UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
FACULTAD DE CIENCIAS
ALMACENES Y MINERÍA DE DATOS

Procesamiento analítico en línea

Gerardo Avilés Rosas
gar@ciencias.unam.mx

OLTP vs. OLAP

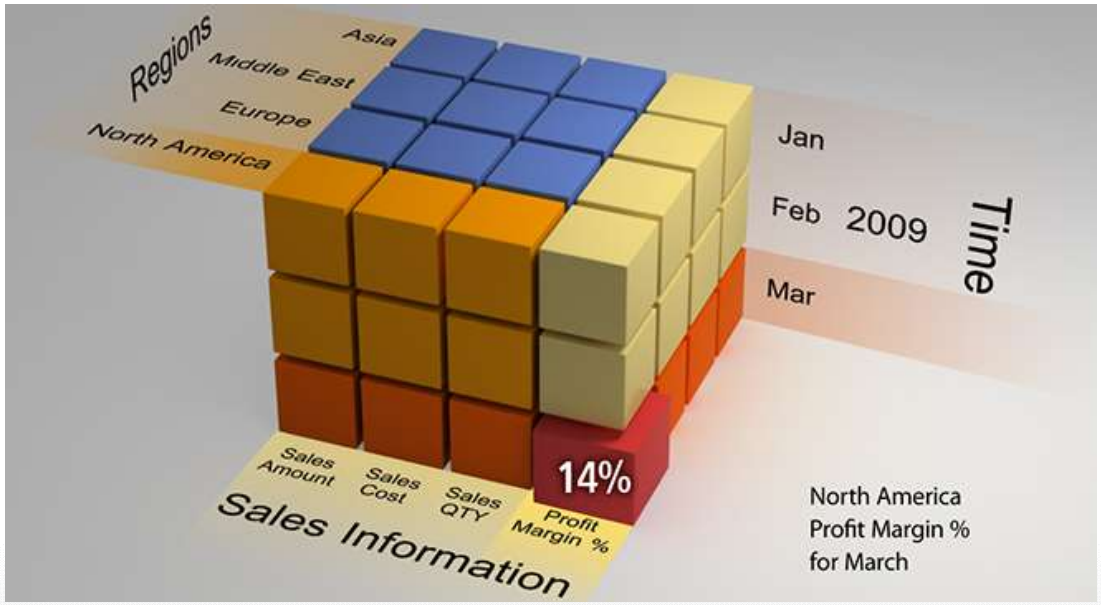
Procesamiento de transacciones en línea (OLTP)

- **Muchas** consultas **"pequeñas"** sobre una cantidad **"pequeña"** de tuplas de varias tablas que requieren unirse.
- **Altamente volátil**. El sistema siempre está disponible para actualizaciones y/o consultas.
- **Volumen pequeño** de datos → unos cuantos históricos
- **Modelo de datos** complejo → normalizado

…OLTP vs. OLAP

Procesamiento analítico en línea (OLAP)

- **Menos consultas**, pero más grandes, generalmente requieren rastrear una **gran cantidad** de datos y hacer **agregaciones**.
- **Lecturas frecuentes** y **variante en el tiempo → actualizaciones frecuentes** (*diariamente, semanalmente*)
- Operaciones en dos fases: **lectura o actualización**
- **Grandes volúmenes** de datos → *perspectiva histórica*
- Modelo de **datos sencillo** → *multidimensional/denormalizado*



Online Analytic Processing

- Se trata de un proceso computacional que permite al usuario **extraer fácil y de manera selectiva** datos, para presentarlos desde distintos puntos de vista.
- Permite **analizar información** proveniente de múltiples fuentes de datos heterogéneas al mismo tiempo.
- Suele almacenarse en bases de **datos multidimensionales**.
- Las consultas que puede ejecutar son complejas debido a que:
  - ✓ Toman grandes cantidades de datos.
  - ✓ Pueden descubrir patrones y tendencias en los datos.
  - ✓ Típicamente son costosas con respecto al tiempo.
  - ✓ Son conocidas como consultas de apoyo a la toma decisiones.

Cubos OLAP

- Se trata de la forma más popular para analizar información proveniente de **bases de datos multidimensionales**.
- Básicamente, un **cubo** es una estructura de datos organizada mediante **jerarquías**. En la intersección de las dimensiones se encuentran las **medidas** y cada una de ellas se puede evaluar en cualquiera de los niveles de las jerarquías:

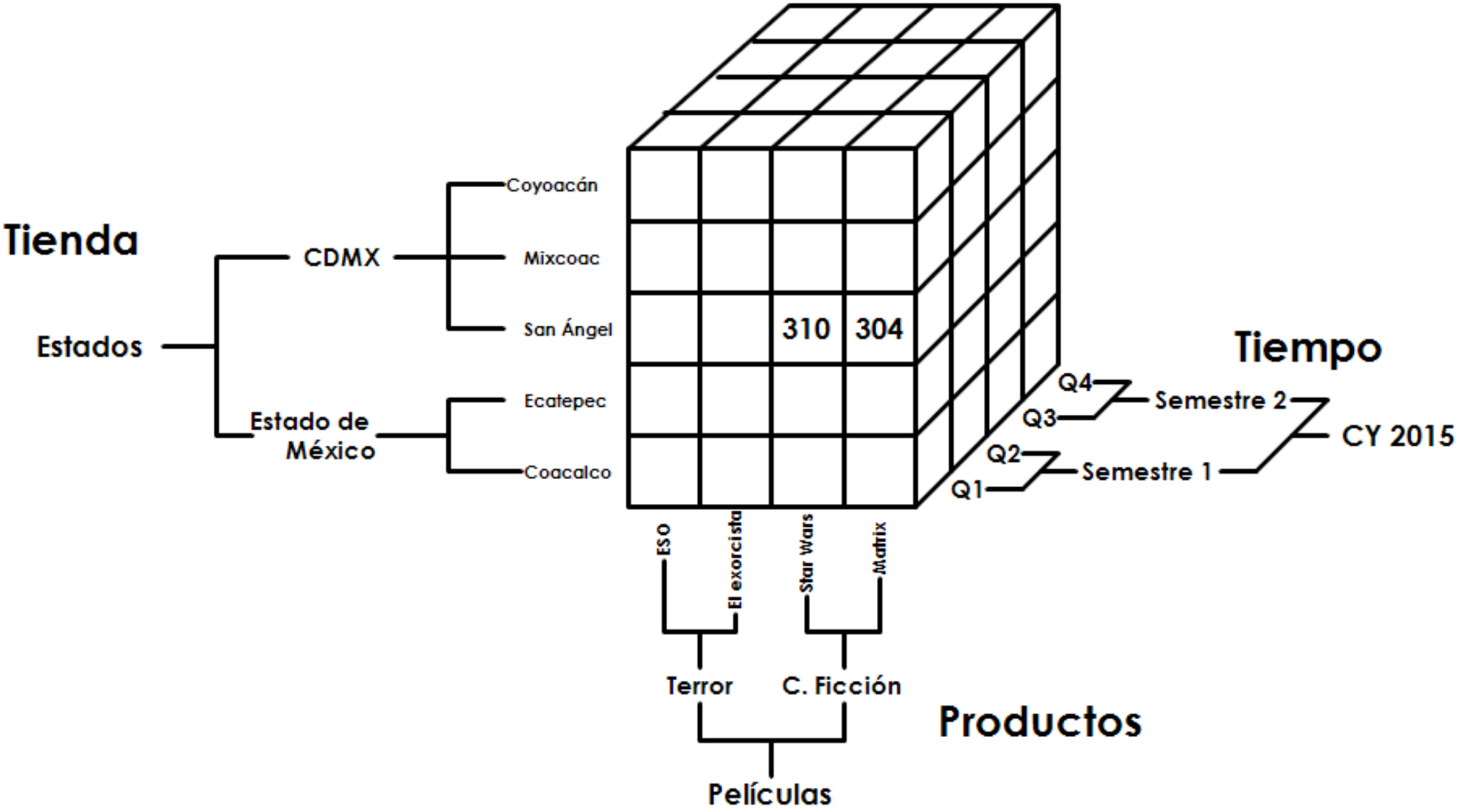
  *Analizar las **ventas** diaria, mensual o anualmente, para un cliente, una región o un país*.

- Tienen la capacidad de **analizar** y **explorar** los datos:

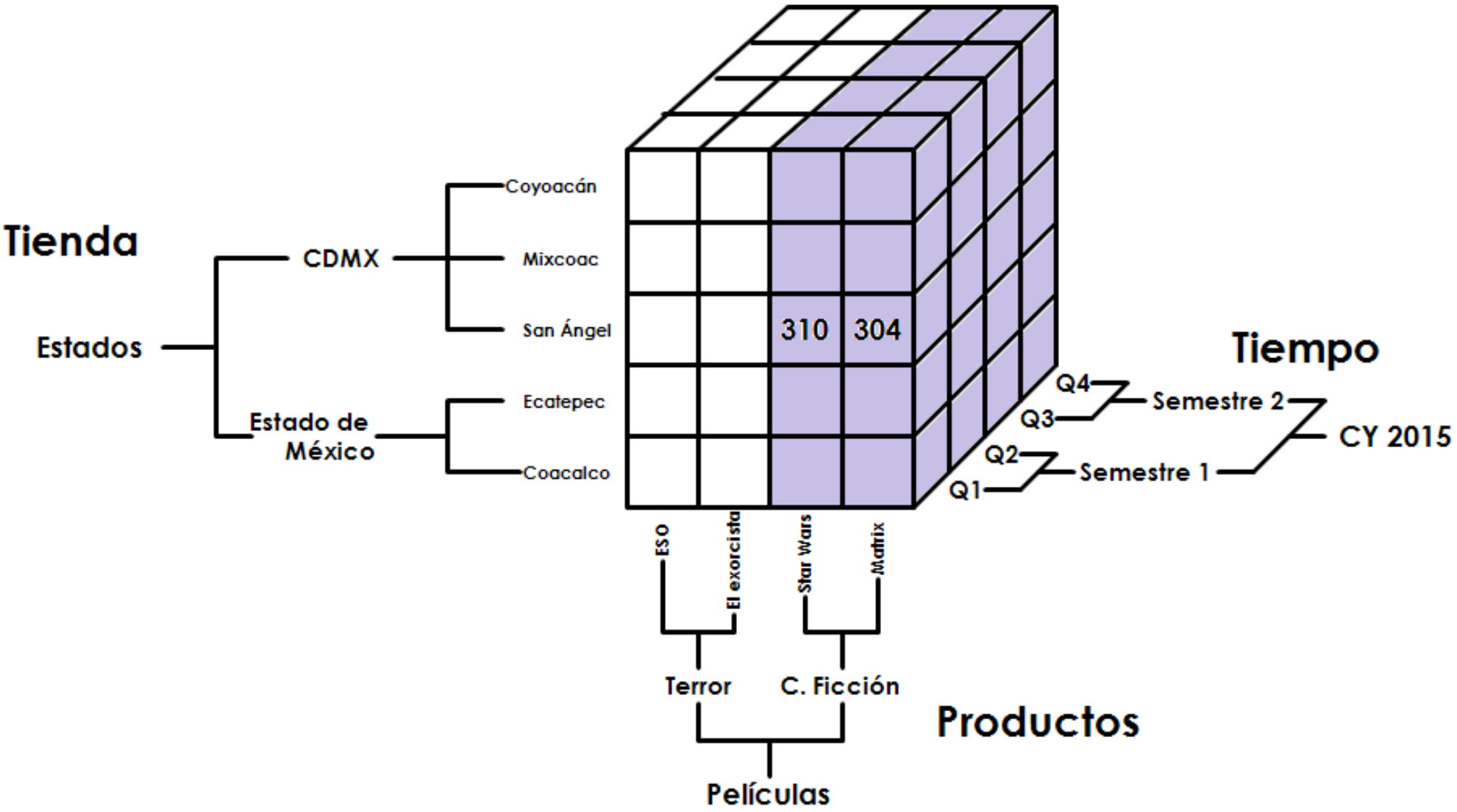
  *Permiten cambiar el enfoque del **¿qué esta pasando?** (enfoque relacional) al **¿por qué esta pasando?** (enfoque multidimensional)*.

- Las herramientas con capacidades **OLAP** proporcionan **análisis interactivo** a través de las diferentes **dimensiones** de los datos.

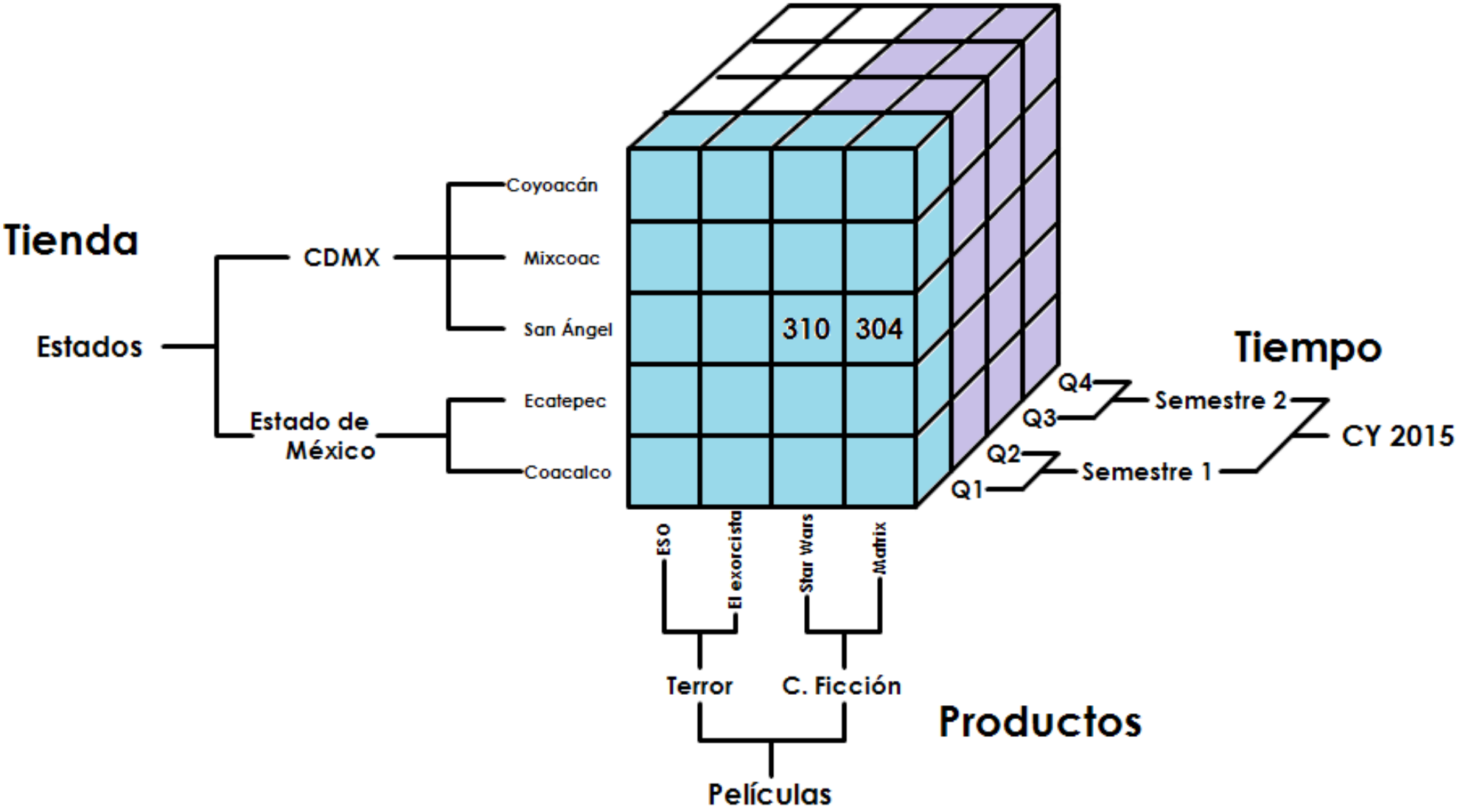
...Cubos OLAP


Tienda
Estados
CDMX
Estado de México
Coyoacán
Mixcoac
San Ángel
Ecatepec
Coacalco
310
304
Tiempo
Q4
Q3
Q2
Q1
Semestre 2
Semestre 1
CY 2015
ESO
El exorcista
Star Wars
Matrix
Terror
C. Ficción
Películas
Productos

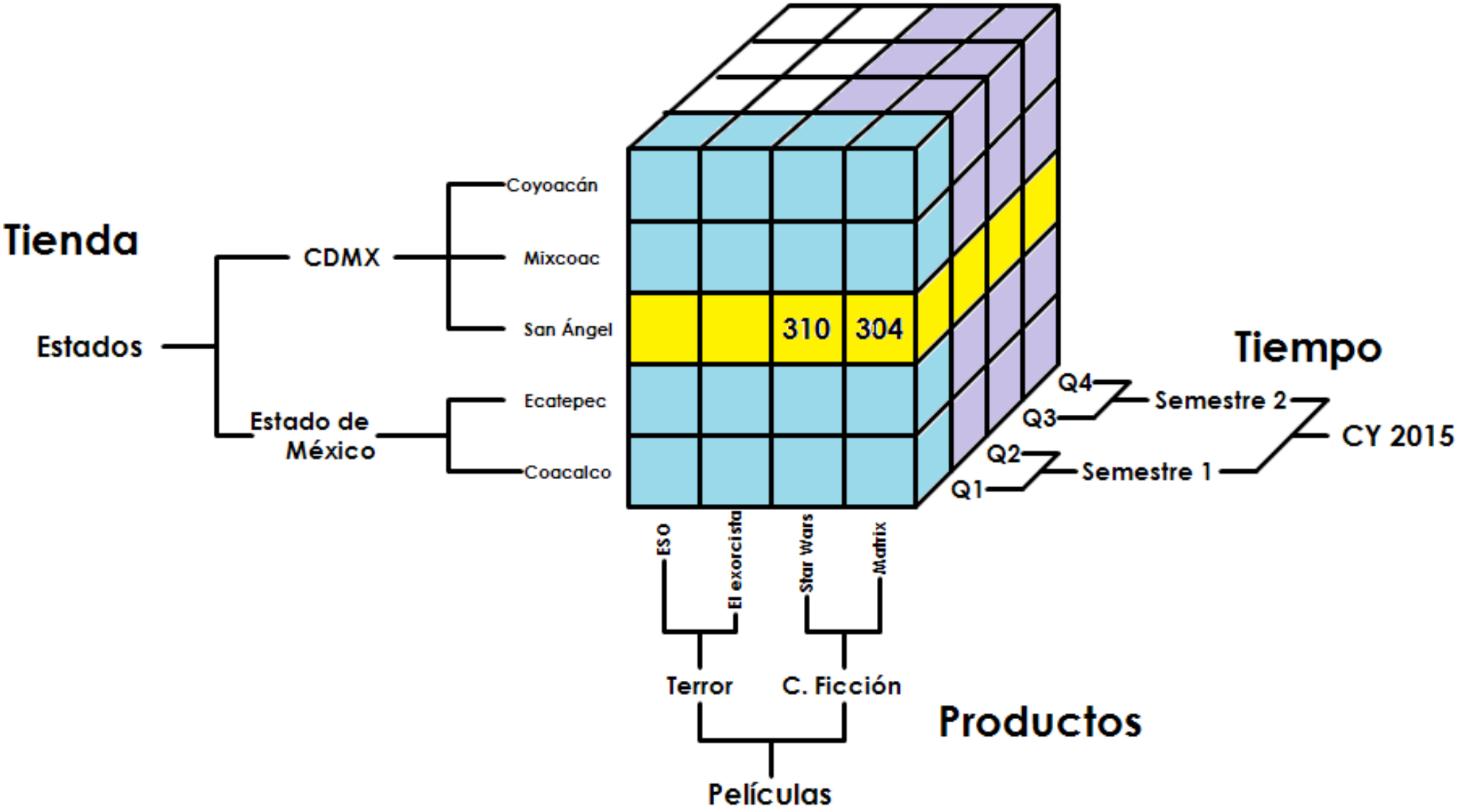
…Cubos OLAP


Tienda
Estados
CDMX
Estado de México
Coyoacán
Mixcoac
San Ángel
Ecatepec
Coacalco
310
304
Tiempo
Q4
Q3
Q2
Q1
Semestre 2
Semestre 1
CY 2015
ESO
El exorcista
Star Wars
Matrix
Terror
C. Ficción
Películas
Productos

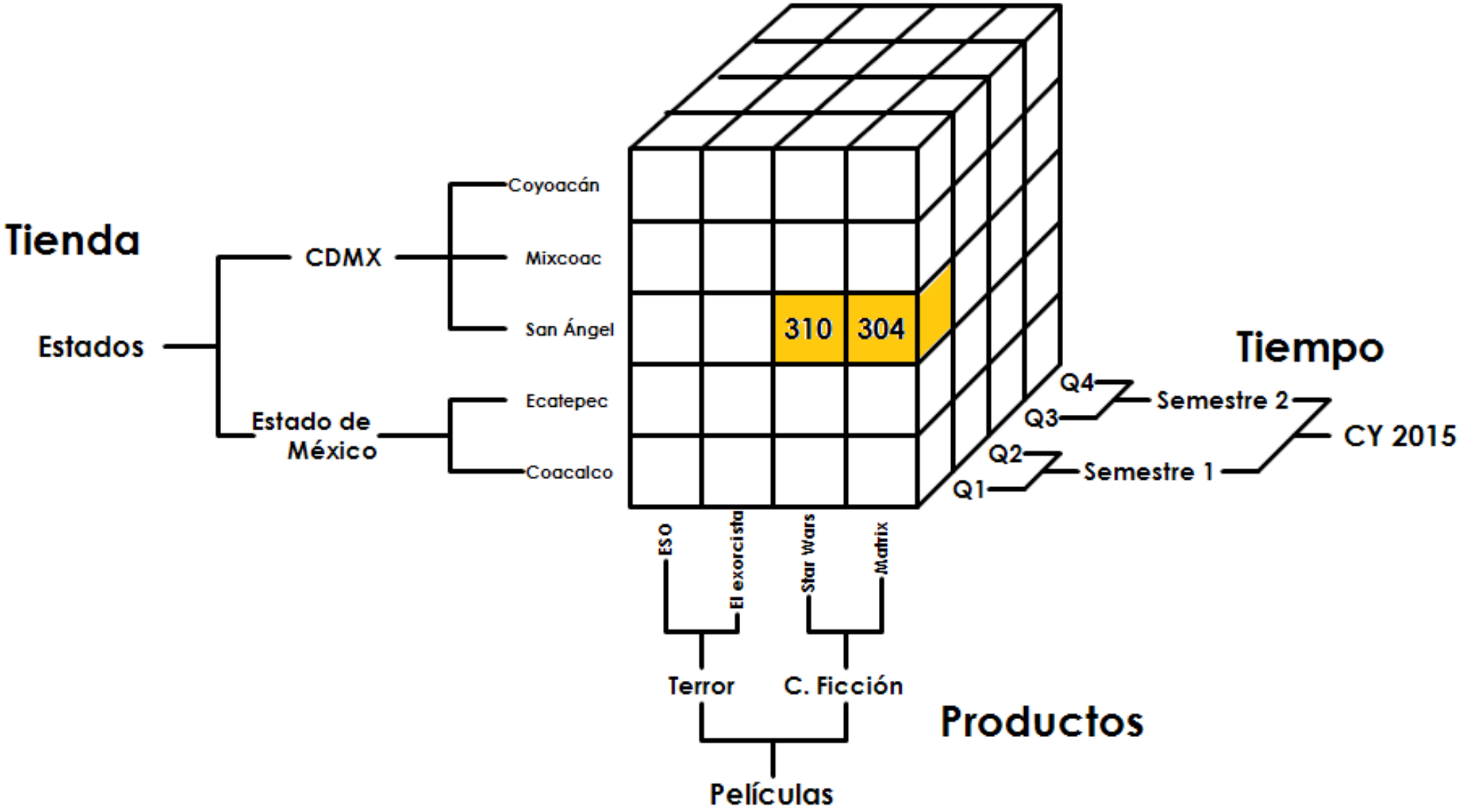
…Cubos OLAP


Tienda
Estados
CDMX
Estado de México
Coyoacán
Mixcoac
San Ángel
Ecatepec
Coacalco
310
304
Tiempo
Q4
Q3
Q2
Q1
Semestre 2
Semestre 1
CY 2015
ESO
El exorcista
Star Wars
Matrix
Terror
C. Ficción
Películas
Productos

…Cubos OLAP


Tienda
Estados
CDMX
Estado de México
Coyoacán
Mixcoac
San Ángel
Ecatepec
Coacalco
310
304
Tiempo
Q4
Q3
Q2
Q1
Semestre 2
Semestre 1
CY 2015
ESO
El exorcista
Star Wars
Matrix
Terror
C. Ficción
Películas
Productos

...Cubos OLAP


Tienda
Estados
CDMX
Estado de México
Coyoacán
Mixcoac
San Ángel
Ecatepec
Coacalco
310
304
Tiempo
Q4
Q3
Q2
Q1
Semestre 2
Semestre 1
CY 2015
ESO
El exorcista
Star Wars
Matrix
Terror
C. Ficción
Películas
Productos

Ventajas

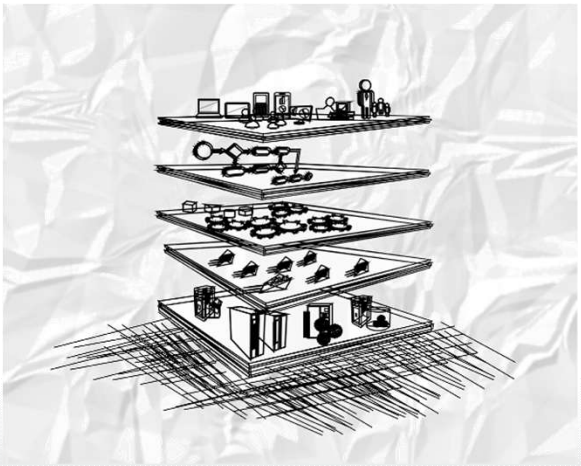
El uso de **cubos OLAP** tiene dos ventajas fundamentales:

- **Facilidad de uso**

Una vez construido el cubo, el usuario de negocio puede consultarlo con facilidad, ***incluso si se trata de un usuario con escasos o nulos conocimientos técnicos***. La estructura jerárquica es sumamente fácil de comprender. El cubo se convierte en una gran ***"tabla dinámica"*** que el usuario puede consultar en cualquier momento.

- **Rapidez de respuesta**

Habitualmente, el cubo tiene distintas **agregaciones precalculadas**, por lo que los tiempos de respuesta son muy cortos.



Desventajas

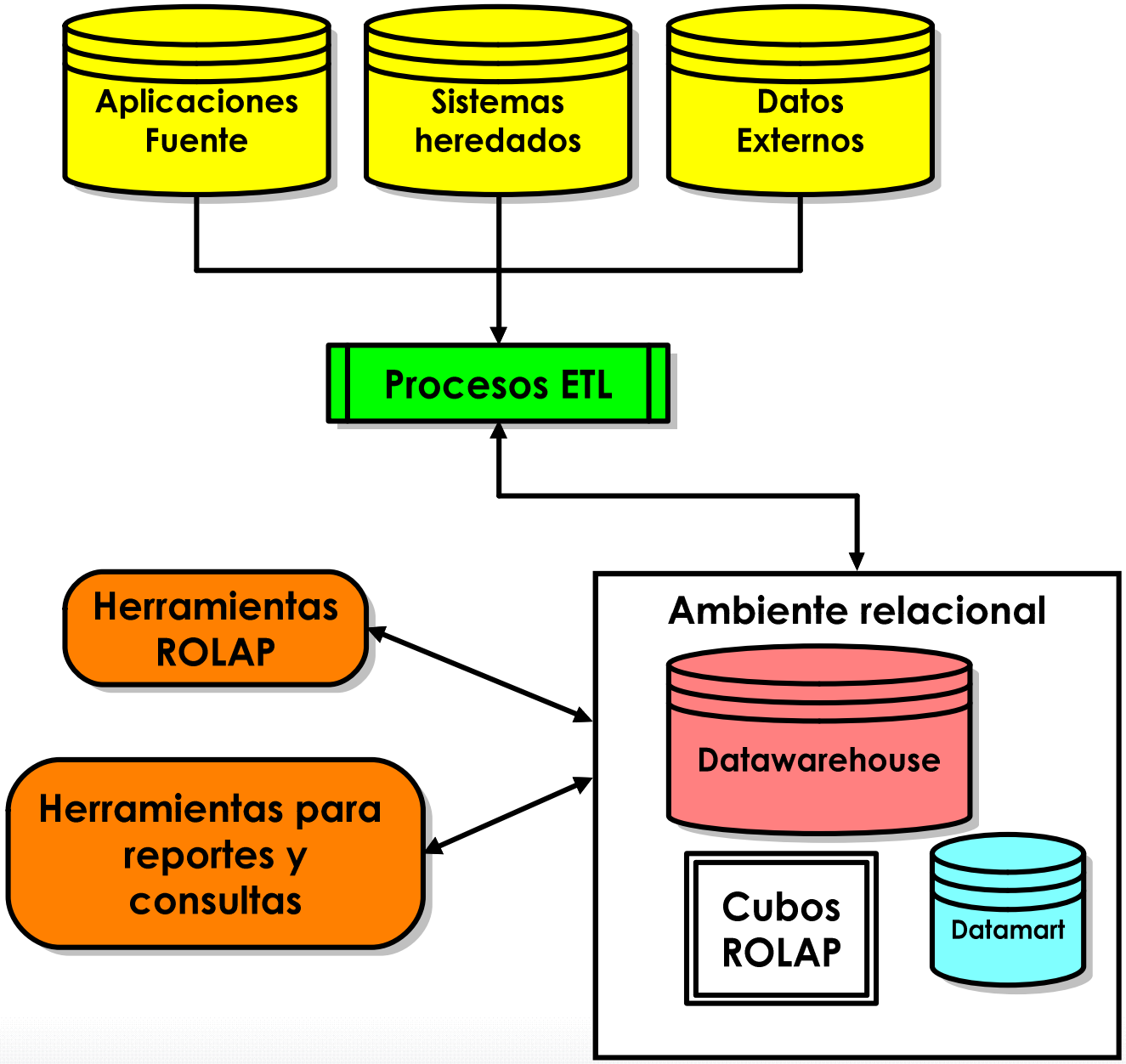
- El cubo es estructura adicional de datos que se debe **mantener** y en algunos caso **actualizar** (esto supone un gasto extra de recursos: *servidores, discos, procesos de carga, etc.*)
- El modelo de negocio no siempre se adapta bien en un modelo basado en jerarquías, por ejemplo:
  - *Una semana no pertenece a un único mes.*
  - *Las zonas de venta no tienen por qué coincidir con la estructura de regiones de cada país.*
  - *Se puede tener a varios responsables pueden encargarse de una misma tienda.*
  - *Distintos departamentos de la compañía pueden utilizar distintas agrupaciones de los productos.*

Implementación: ROLAP

- En este tipo de plataforma se almacenan los datos en una **base de datos relacional**, lo que implica que no es necesario que los datos se repliquen en un almacenamiento separado para el análisis.
- Los cálculos se realizan en una **BD relacional**, con **grandes volúmenes** de datos y tiempos de navegación no predecibles.
- El sistema **ROLAP** utiliza una arquitectura de tres niveles:

1. El **nivel de base de datos** utiliza bases de datos relacionales para el manejo, acceso y obtención de datos.
2. El **nivel de aplicación** es el motor que ejecuta las consultas multidimensionales de los usuarios.
3. El **motor ROLAP** se integra con niveles de presentación, a través de los cuales los usuarios realizan los análisis OLAP.

…Implementación: ROLAP



…Implementación: ROLAP

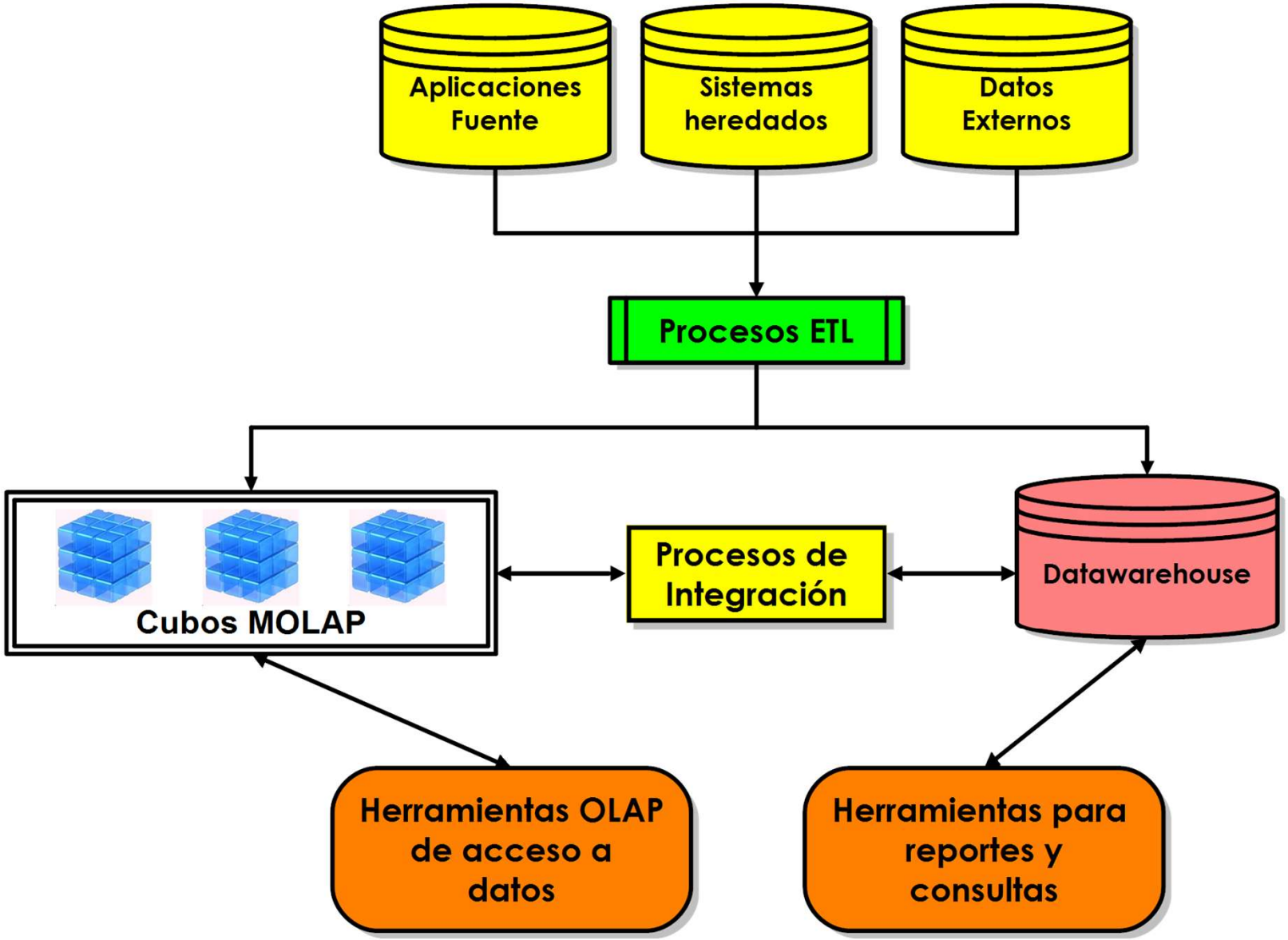
- Los datos se cargan desde el sistema operacional y se crean **índices** para optimizar los tiempos de acceso a las consultas.
- Los análisis multidimensionales se transforman dinámicamente a **consultas SQL**. Los resultados se relacionan mediante **tablas cruzadas** y conjuntos multidimensionales.
- Esta arquitectura usa **datos precalculados** (*siempre que estén disponibles*), o bien, generarlos dinámicamente desde los datos elementales.
- Como se accede directamente a los datos del DWH, soportan técnicas de optimización de accesos (*acelerar consultas*): ***particionado de los datos a nivel de aplicación, denormalización y joins múltiples***.

Implementación: MOLAP

- Los datos son **replicados** en plataformas con un almacenamiento construido a propósito que asegura mayor velocidad en los análisis.
- Los cálculos se llevan a cabo en un servidor con una **base de datos multidimensional**, partiendo de la premisa que un sistema **OLAP** estará mejor implementado si se almacenan los datos de forma multidimensional.
- El sistema **MOLAP** utiliza una arquitectura de **dos niveles**:

1. La **base de datos multidimensional** es la encargada del manejo, acceso y obtención de los datos.
2. El **nivel de aplicación** es el responsable de la ejecución de los requerimientos OLAP. El **nivel de presentación** se integra con el de aplicación y proporciona un interfaz a través del cual los usuarios finales visualizan los análisis OLAP. Una arquitectura cliente/servidor permite a varios usuarios acceder a la misma base de datos multidimensional.

…Implementación: MOLAP

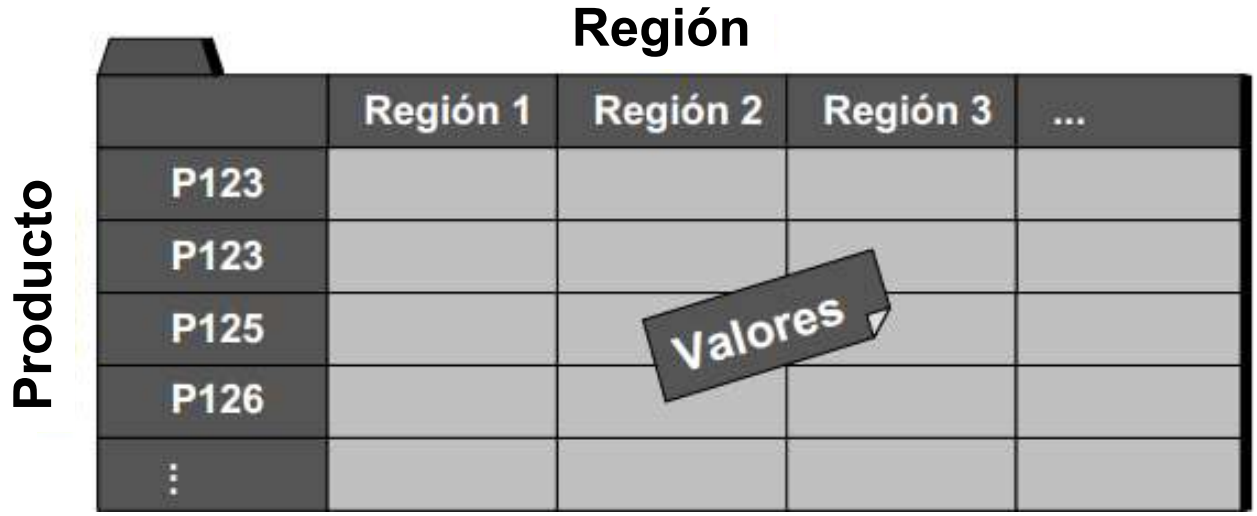


…Implementación: MOLAP

- La información procedente de los sistemas operacionales, se carga en el sistema **MOLAP**, mediante una serie de rutinas batch. Una vez cargados los datos BDMD, se realizan una serie de cálculos en batch, para obtener los datos agregados.
- Se manejan **índices** y **tablas hash** para mejorar los tiempos de accesos en las consultas.
- La arquitectura MOLAP requiere **cálculos intensivos** de compilación: *lee datos precompilados, y tiene capacidades limitadas de crear agregaciones dinámicamente o de encontrar agregaciones que no se hayan precalculado y/o almacenado previamente*.

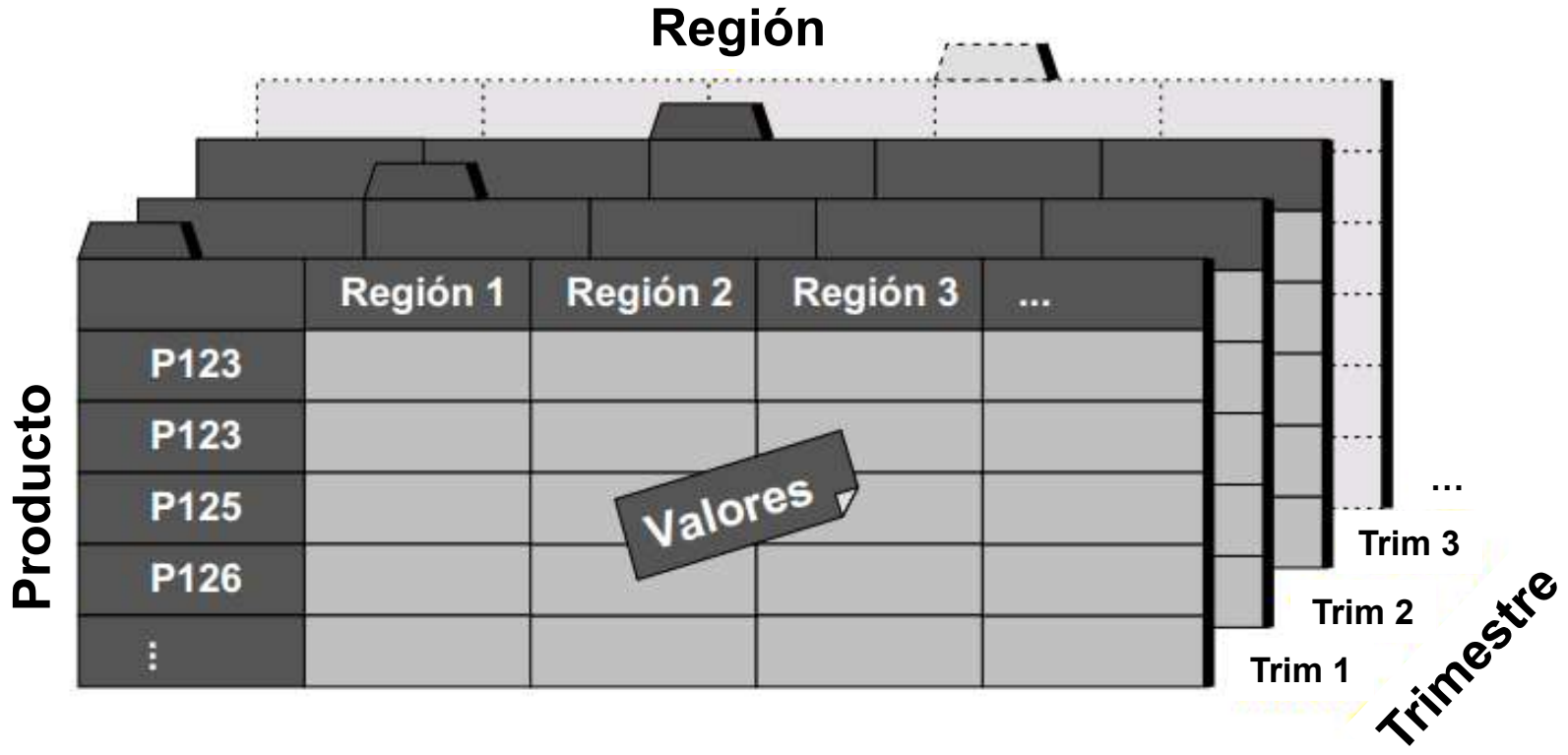
Modelo de datos

- Es un conjunto de conceptos que pueden usarse para describir la estructura de un **data warehouse**.
- La estructura corresponde con los tipos y estructuras de datos, sus relaciones, restricciones que deberían permitir a los datos.
- Por ejemplo, en una hoja de cálculo podemos encontrar una **matriz de dos dimensiones**:



…Modelo de datos

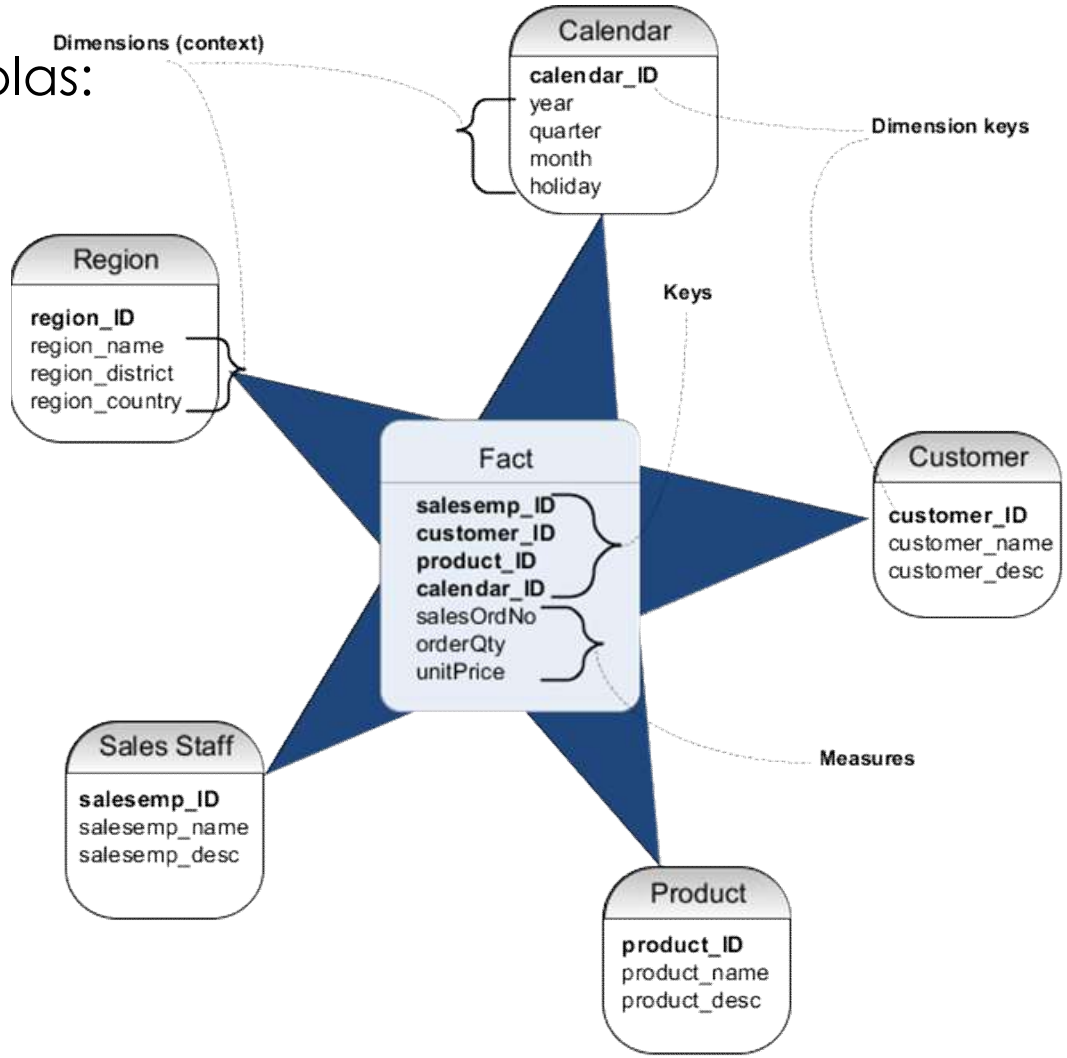
- Siguiendo con el mismo ejemplo, si añadimos una dimensión más, tendríamos una **matriz de tres dimensiones**:



- De esta forma, las herramientas de explotación OLAP han adoptado un modelo multidimensional de los datos.

…Modelo de datos

- El **modelo multidimensional** es un método basado en el **modelo relacional**.
- Se compone de dos tipos de tablas:
  - Varias **tablas de dimensión**, cada una formada por tuplas de atributos que permitirán describir medidas.
  - Una **tabla de hecho** *(pueden ser más)*, compuesta por tuplas, una por cada hecho registrado. Los hechos contienen **medias u observaciones** y se relacionan con las tablas de dimensión a través de **llaves foráneas**.



Procesamiento analítico

- Los datos en el **modelo multidimensional** son percibidos y manejados como si estuvieran almacenados en una **matriz de varias dimensiones**.
- Este procesamiento requiere invariablemente algún tipo de agregación de datos y generalmente, desde distintas perspectivas:

  - **El problema fundamental de este tipo de procesamiento, es la cantidad de agrupamientos, la cual llega a ser muy grande rápidamente y los usuarios deben considerarlos todos o casi todos.**

- El **lenguaje de consulta estructurado** (SQL) soporta la agregación que se requiere, sin embargo, cada consulta individual produce como resultado una única tabla (**todas las filas en la tabla tiene por ende, una misma forma y misma interpretación**):

n agrupamientos = n consultas = n tablas

…Procesamiento analítico

Por ejemplo, tenemos información de los prestamos que se realizan en las sucursales de un banco, pertenecientes a un estado:

PRÉSTAMOS			
ESTADO	SUCURSAL	NUM_PRES	IMPORTE
GUANAJUATO	SALAMANCA	P-2808	492672.00
GUANAJUATO	ACAMBARO	P-5054	173768.00
QUINTANA ROO	CANCUN	P-3557	119835.00
OAXACA	TEHUANTEPEC	P-7287	474981.00
COAHUILA	TULIPANES	P-6768	142285.00
VERACRUZ	POZA RICA	P-3968	413654.00
GUANAJUATO	ACAMBARO	P-6703	366239.00
GUANAJUATO	IRAPUATO	P-4270	230897.00
PUEBLA	ACATZINGO	P-6454	142175.00
SAN LUIS POTOSÍ	SAN LUIS POTOSI	P-8627	292952.00
MORELOS	OAXTEPEC	P-1829	10115.00
NUEVO LEÓN	MONTERREY	P-8021	467128.00
YUCATÁN	PROGRESO	P-5106	73258.00
SINALOA	CULIACAN	P-2783	119978.00
TAMAULIPAS	TAMPICO	P-2857	238885.00
MICHOACÁN	URUAPAN	P-1052	460425.00
MICHOACÁN	PATZCUARO	P-7248	477315.00
ZACATECAS	FRESNILLO	P-6830	281662.00

…Procesamiento analítico

Nos interesan las siguientes consultas sobre la tabla:

- Obtener la cantidad total de prestamos que se han otorgado a nivel nacional.
- Obtener la cantidad de prestamos que se han otorgado por estado.
- Obtener la cantidad de prestamos por sucursal.
- Obtener la cantidad de prestamos por estado y sucursal.

Recordemos un poco de SQL…

…Procesamiento analítico

- Obtener la cantidad total de prestamos que se han otorgado a nivel nacional:

```
select count(num_pres)
from prestamos;
```

COUNT(NUM_PRES)
4865

…Procesamiento analítico

- Obtener la cantidad de prestamos que se han otorgado por estado.

```
select estado,count(num_pres)
from prestamos
group by estado
order by estado;
```

ESTADO	COUNT(NUM_PRES)
CHIAPAS	128
CHIHUAHUA	164
COAHUILA	124
DISTRITO FEDERAL	44
ESTADO DE MÉXICO	506
GUANAJUATO	210
GUERRERO	122
HIDALGO	126
JALISCO	336
MICHOACÁN	250

…Procesamiento analítico

- Obtener la cantidad de prestamos por sucursal.

```
select sucursal,count(num_pres)
from prestamos
group by(sucursal)
order by sucursal;
```

SUCURSAL	COUNT(NUM_PRES)
ACAMBARO	40
ACAPULCO	40
ACATZINGO	40
ALTAMIRA	42
AMECAMECA	42
APODACA	42
ARAGON	40
ARANDAS	40
ATLACOMULCO	40
BOCA DEL RIO	40
BONAMPAK	44
BUENAVISTA	44

…Procesamiento analítico

- Obtener la cantidad de prestamos por estado y sucursal.

```
select estado,sucursal,count(num_pres)
from prestamos
group by estado,sucursal
order by estado,sucursal;
```

ESTADO	SUCURSAL	COUNT(NUM_PRES)
CHIAPAS	BONAMPAK	44
CHIAPAS	TONALA	44
CHIAPAS	VILLAFLORES	40
CHIHUAHUA	CAMARGO	44
CHIHUAHUA	DELICIAS	40
CHIHUAHUA	JIMENEZ	40
CHIHUAHUA	PASO DEL NORTE	40
COAHUILA	LA ROSITA	42
COAHUILA	PIEDRAS NEGRAS	42
COAHUILA	TULIPANES	40
DISTRITO FEDERAL	DIVISION DEL NORTE	44

…Procesamiento analítico

- Las desventajas de este enfoque son obvias:
  - La formulación de estas consultas es tediosa para el usuario: si queremos importe promedio, el mayor importe prestado, etc.
  - La ejecución de todas esas consultas (pasan todas, por los mismos datos) es probablemente costosa en tiempo de ejecución.
- Valdría la pena trata de encontrar una forma de:
  - Solicitar varios niveles de agregación en una sola consulta.
  - Ofrecer a la implementación la oportunidad de calcular todas esas agregaciones de manera más eficiente.
- **SABD** como **Microsoft SQL Server** u **Oracle**, permiten realizar las consultas anteriores en un solo paso a través de la cláusula **GROUP BY**.

…Procesamiento analítico

- La opción **GROUPING SETS**, permite al usuario especificar con exactitud qué agrupamientos específicos van a realizarse:

```
select estado,sucursal,count(num_pres)
from prestamos
group by grouping sets (estado,sucursal)
order by estado,sucursal;
```

Grupo **Estado**

ESTADO	SUCURSAL	COUNT(NUM_PRES)
CHIAPAS	null	128
CHIHUAHUA	null	164
COAHUILA	null	124
DISTRITO FEDERAL	null	44
…	Null	
null	ACAMBARO	40
null	ACAPULCO	40
null	ACATZINGO	40
null	ALTAMIRA	42
null	AMECAMECA	42

Grupo **Sucursal**

…Procesamiento analítico

- La opción **ROLLUP**, es una forma de abreviar ciertas combinaciones de **GROUPING SETS**:

```
select estado,sucursal,count(num_pres)
from prestamos
group by roll up (estado,sucursal)
order by estado,sucursal;
```

ESTADO	SUCURSAL	COUNT(NUM_PRES)
CHIAPAS	BONAMPAK	44
CHIAPAS	TONALA	44
CHIAPAS	VILLAFLORES	40
CHIAPAS	Null	128
CHIHUAHUA	CAMARGO	44
CHIHUAHUA	DELICIAS	40
CHIHUAHUA	JIMENEZ	40
CHIHUAHUA	PASO DEL NORTE	40
CHIHUAHUA	null	164
…	…	…
null	null	4865

Grupo **estado y sucursal**

Grupo **estado**

Nivel **nacional**

…Procesamiento analítico

- El término **ROLLUP** se deriva del hecho de que las cantidades han sido ***enrolladas*** en toda la ***"dimensión"*** del estado.
- En general **GROUP BY ROLLUP (A,B,…,Z)** *significa agrupar en todas las combinaciones siguientes*:

(A,B,…,Z)

(A,B,…)

(A,B)

(A)

()

- Como se puede observar, hay muchos ***"enrollar en la dimensión A"*** distintos, dependiendo de qué otras columnas son mencionadas en la lista **ROLLUP**.
- Es importante mencionar que **ROLLUP (A,B)** tiene un significado distinto que ROLLUP **(B,A)**.

…Procesamiento analítico

- La opción **CUBE**, es otra forma de abreviar ciertas combinaciones de **GROUPING SETS**:

```
select estado,sucursal,count(num_pres)
from prestamos
group by cube (estado,sucursal)
order by estado,sucursal;
```

ESTADO	SUCURSAL	COUNT(NUM_PRES)
CHIAPAS	BONAMPAK	44
CHIAPAS	TONALA	44
CHIAPAS	VILLAFLORES	40
CHIAPAS	null	128
…	…	…
null	ACAMBARO	40
null	ACAPULCO	40
null	ACATZINGO	40
null	ALTAMIRA	42
null	AMECAMECA	42
null	…	…
null	null	4865

Grupo **estado y sucursal**

Grupo **estado**

Grupo **sucursal**

Nivel **nacional**

…Procesamiento analítico

- El término **CUBE** se deriva del hecho de que en la terminología OLAP, los valores de datos pueden ser percibidos como si estuvieran almacenados en la celdas de una matriz multidimensional.
- En el caso que estamos revisando, el cubo es de dos dimensiones: estado y sucursal.
- **GROUP BY CUBE (A,B,C,…,Z)** significa "*agrupar por todos los subconjuntos posibles del conjunto (A,B,C,…,Z)*".

…Procesamiento analítico

- Con frecuencia, los productos OLAP muestran los resultados, no como tablas estilo SQL, sino como **referencias cruzadas**:

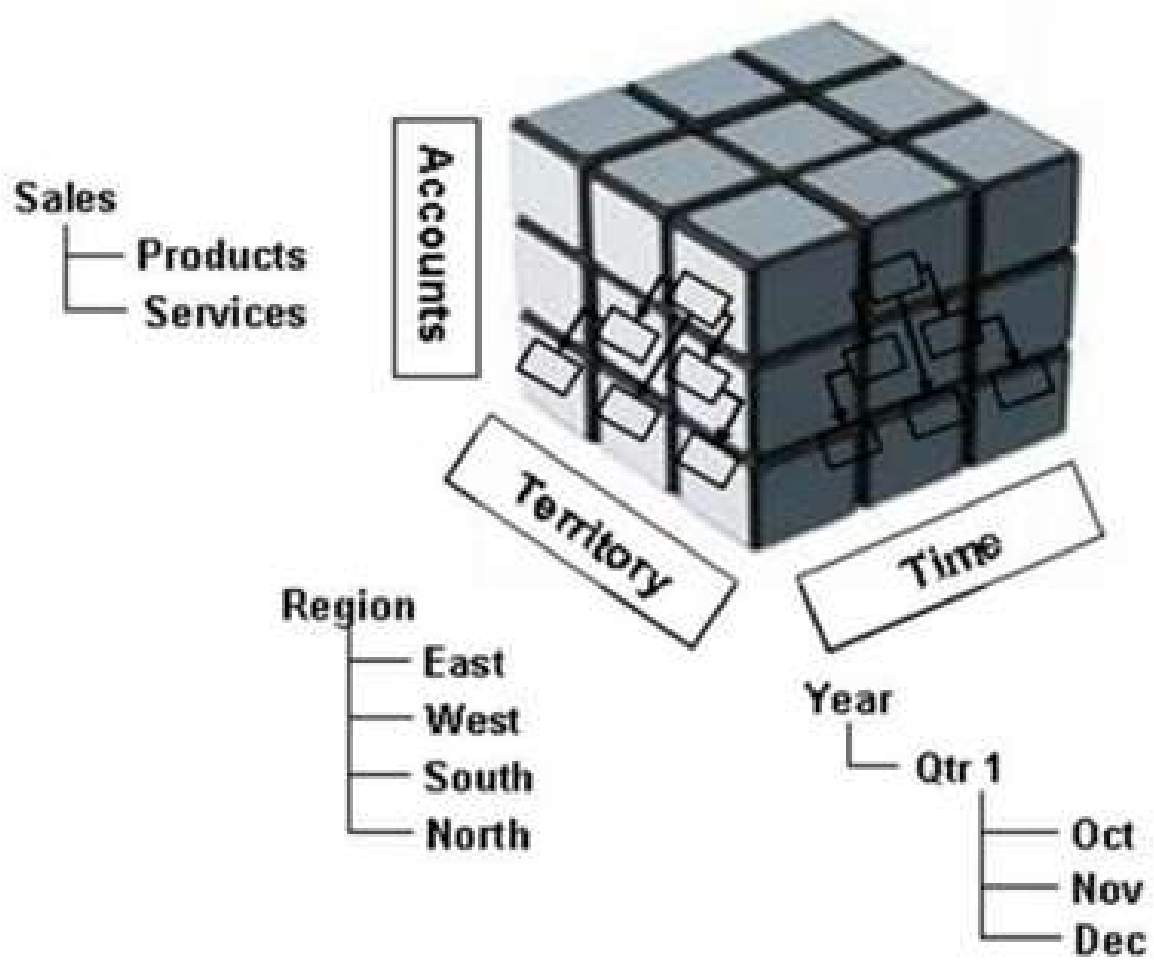
	BONAMPAK	TONALA	VILLAFLORES	CAMARGO	DELICIAS	…	TOTAL
CHIAPAS	44	44	40	0	0	…	128
CHIHUAHUA	0	0	0	44	40		164
…	…	…	…	…	…	…	…
TOTAL	44	44	40	44	40	…	4865

```
select estado,nombresucursal as sucursal,año,
       [1] as 'Q1',[2] as 'Q2',[3] as 'Q3', [4] as 'Q4',
       [1] + [2] + [3] + [4] as "Total año”
from (select estado,nombresucursal,año,trimestre,
             numprestamo from olap) res
pivot(count(numprestamo) for trimestre in [1],[2],[3],[4])) pvt;
--válido en SQL Server
```

- Se trata de una forma más compacta y legible de representar el resultado.
- No se trata de una relación, sino de un **informe**.
- Las dimensiones se tratan como variables independientes, mientras que las intersecciones, los valores para las variables dependientes correspondientes.

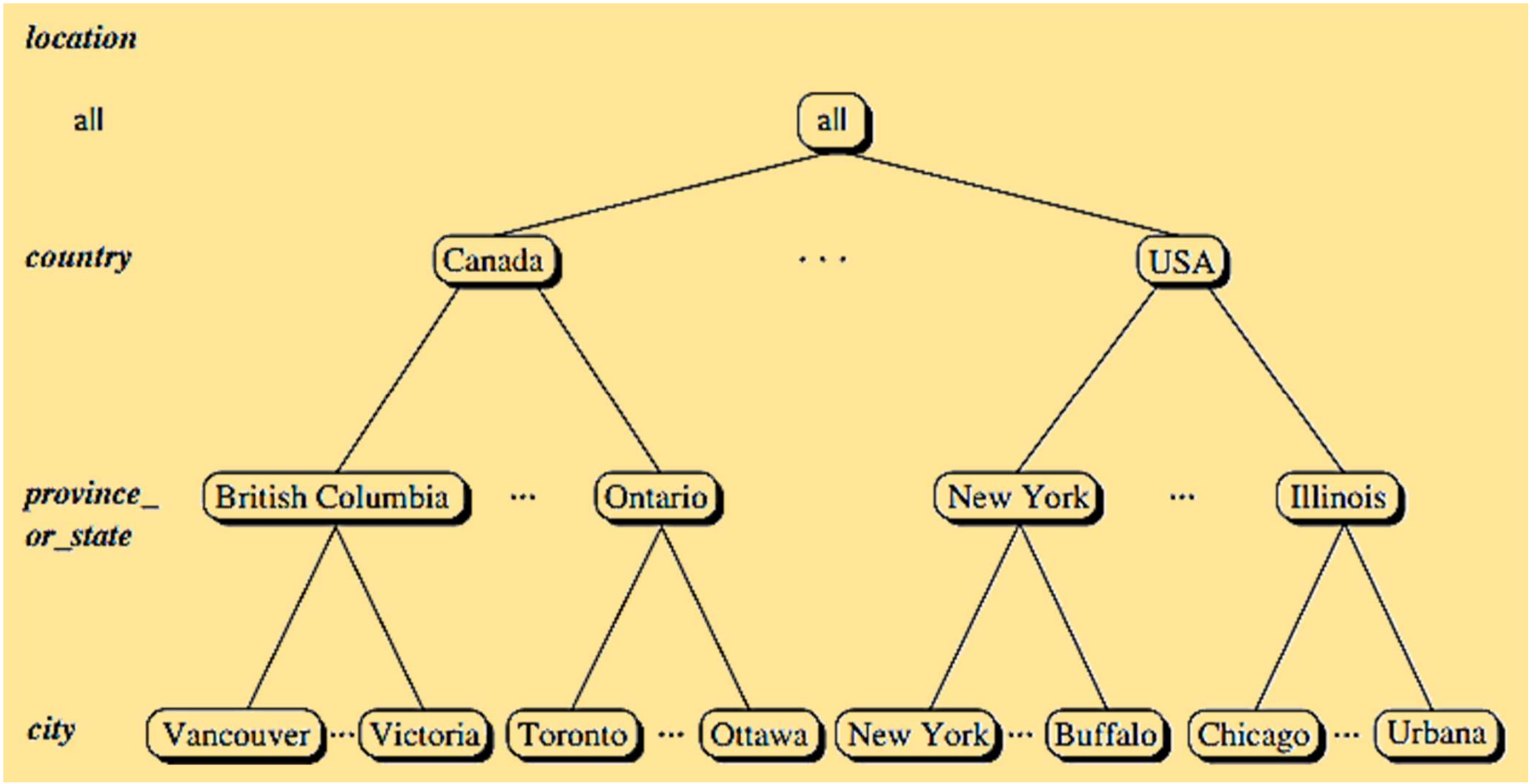
Jerarquía de conceptos

- El **modelo multidimensional** permite representar de una manera muy sencilla **jerarquías:**



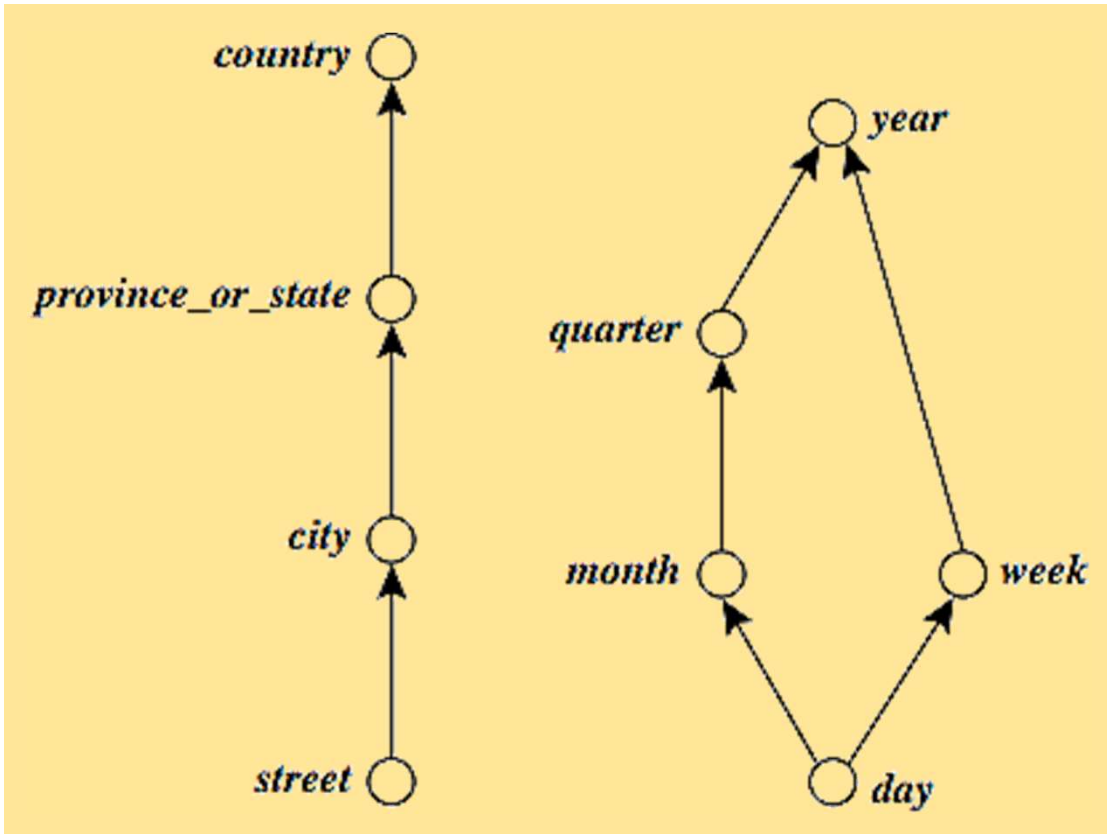
...Jerarquía de conceptos

- Define una secuencia de mapeos que van de un conjunto de **conceptos de bajo nivel** a **conceptos de alto nivel:**



…Jerarquía de conceptos

- Los conceptos pueden relacionarse por medio de relaciones de orden totales o parciales:

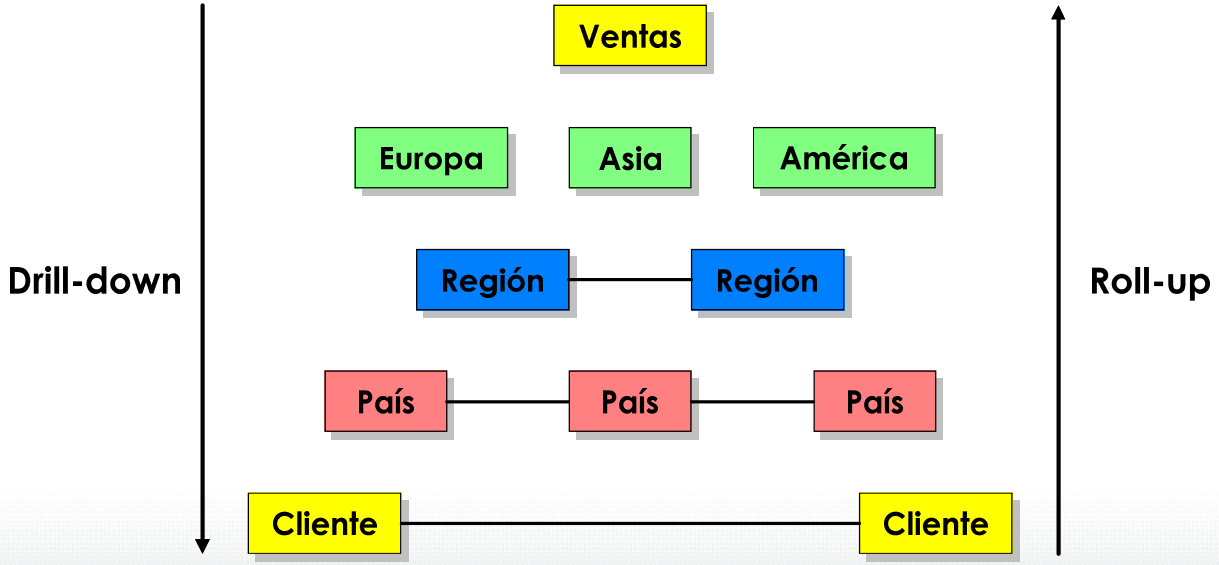


...Jerarquía de conceptos

- Las jerarquías permiten dos tipos de exploraciones:
  - ✓ **Ascendentes** (*roll-up*)

    Permite desplazar la jerarquía hacia arriba, agrupándola en unidades mayores a través de una dimensión, por ejemplo, resumir los datos semanales en trimestrales o anuales.
  - ✓ **Descendentes** (*drill-down*)

    Ofrece la función contraria es decir, de grano más fino; por ejemplo, detallando las ventas del país, por regiones y estás, a su vez, por estados, etc.



Medidas

- Una **medida** en un cubo de datos, es una **función** que puede ser evaluada en cada punto del espacio de datos.
- El valor de la medida es calculada para un punto dado por medio de agregaciones de datos que corresponden a una dimensión en particular.
- Las medidas se pueden organizar en **tres categorías**, dependiendo de la función con la que son calculadas:
  - **Distributivas:** aquellas funciones que pueden ser calculadas de forma distributiva, por ejemplo, **sum()**, **count()**, **min()** y **max()**.
  - **Algebraicas:** aquellas funcionen que pueden ser calculadas por una ***función algebraica*** con ***M argumentos***, cada uno de los cuales se obtiene de aplicar funciones de agregación distributivas, por ejemplo, **avg()**.
  - **Holísticas:** una función de agregación es ***holística*** si no existe una función algebraica con M argumentos que caracterice su cálculo, por ejemplo, **median(), mode(), rank()**