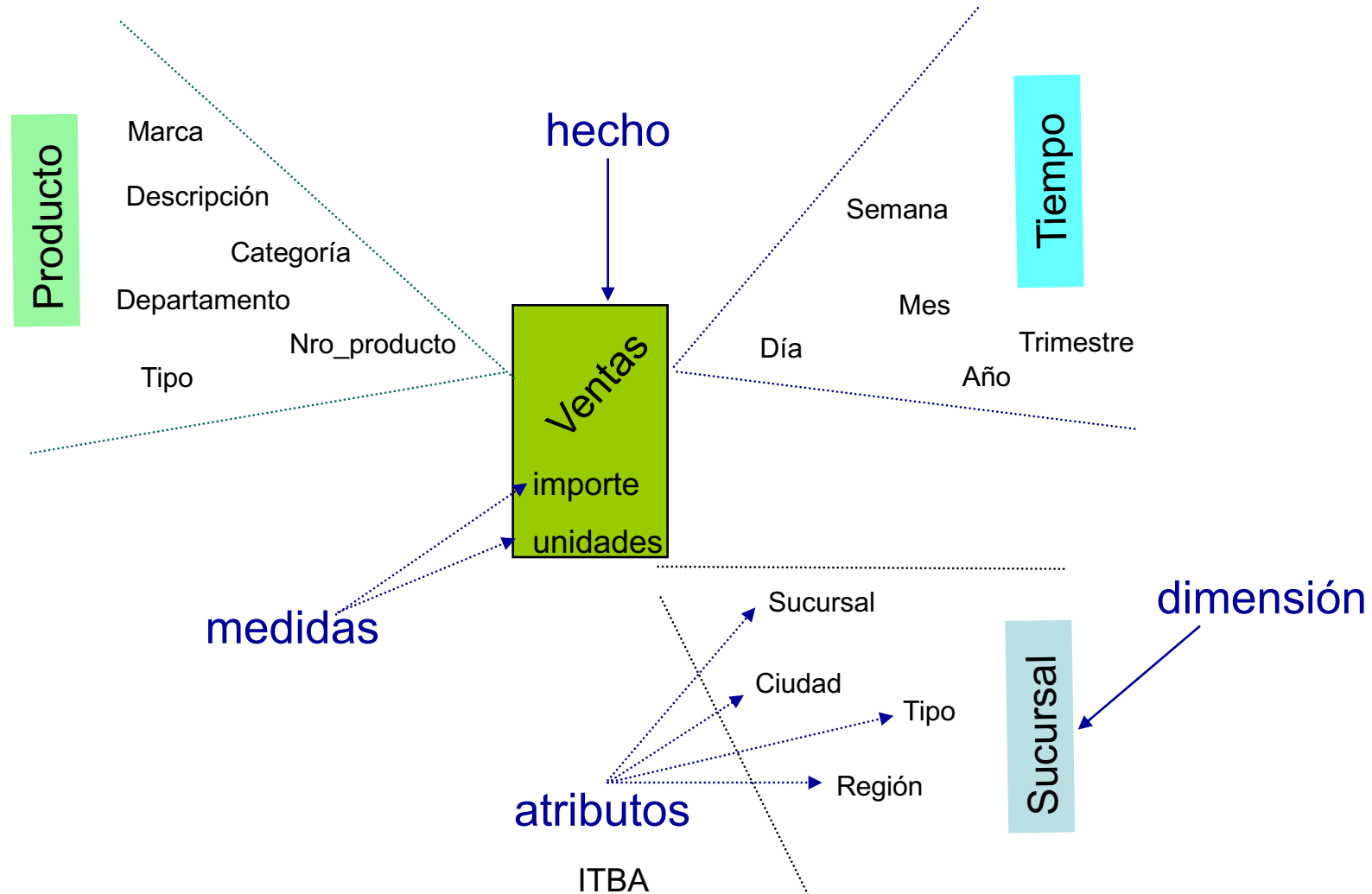
Hechos y dimensiones



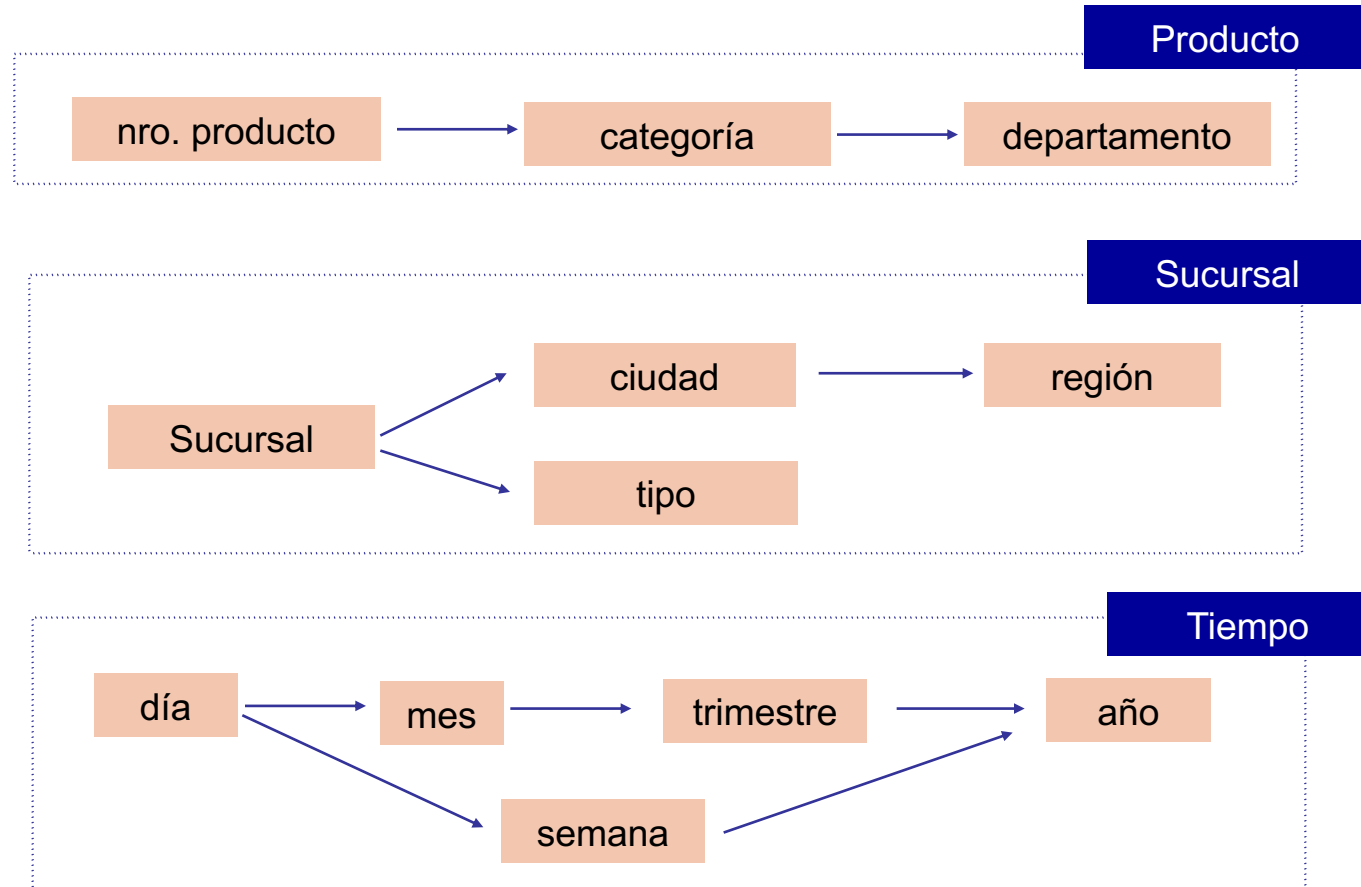
Hechos y dimensiones



Hechos, dimensiones, métricas

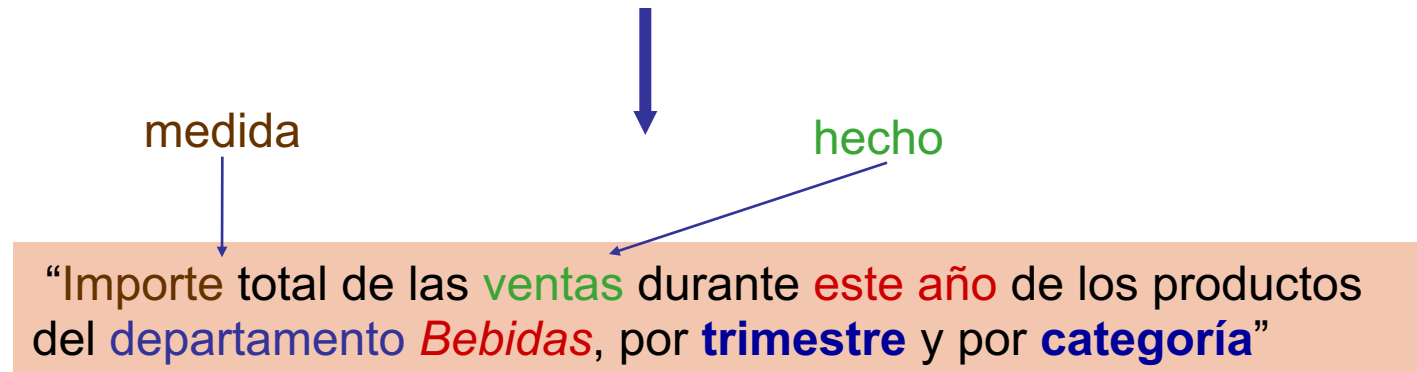


Jerarquías de agregación



Explotación del DW: OLAP

Una consulta a un DW consiste generalmente en la obtención de **medidas** sobre los **hechos** parametrizadas por atributos de las **dimensiones** y restringidas por **condiciones** impuestas sobre las dimensiones

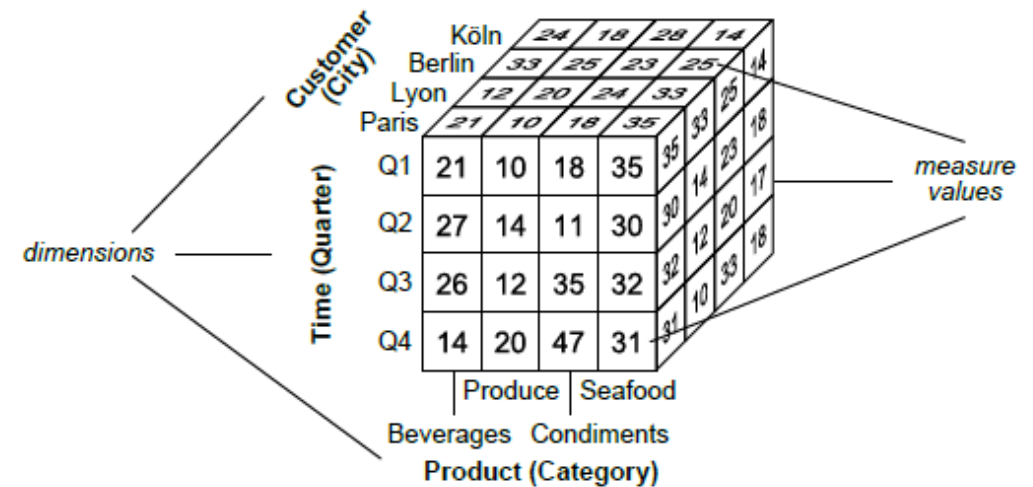


Restricciones: productos del departamento Bebidas, ventas durante este año

Parámetros de la consulta: por categoría de producto y por trimestre

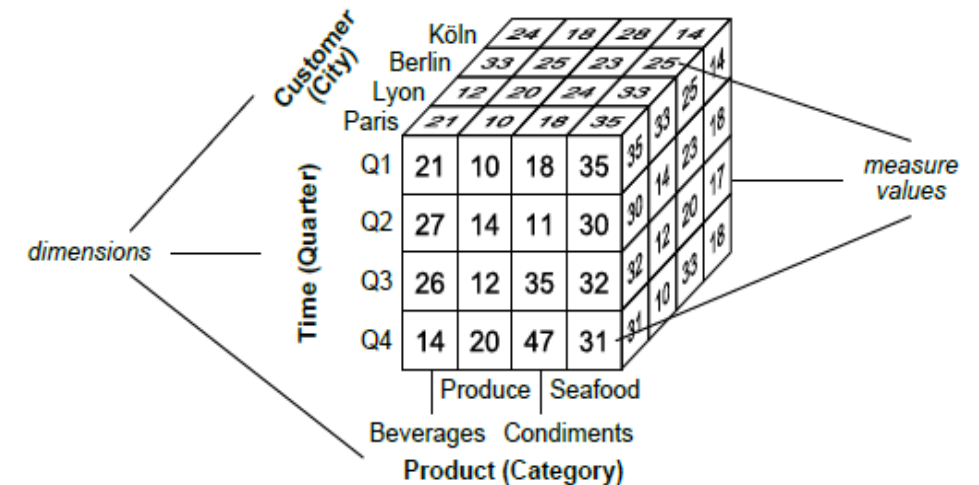
El Modelo Multidimensional

- Percibe los datos como un espacio multidimensional (MD): el **data cube**
- Data cube: compuesto por **dimensions** y hechos (facts)
- **Dimensiones**: Perspectives utilizadas para analizar los datos
- **Atributos**: describen a las dimensiones
 - Por ejemplo, la **dimension Producto** puede tener atributos **ProductNumber** y **UnitPrice** (not se muestran en la figura)
- The **celdas** or **facts** tienen asociados valores numéricos llamados **métricas (measures)**
 - Ejemplo: **cantidad de unidades vendidas por categoría y ciudad del cliente**
- **Dimensiones**: **Product**, **Time**, and **Customer**,
- Métrica: **Quantity**



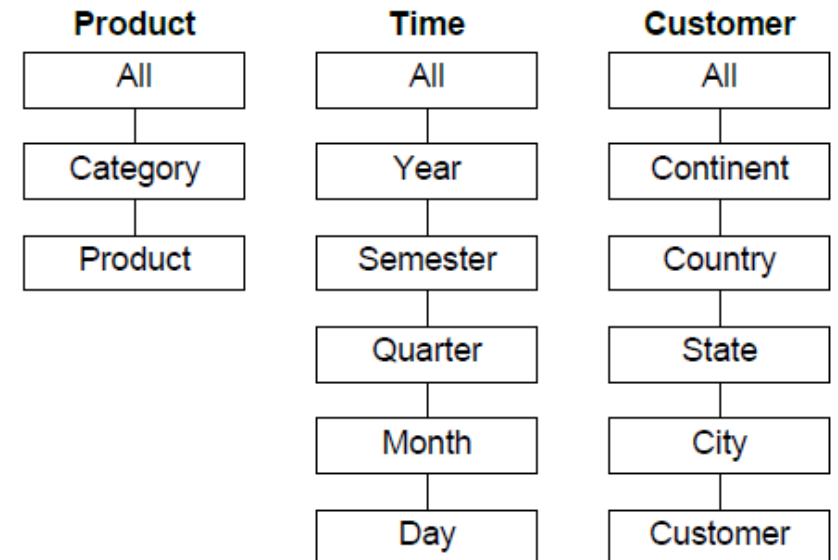
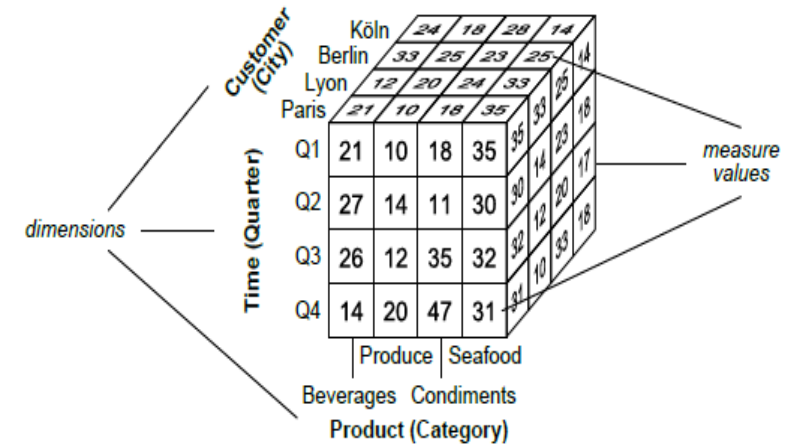
El Modelo Multidimensional

- Importante: Definir la granularidad del data cube
 - En el ejemplo: Trimestre, Ciudad, Categoría
 - Este cubo no es el de granularidad más fina, sino una agregación
- **Dimensiones:** Elementos de una dimension, a una cierta granularidad se denominan “miembros”
 - Ejemplo: Paris es un miembro del nivel “City” de la dimension “Customer”
- **Métricas:** En la granularidad definida, la métrica de la coordenada (Q1, Paris, Beverages) tiene un valor de 21
 - Por ejemplo, la dimension **Producto** puede tener atributos **ProductNumber** y **UnitPrice** (not se muestran en la figura)



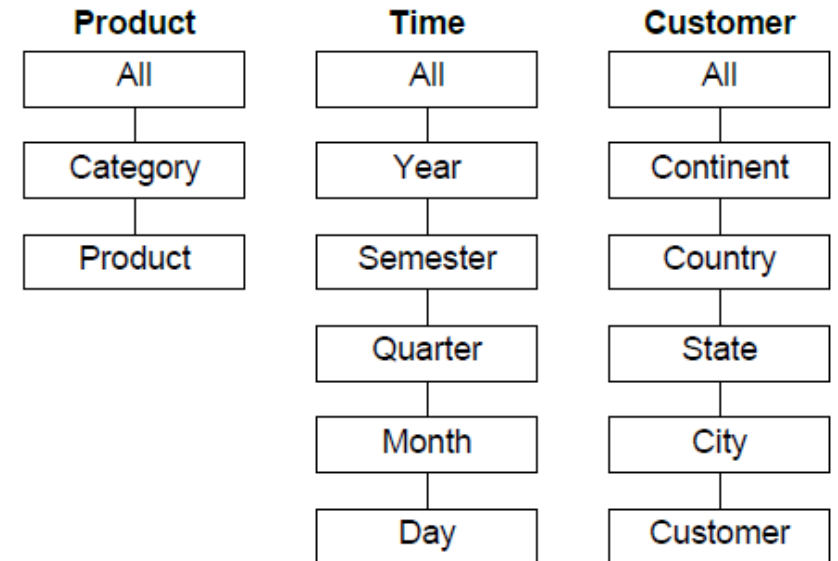
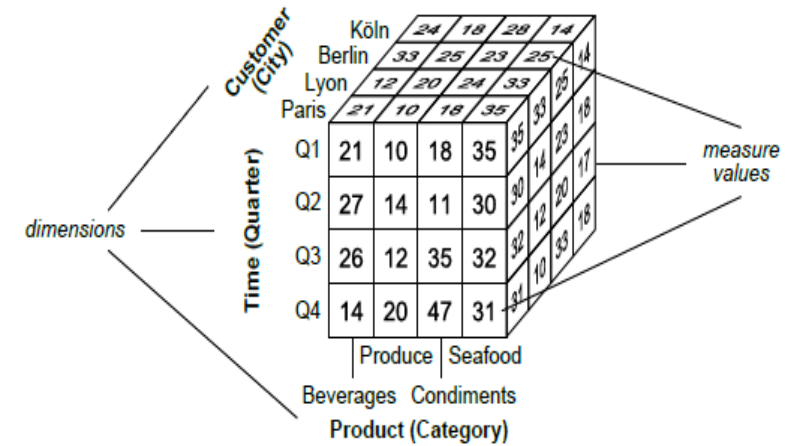
El Modelo Multidimensional

- Las Jerarquías permiten visualizar los datos a diferentes niveles de granularidad
- La estructura jerárquica de una dimension se denomina esquema
- Definir las jerarquías de agregación es fundamental para garantizar la correctitud de las sumalizaciones
- En el ejemplo se muestran las jerarquías de las dimensiones Product, Time, Customer



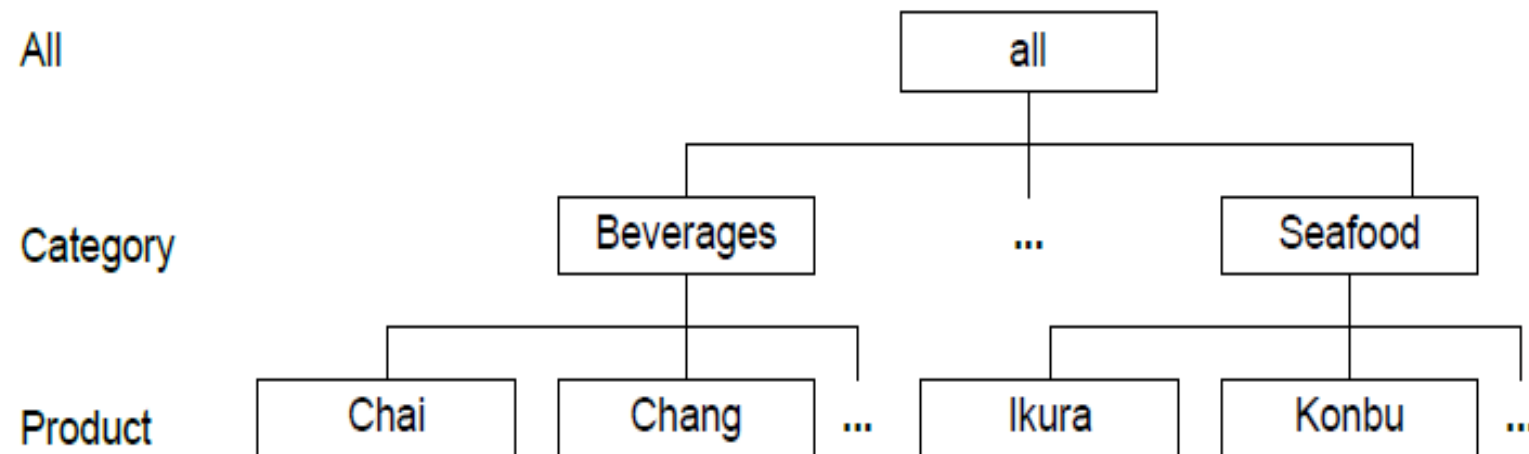
El Modelo Multidimensional

- El **cubo base** está compuesto por los niveles inferiores de las dimensiones
- En el ejemplo, la estructura del cubo base es (Product, Day, Customer)
- Existen muchos posibles cubos con estas 3 dimensiones, partiendo del cubo base:
(Product, Day, Customer), (Product, Day, City), (Product, Day, State),..., (Category, Day, Customer), (Category, Day, City),..., (Category, Year, Continent)



El Modelo Multidimensional

- Los miembros de los niveles de una dimension conforman sus instancias.
- Para la dimension Product Podemos tener, por ejemplo:



El Modelo Multidimensional: Métricas

- Dos clasificaciones ortogonales
- Según su aditividad
 - **Aditivas**: Las más comunes: monto de ventas, costo, salario.
 - **Semi-aditivas**: Cantidad en stock (inventario). No se pueden sumar a lo largo del tiempo.
 - **No-aditivas**: No se pueden sumarizar usando la adición (precio unitario, tasa de cambio)
- Según su distributividad
 - **Distributivas**: Definidas por funciones de agregación que se pueden computar en forma distribuida
 - Count() es distributive; Count distinct() no es distributive
 - $S = \{3,3,4,5,8,4,7,3,8\}$; Count distinct (S) = 5; $S = \{3,3,4\}, \{5,8,4\}, \{7,3,8\}$. Si aplicamos Count distinct a cada subconjunto y lo sumamos, el resultado es 8.
 - **Algebraica**: Funciones escalares de funciones distributivas (ej: Promedio)
 - **Holísticas**: No se pueden computar a partir de sub-agregados (mediana, ranking)
- **Sumarizabilidad**: la correcta agregación de las métricas a lo largo de una dimensión

Operaciones OLAP

- ROLL-UP: Sumariza un cubo a lo largo de una dimension, hasta un nivel, usando una función de agregación

A 3D cube diagram illustrating a roll-up operation. The vertical axis is labeled 'Time (Quarter)' with values Q1, Q2, Q3, and Q4. The horizontal axis is labeled 'Product (Category)' with values Produce, Seafood, Beverages, and Condiments. The depth axis is labeled 'Customer (City)' with values Köln, Berlin, Lyon, and Paris. The cube contains numerical data for each combination of these dimensions.

Time (Quarter)	Customer (City)	Produce	Seafood	Beverages	Condiments
Q1	Köln	21	10	18	35
Q1	Berlin	33	25	23	25
Q1	Lyon	12	20	24	33
Q1	Paris	21	10	18	35
Q2	Köln	27	14	11	30
Q2	Berlin	30	14	20	17
Q2	Lyon	32	12	35	32
Q2	Paris	31	10	33	18
Q3	Köln	26	12	35	32
Q3	Berlin	32	12	35	32
Q3	Lyon	31	10	33	18
Q3	Paris	31	10	33	18
Q4	Köln	14	20	47	31
Q4	Berlin	31	10	33	18
Q4	Lyon	31	10	33	18
Q4	Paris	31	10	33	18

ROLL-UP (C, Customer, Country, SUM)



A 3D cube diagram showing the result of the roll-up operation. The vertical axis is labeled 'Time (Quarter)' with values Q1, Q2, Q3, and Q4. The horizontal axis is labeled 'Product (Category)' with values Produce, Seafood, Beverages, and Condiments. The depth axis is labeled 'Customer (Country)' with values Germany and France. The cube contains numerical data for each combination of these dimensions.

Time (Quarter)	Customer (Country)	Produce	Seafood	Beverages	Condiments
Q1	Germany	57	43	51	39
Q1	France	33	30	42	68
Q2	Germany	39	26	41	44
Q2	France	33	30	42	68
Q3	Germany	30	22	46	44
Q3	France	33	30	42	68
Q4	Germany	25	29	49	41
Q4	France	33	30	42	68

Operaciones OLAP

- DRILL – DOWN: Desagrega un cubo a lo largo de una dimension, hasta un nivel
- En general es la inversa del rollup

A 3D cube diagram illustrating a drill-down operation. The vertical axis is labeled 'Time (Quarter)' with values Q1, Q2, Q3, and Q4. The horizontal axis is labeled 'Customer (City)' with values Köln, Berlin, Lyon, and Paris. The depth axis is labeled 'Product (Category)' with values Produce, Seafood, Beverages, and Condiments. The cube is divided into a 4x4x4 grid of cells, each containing a numerical value. A large blue arrow points from this cube towards the right, indicating the direction of the drill-down operation.

Customer (City)	Köln	Berlin	Lyon	Paris
Q1	21	10	18	35
Q2	27	14	11	30
Q3	26	12	35	32
Q4	14	20	47	31

Product (Category): Produce, Seafood, Beverages, Condiments

DRILL-DOWN (C, Time, Month)

A 3D cube diagram showing the result of a drill-down operation. The vertical axis is labeled 'Time (Month)' with values Jan, Feb, Mar, ..., Dec. The horizontal axis is labeled 'Customer (City)' with values Köln, Berlin, Lyon, and Paris. The depth axis is labeled 'Product (Category)' with values Produce, Seafood, Beverages, and Condiments. The cube is divided into a 4x4x4 grid of cells, each containing a numerical value. The values are more granular than in the previous cube, reflecting the drill-down to the month level.

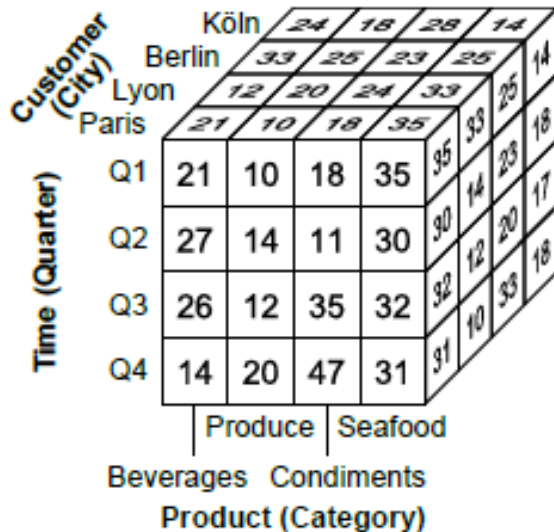
Customer (City)	Köln	Berlin	Lyon	Paris
Jan	7	2	6	20
Feb	8	4	8	8
Mar	6	4	4	7
...
Dec	4	4	16	7

Product (Category): Produce, Seafood, Beverages, Condiments

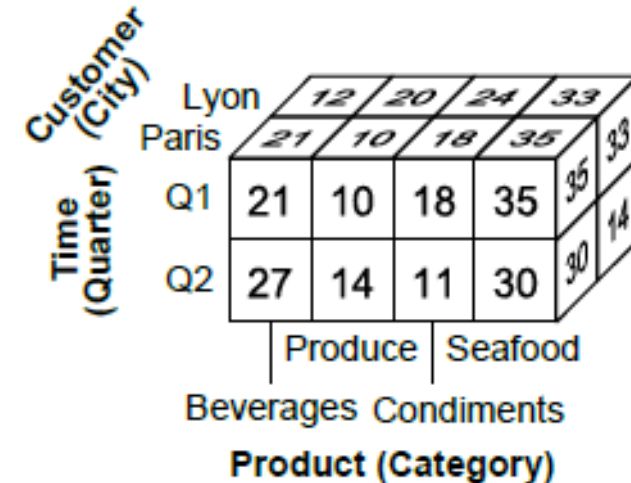
Equivalente a sumarizar el cubo base hasta el nivel **Month**

Operaciones OLAP

- DICE: Obtiene un subcubo que satisface una condición Booleana
- La dimension a eliminar debe tener un único miembro



DICE (C, (city='Paris' OR Lyon)
and Quarter= 'Q3' or 'Q4')



Operaciones OLAP

- SLICE: Elimina una dimension del cubo
- La dimension a eliminar debe tener un único miembro. Si previamente se hizo un DICE por City = 'Paris', se toma simplemente una “rodaja” del cubo.

Customer (City)

Time (Quarter)

Product (Category)

	Köln	Berlin	Lyon	Paris
Q1	21	10	18	35
Q2	27	14	11	30
Q3	26	12	35	32
Q4	14	20	47	31

SLICE (C, Customer)



Time (Quarter)

Q1	21	10	18	35
Q2	27	14	11	30
Q3	26	12	35	32
Q4	14	20	47	31

Product (Category)