UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
FACULTAD DE CIENCIAS

DIPLOMADO EN MINERÍA DE DATOS

Módulo 3. *Procesamiento analítico de datos*

Procesamiento analítico en línea

Gerardo Avilés Rosas
gar@ciencias.unam.mx

OLTP vs. OLAP

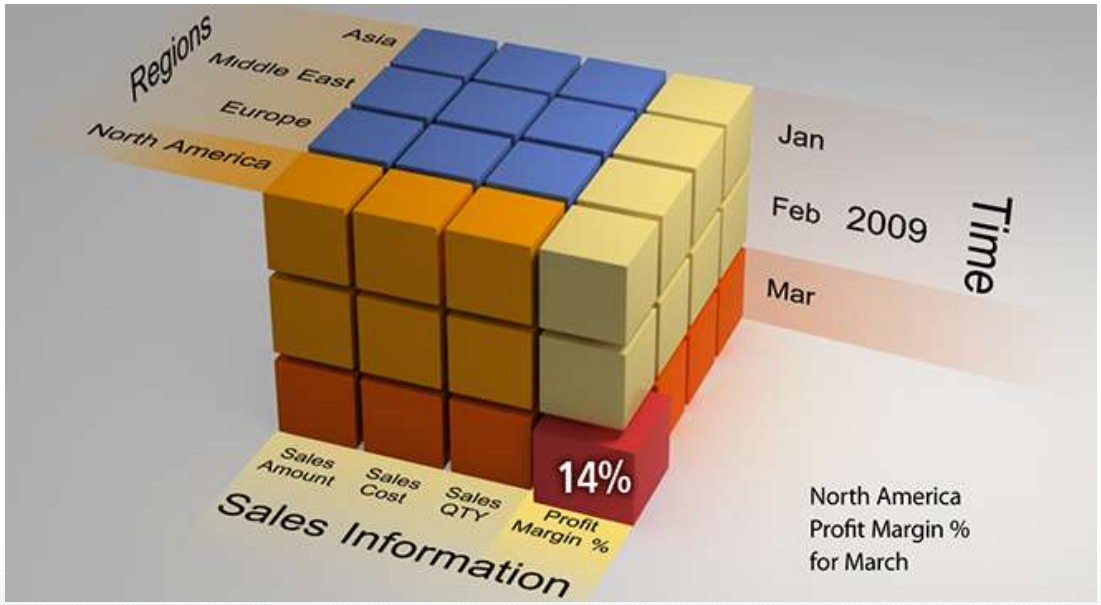
Procesamiento de transacciones en línea (OLTP)

- **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:

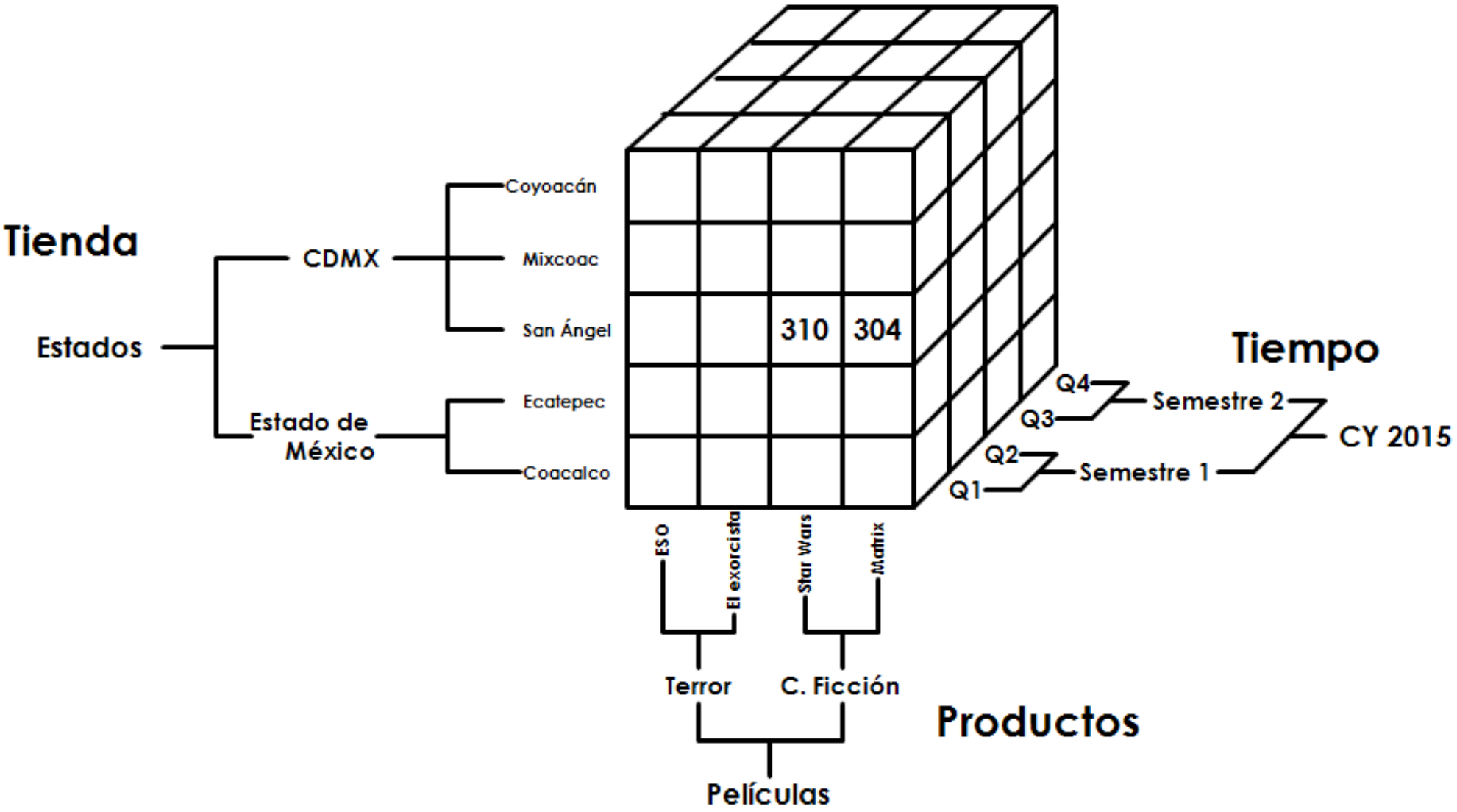
  *Analizar las **ventas** diaria, mensual o anualmente, para un cliente, una región o un país.*

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

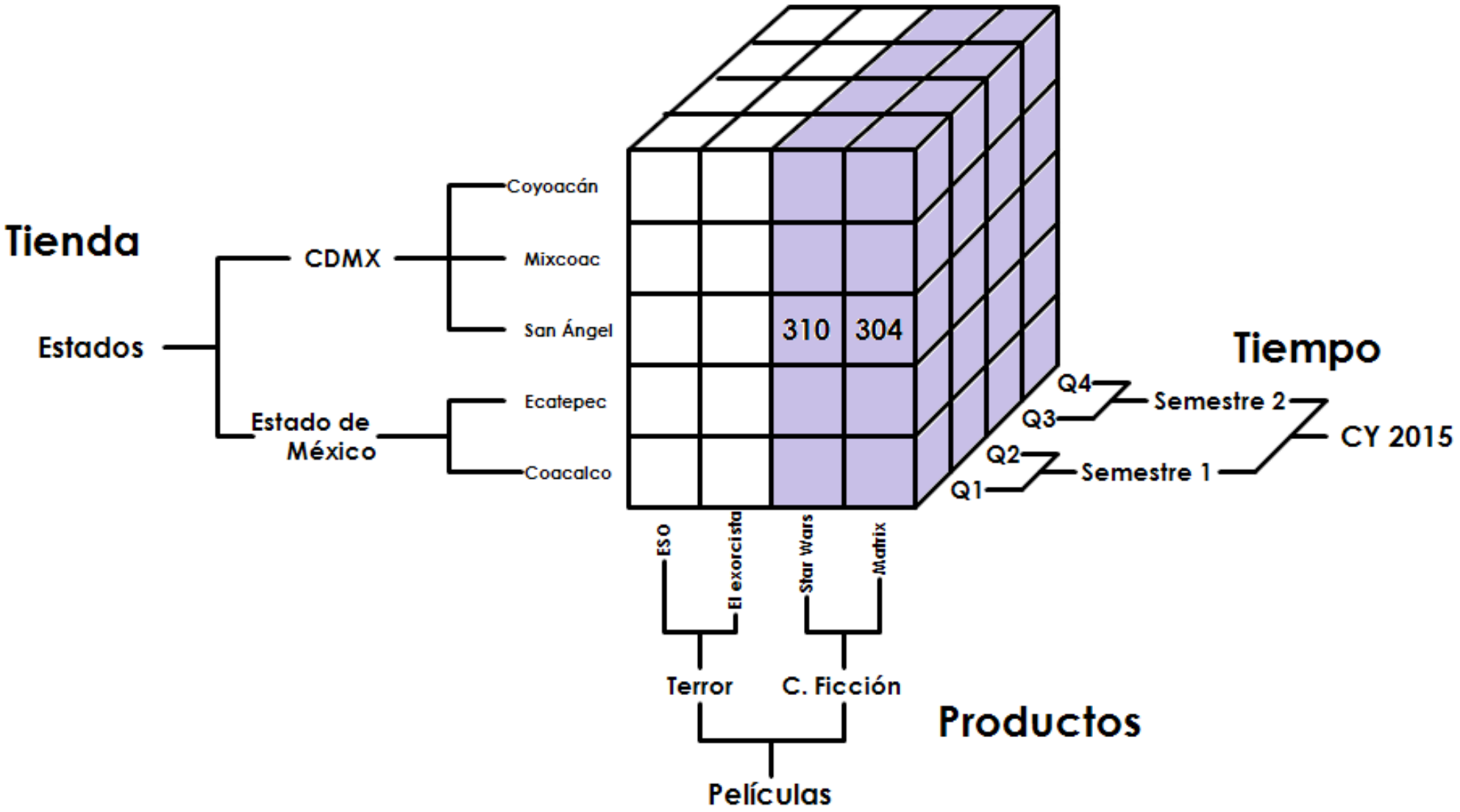
  *Permiten cambiar el enfoque del **¿qué esta pasando?** (enfoque relacional) al **¿por qué esta pasando?** (enfoque multidimensional)*.

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

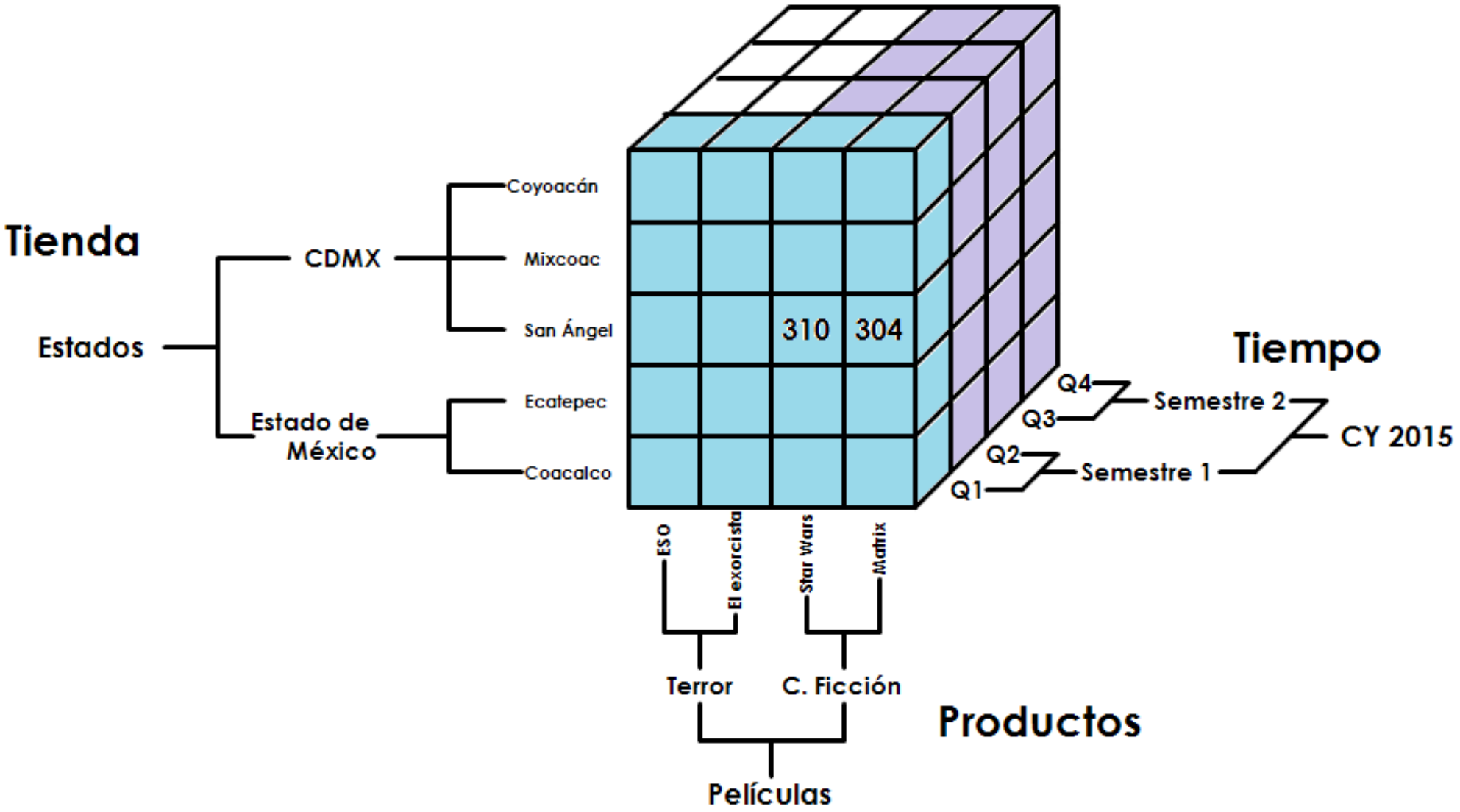
...Cubos OLAP


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

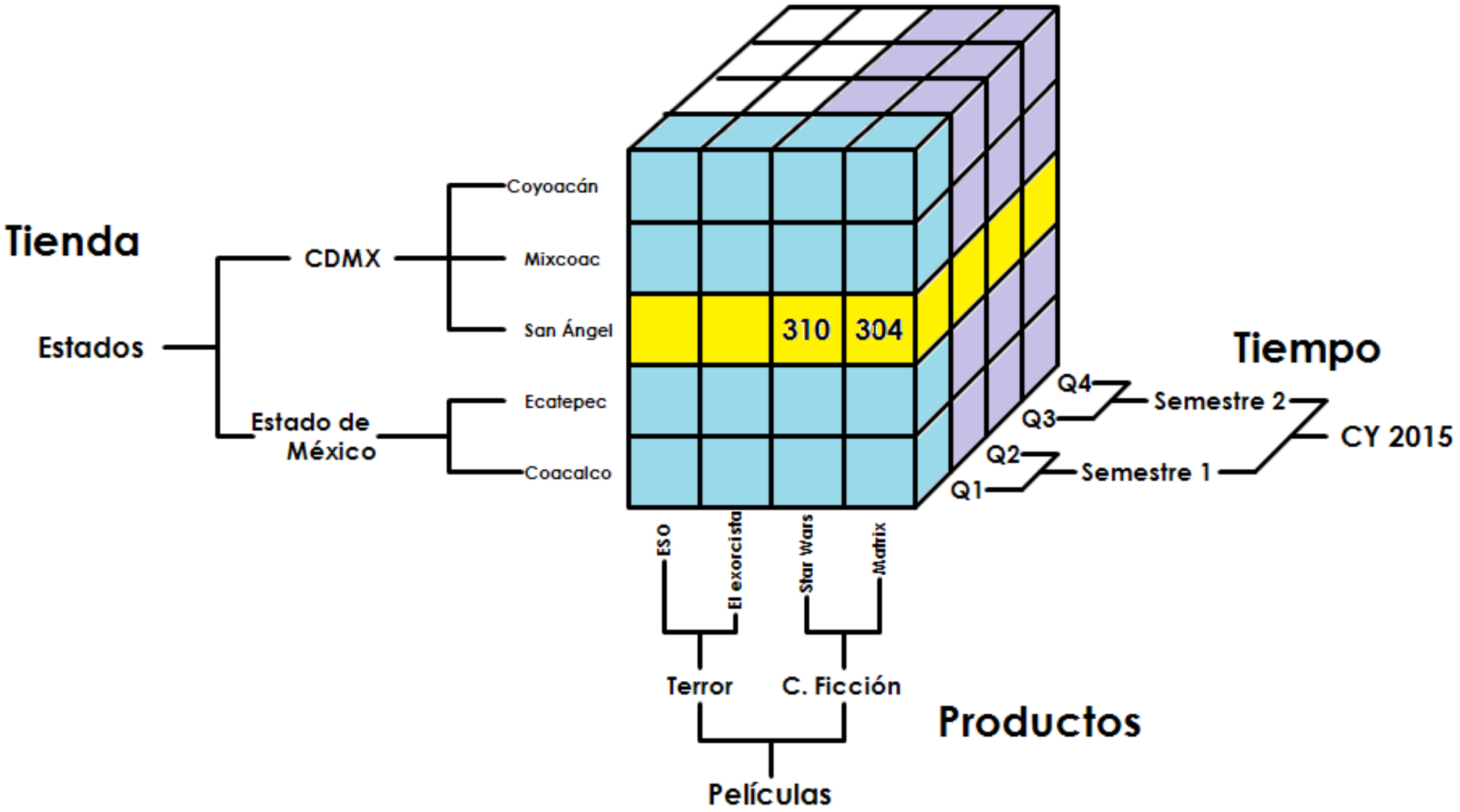
...Cubos OLAP


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

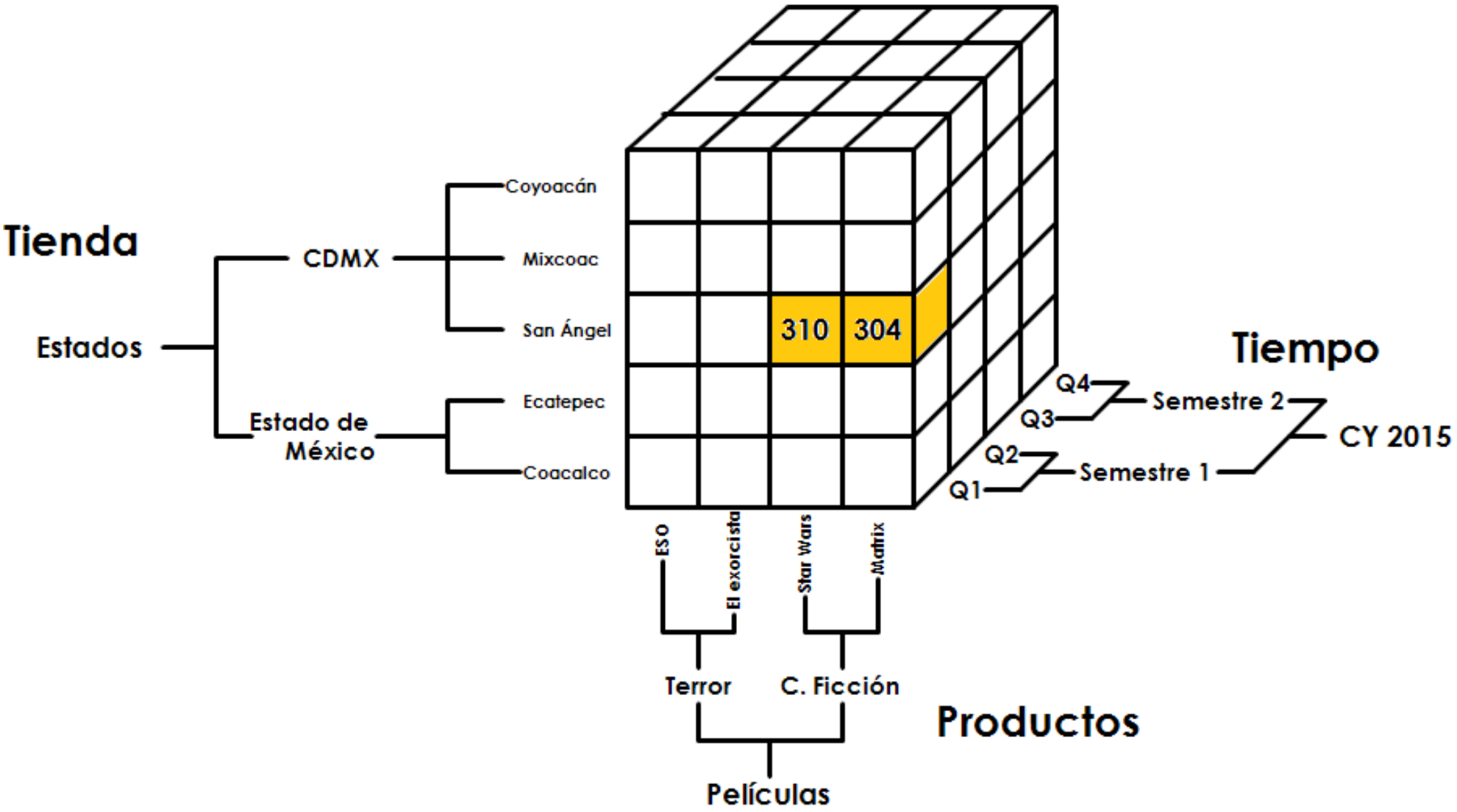
...Cubos OLAP


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

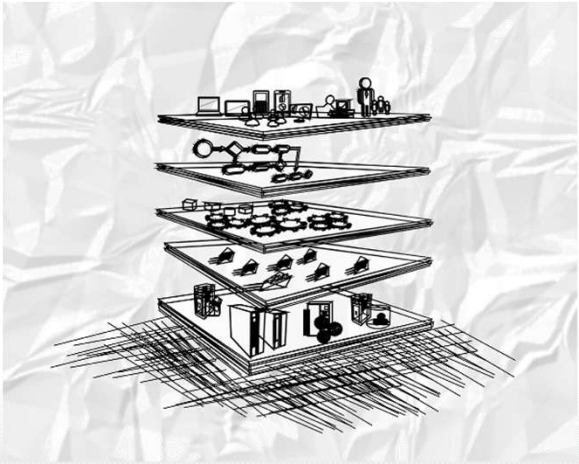
El uso de **cubos OLAP** tiene dos ventajas fundamentales:

- **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

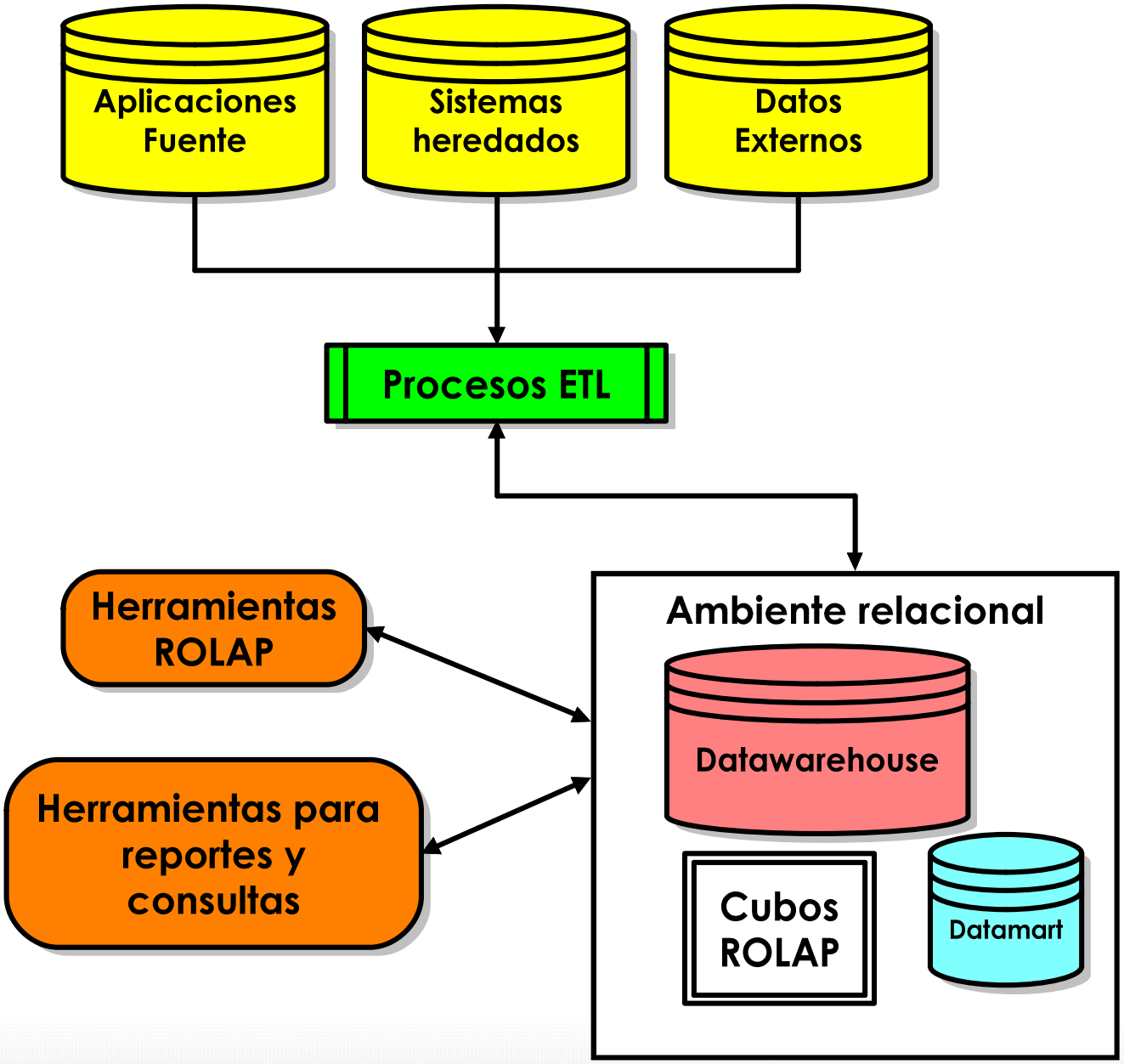
- El cubo es estructura adicional de datos que se debe **mantener** y en algunos caso **actualizar** (esto supone un gasto extra de recursos: *servidores, discos, procesos de carga, etc.*)
- El modelo de negocio no siempre se adapta bien en un modelo basado en jerarquías, por ejemplo:
  - *Una semana 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

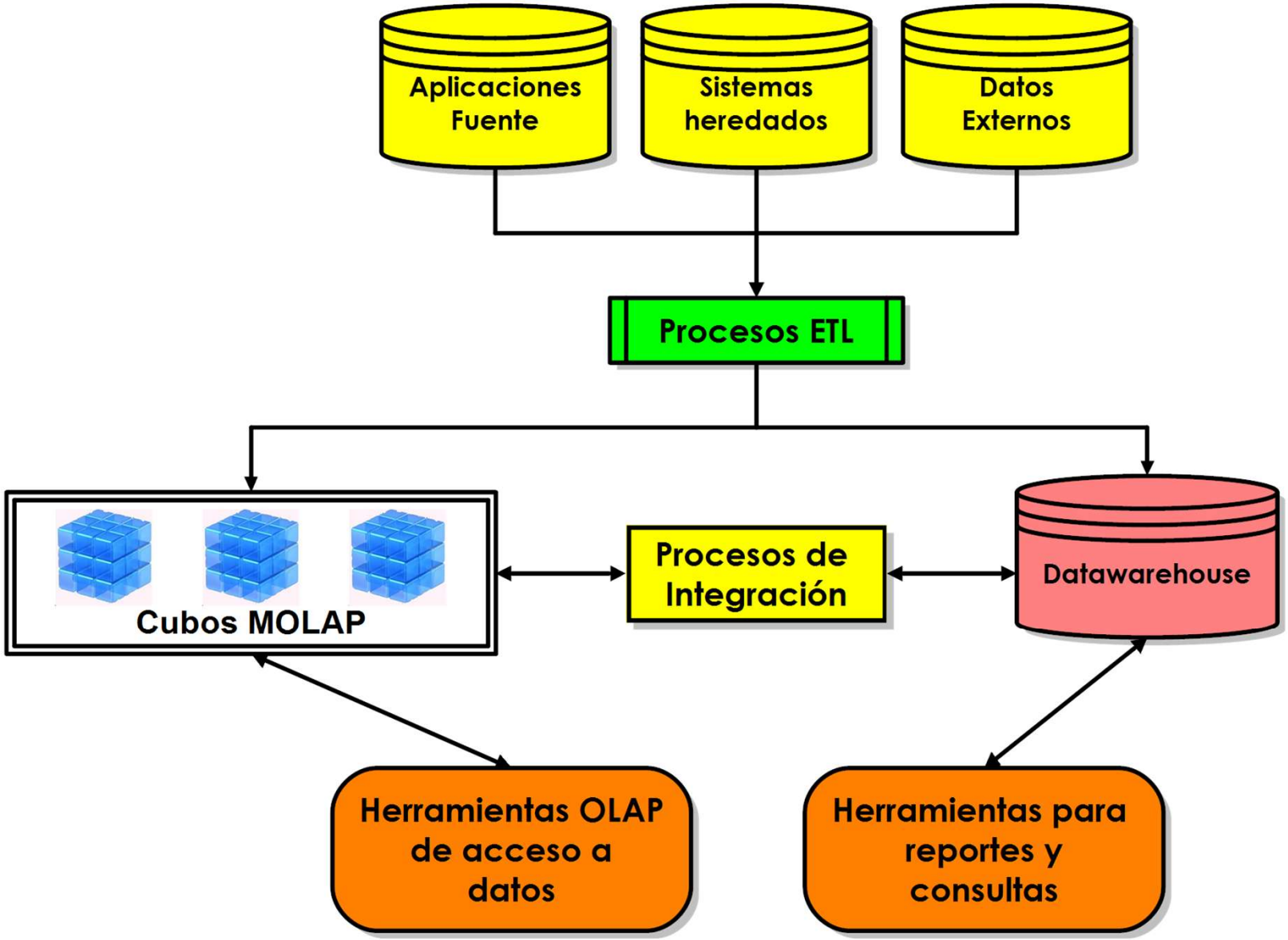
- Los datos se cargan desde el sistema operacional y se crean **índices** para optimizar los tiempos de acceso a las consultas.
- Los análisis multidimensionales se transforman dinámicamente a **consultas SQL**. Los resultados se relacionan mediante **tablas cruzadas** y conjuntos multidimensionales.
- 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

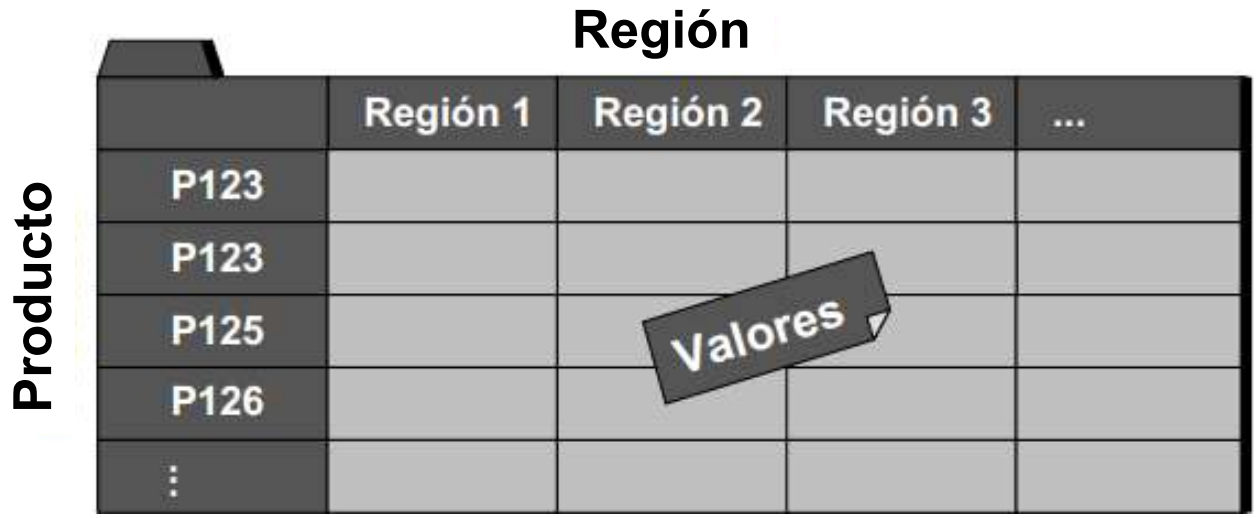


…Implementación: MOLAP

- 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*.

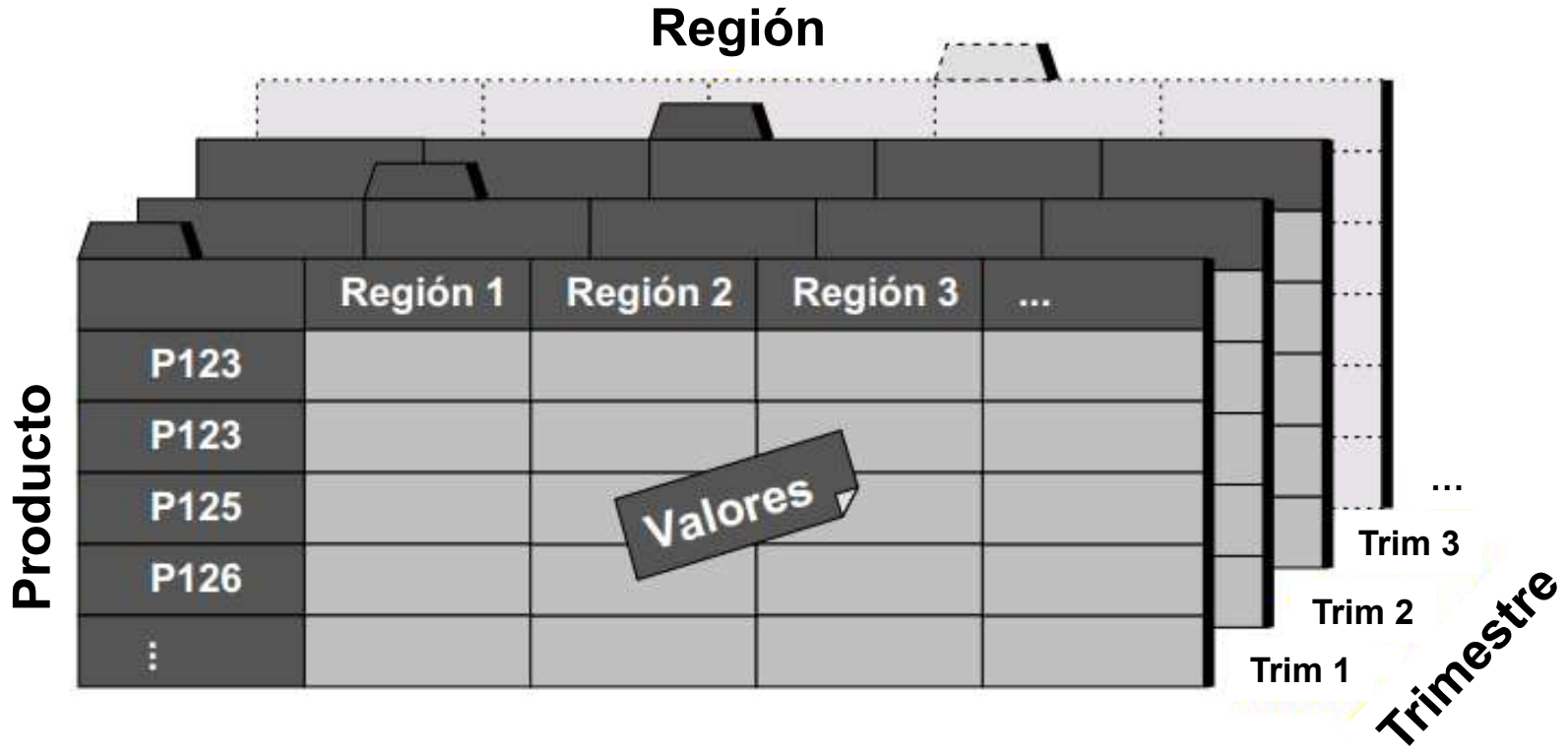
Modelo de datos

- 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

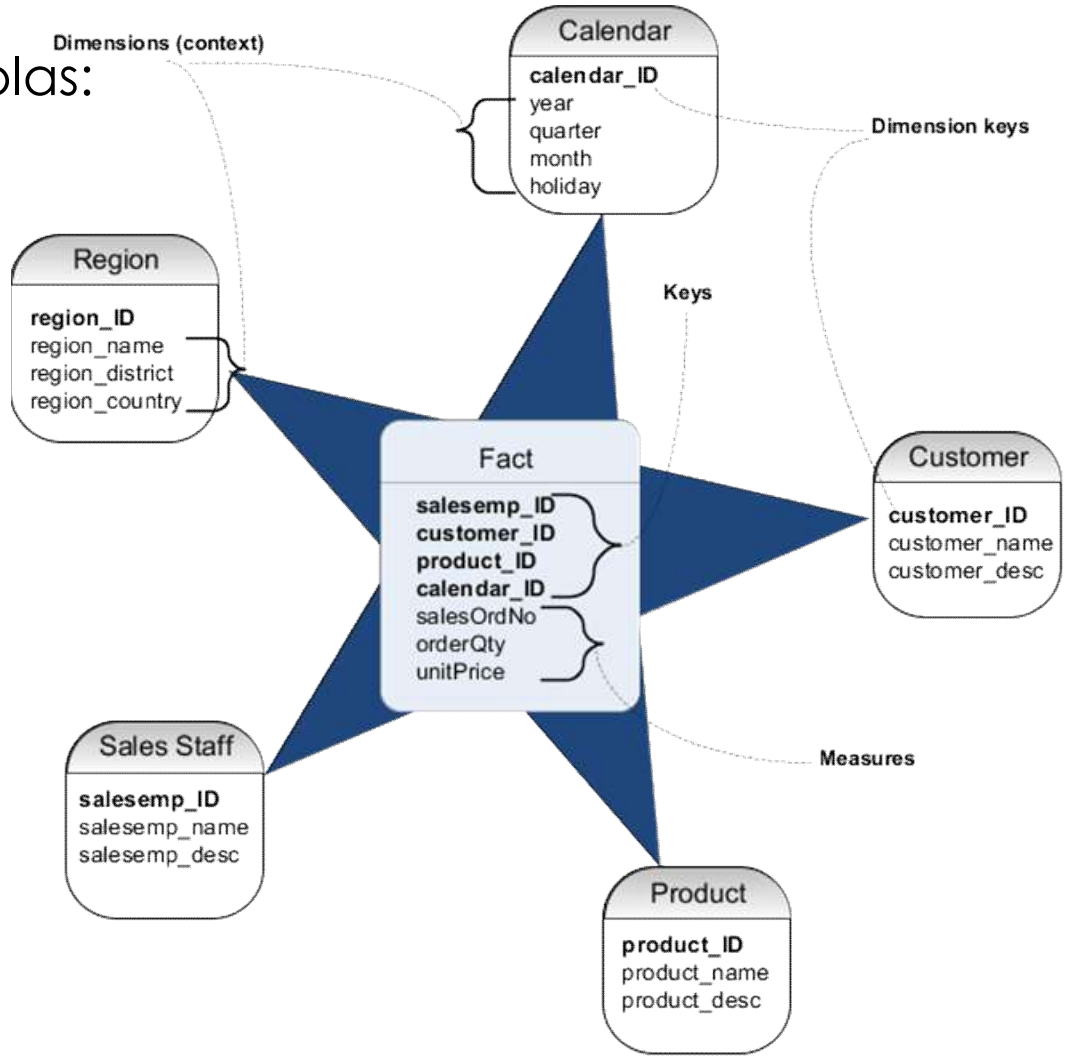
- Siguiendo con el mismo ejemplo, si añadimos una dimensión más, tendríamos una **matriz de tres dimensiones**:



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

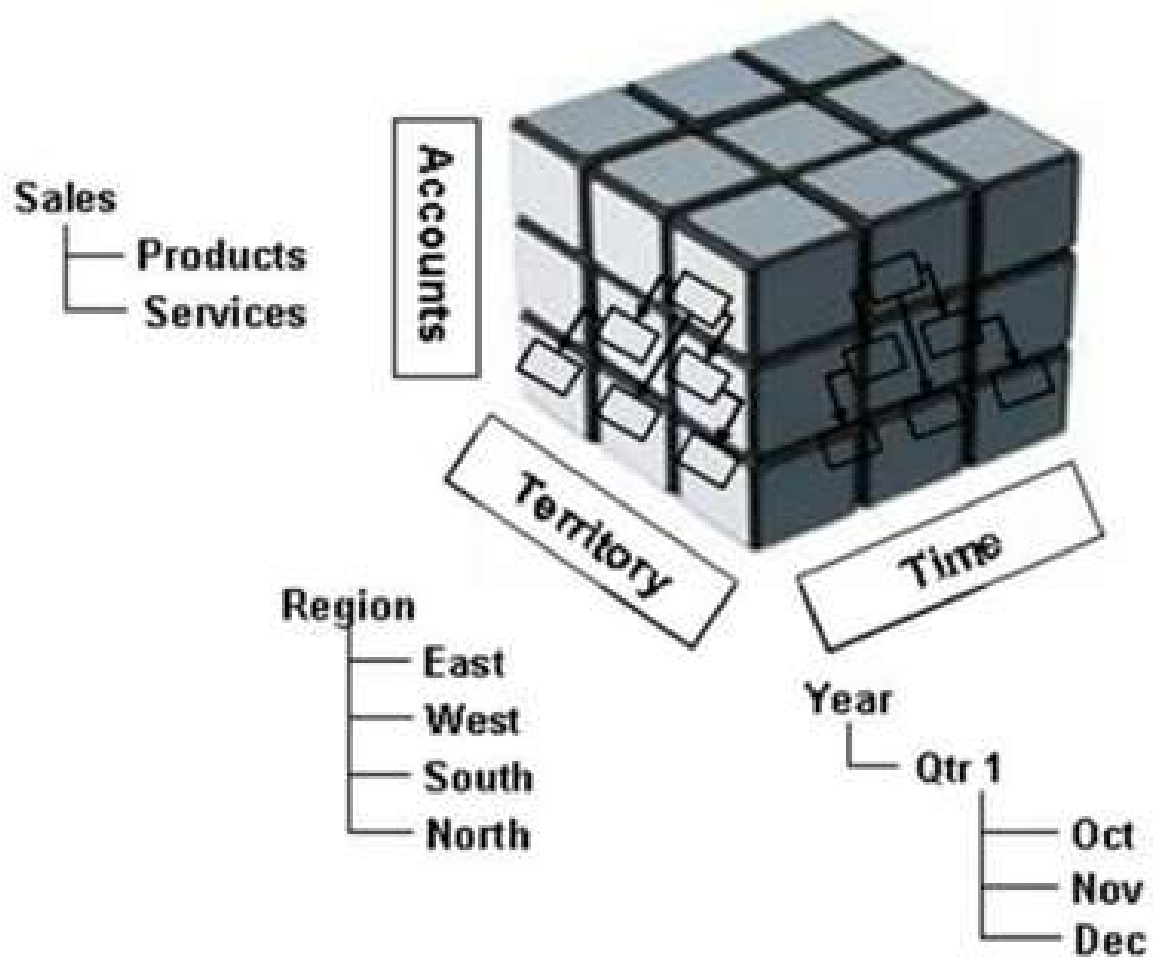
…Modelo de datos

- 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**.



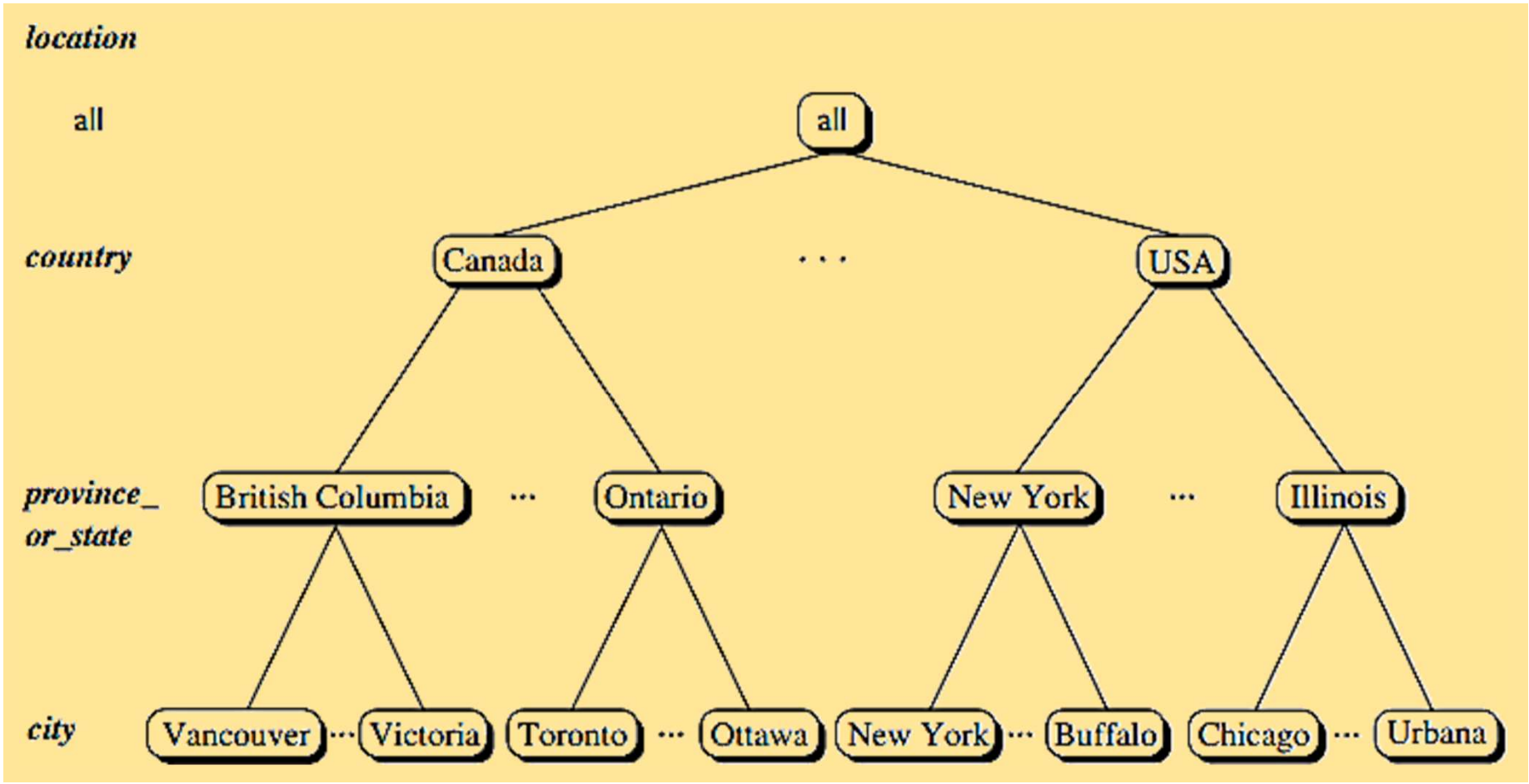
Jerarquía de conceptos

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



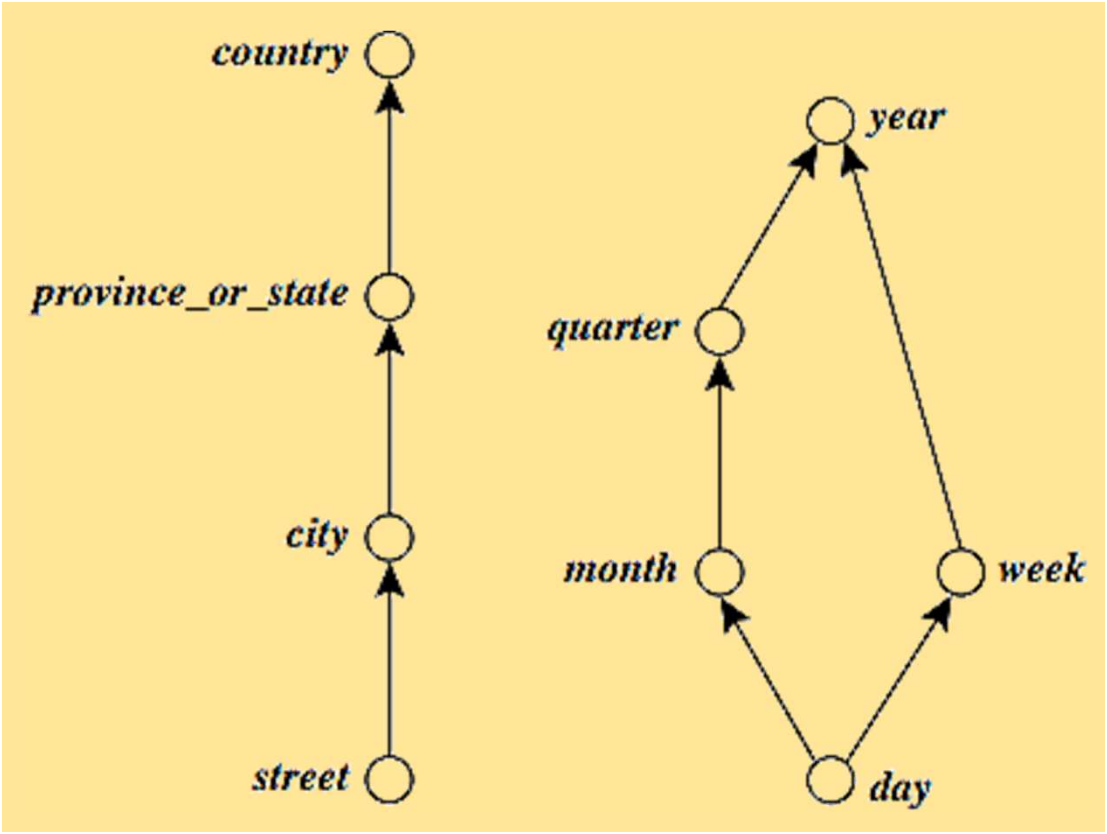
…Jerarquía de conceptos

- 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:

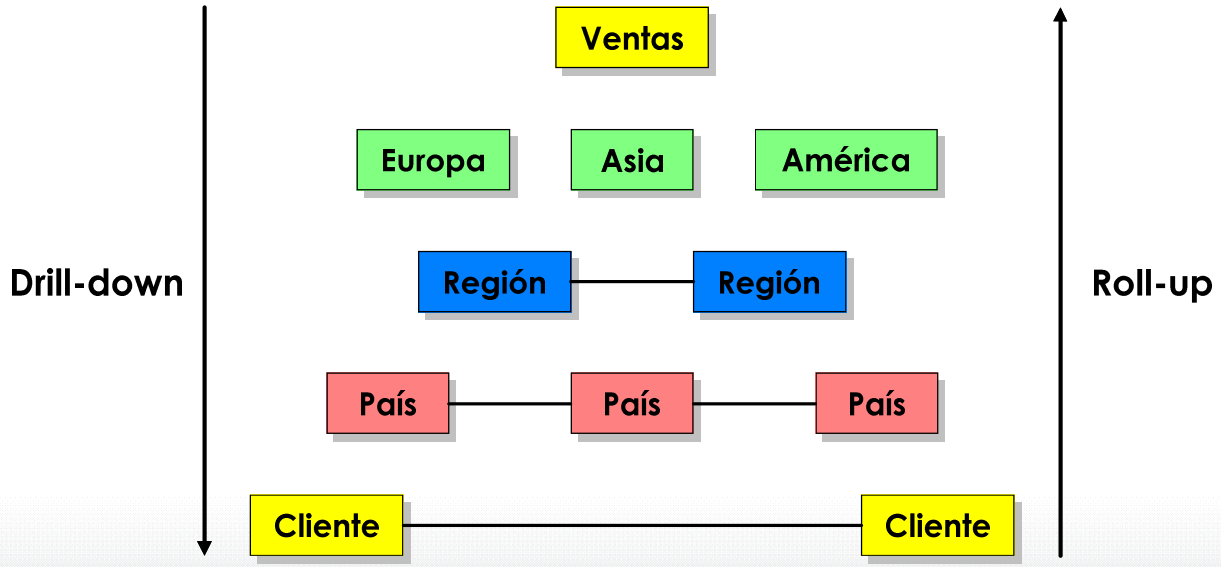


…Jerarquía de conceptos

- 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.




¡Gracias!