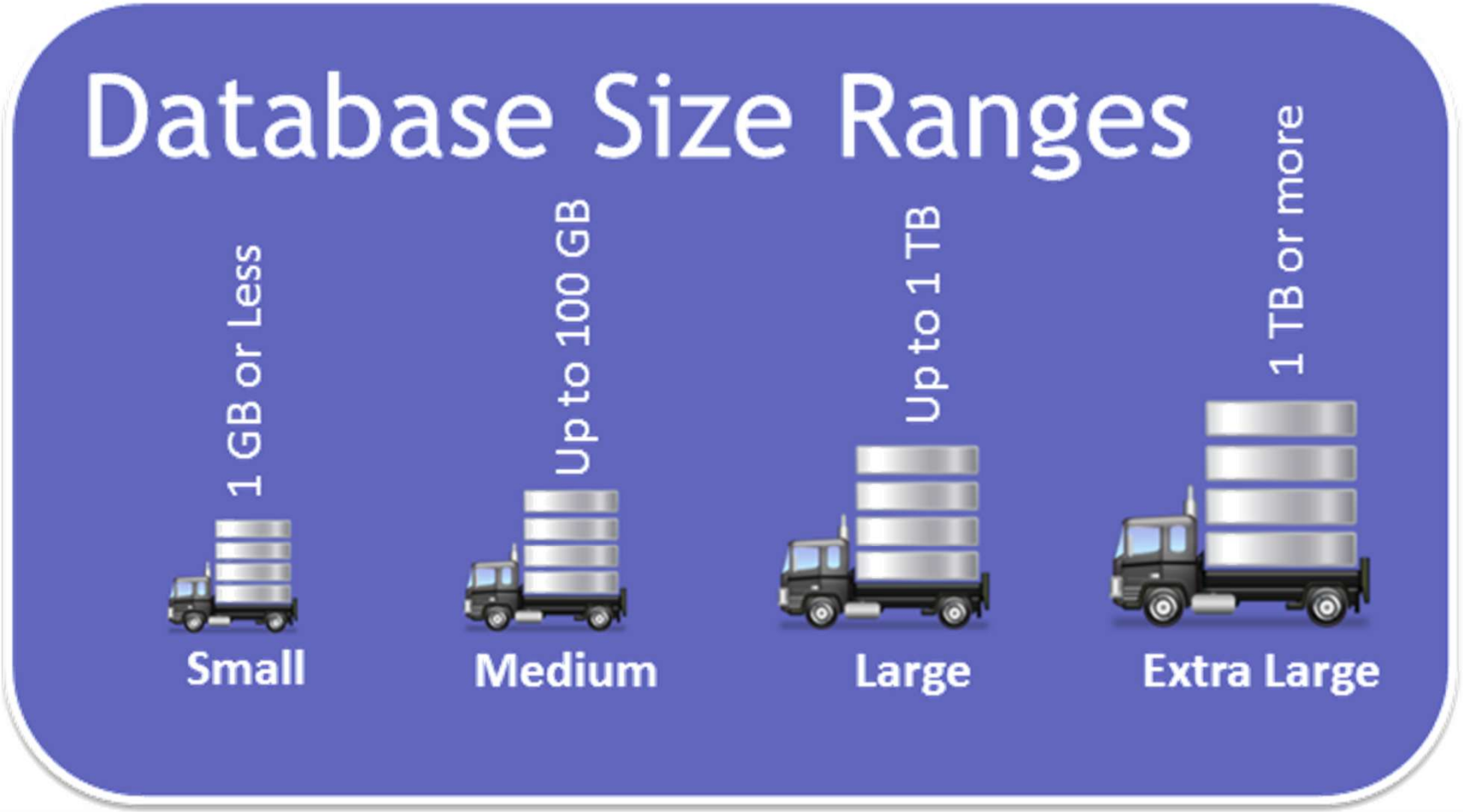
UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
FACULTAD DE CIENCIAS
FUNDAMENTOS DE BASES DE DATOS

Introducción a la Inteligencia de Negocios

Gerardo Avilés Rosas
gar@ciencias.unam.mx

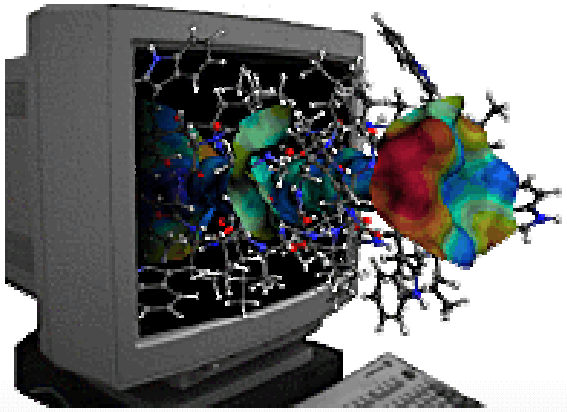
Introducción

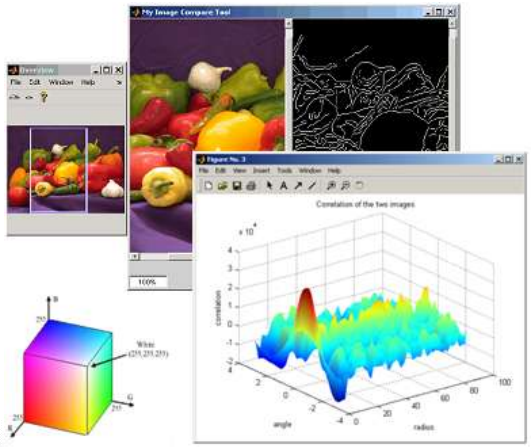
Cada día crece, en forma espectacular, la cantidad de datos que se generados y se registran:



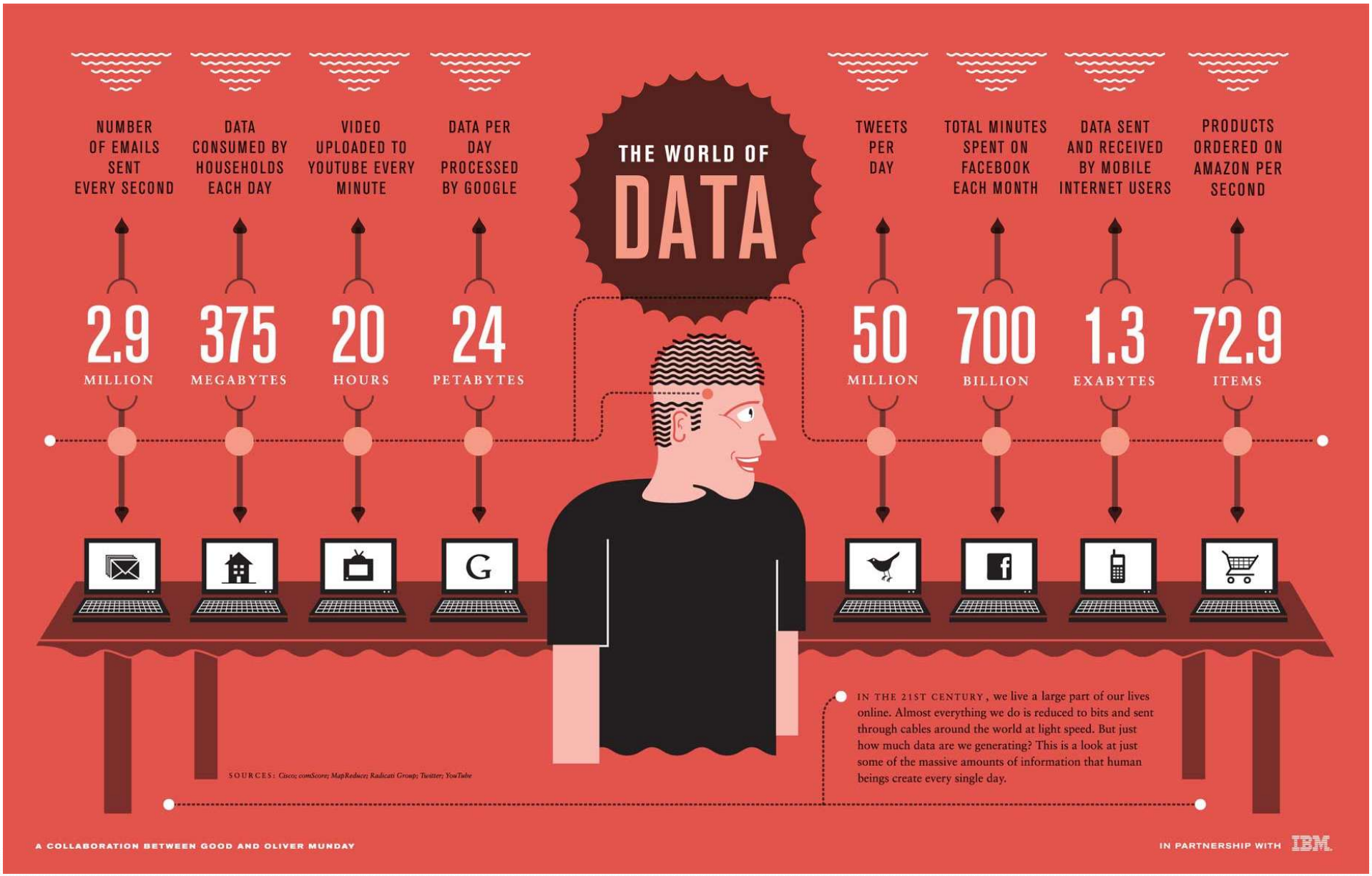
...Introducción

Fuentes de datos





...Introducción



…Introducción

Conclusión: estamos ahogados en datos…

…Introducción

Las aplicaciones sobre **Bases de Datos** son muy importantes para la vida de una **organización**:

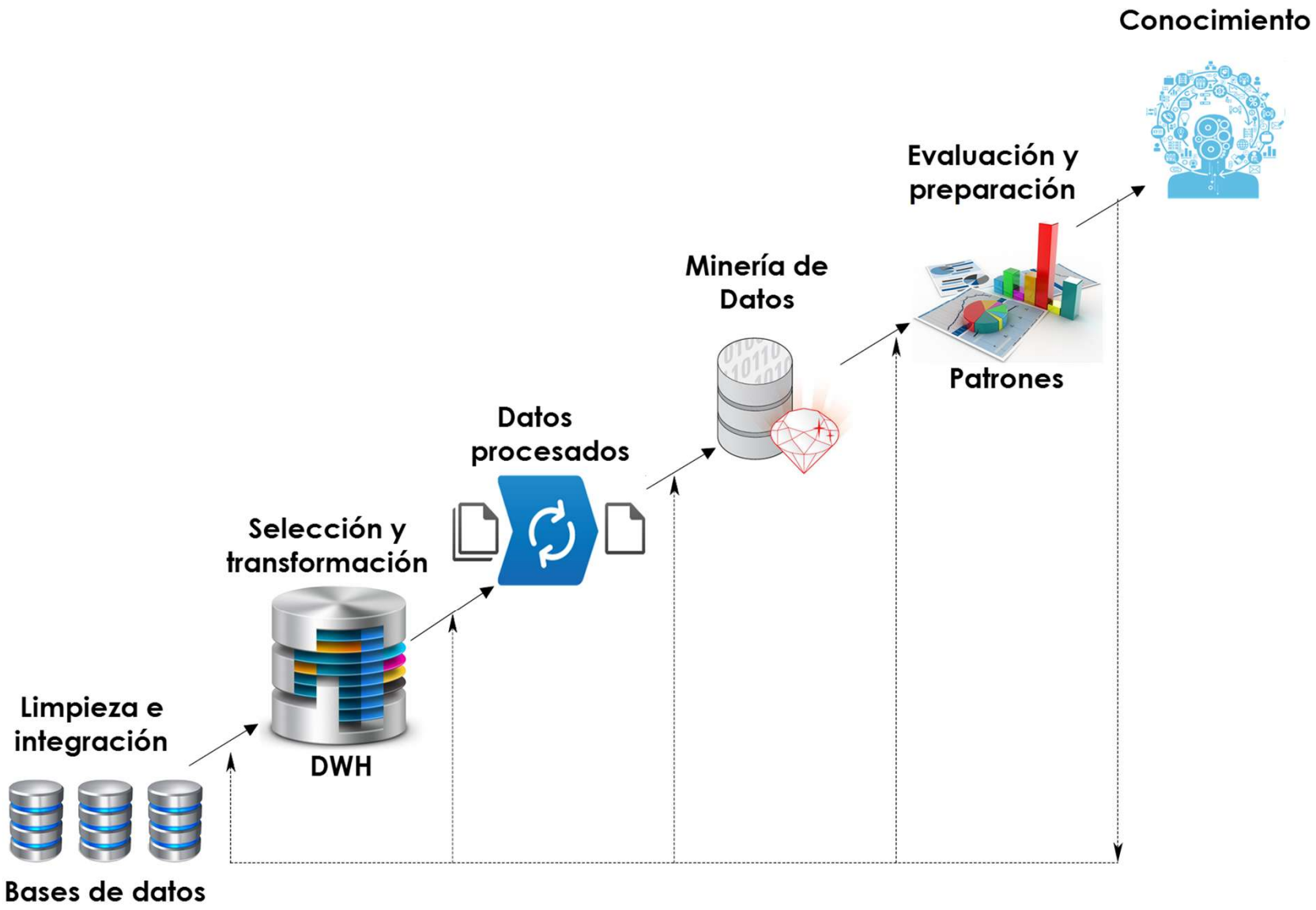
- Soportan las operaciones del **día a día** de los negocios.
- **Reúnen, almacenan y procesan** todos los datos necesarios para la ejecución exitosa de las operaciones diarias rutinarias.
- Proporcionan información **en línea** y producen **reportes** para monitorear y realizar los negocios.
- Se han desarrollado métodos eficientes para el proceso de **transacciones en línea** (*OLTP*), donde una consulta se ve como una transacción de solo lectura.

Sin estos sistemas de cómputo, los negocios **no pueden sobrevivir**.

…Introducción

- Al expandirse los negocios, **la complejidad de estos crece**; los ejecutivos requieren información para ser competitivos y mejorar su línea de producción.
- En los **90s** se empieza a tomar ventaja competitiva con la construcción de sistemas de **almacenamiento de datos** (*DWH*).
- El **almacén de datos** es visto como un **repositorio** de **fuentes de datos heterogéneas** bajo un esquema uniforme en un **solo sitio**, que facilita a los ejecutivos la toma de decisiones:
  - ✓ *Incluye **limpieza de datos**, **integración de datos** y **procesamiento analítico en línea** (OLAP), **técnicas de análisis** (resúmenes, consolidación y agregación) así como la habilidad de ver la información desde diferentes ángulos.*
- La **minería de datos** es la **extracción** o **minado** de conocimiento de grandes volúmenes de datos.
  - ✓ *Se trata de un proceso que intenta **descubrir patrones** (útiles, inesperados) en **grandes volúmenes** de datos.*

Knowledge Discovery in Databases



Información estratégica

¿ Qué es información estratégica?

- *Es la información que permite direccionar una organización hacia el logro de objetivos de negocio*.

¿Quién necesita información estratégica?

- *Los responsables de mantener la competitividad de una empresa.*
- **Ejemplos de objetivos de negocios:**
  - ✓ Conservar su clientela base
  - ✓ Aumentar su clientela en x% en los n años siguientes
  - ✓ Mejorar los niveles de calidad de sus principales productos
  - ✓ Incrementar sus ventas x% en cierta región, etc.
  - ✓ Mejorar el servicio del cliente en…
  - ✓ etc.

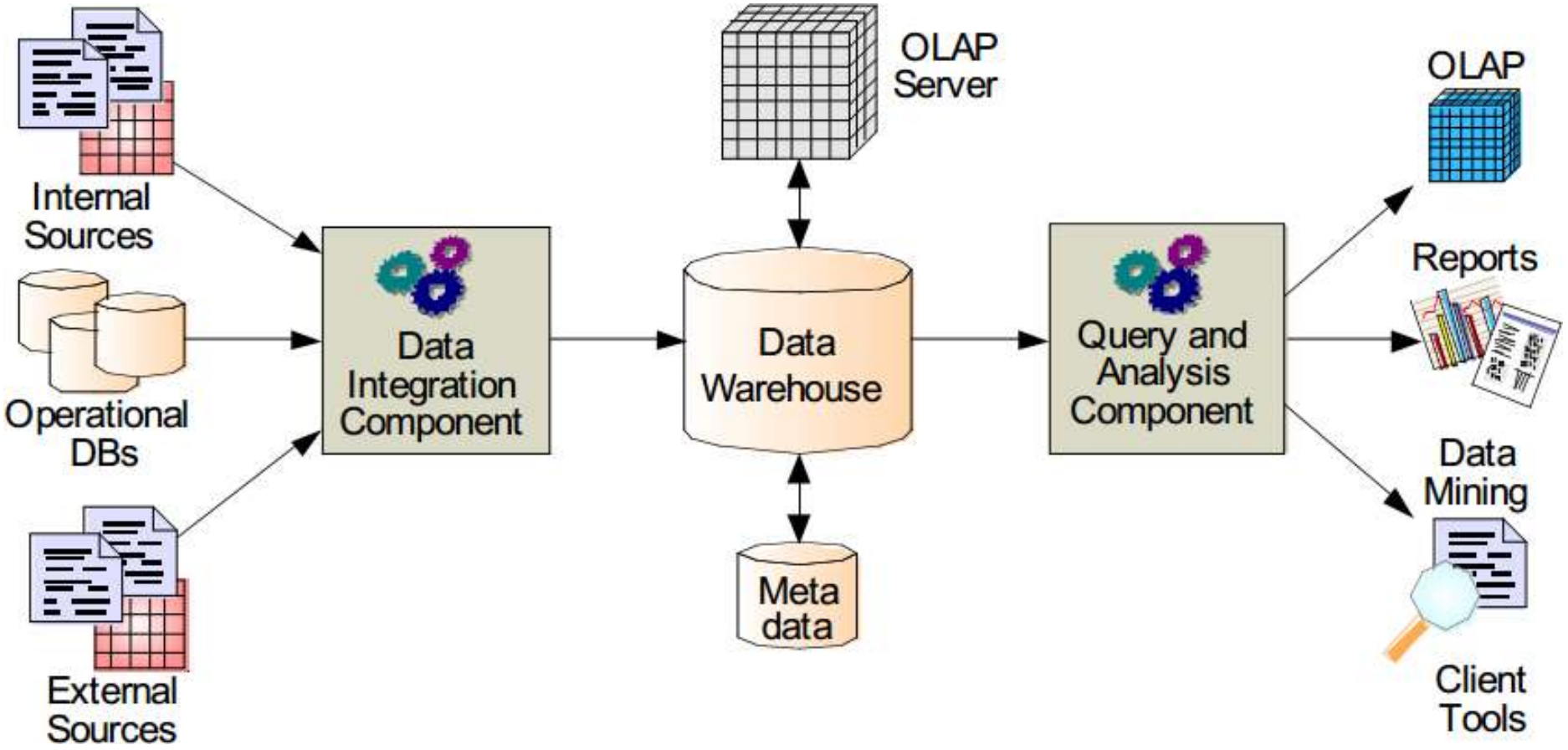
…Información estratégica

- La **información estratégica** no pretende producir una factura, hacer un envío, etc., es más importante para la **salud y supervivencia** de la corporación.
- Las **decisiones críticas** dependen de la información estratégica apropiada de una empresa.
- Características deseadas en la información estratégica:
  - ✓ *Integrada*
  - ✓ *Integridad de datos*
  - ✓ *Accesible*
  - ✓ *Creíble*
  - ✓ *A tiempo*

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 a los usuarios analizar información de bases de datos proveniente de múltiples sistemas 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.

…Online Analytic Processing

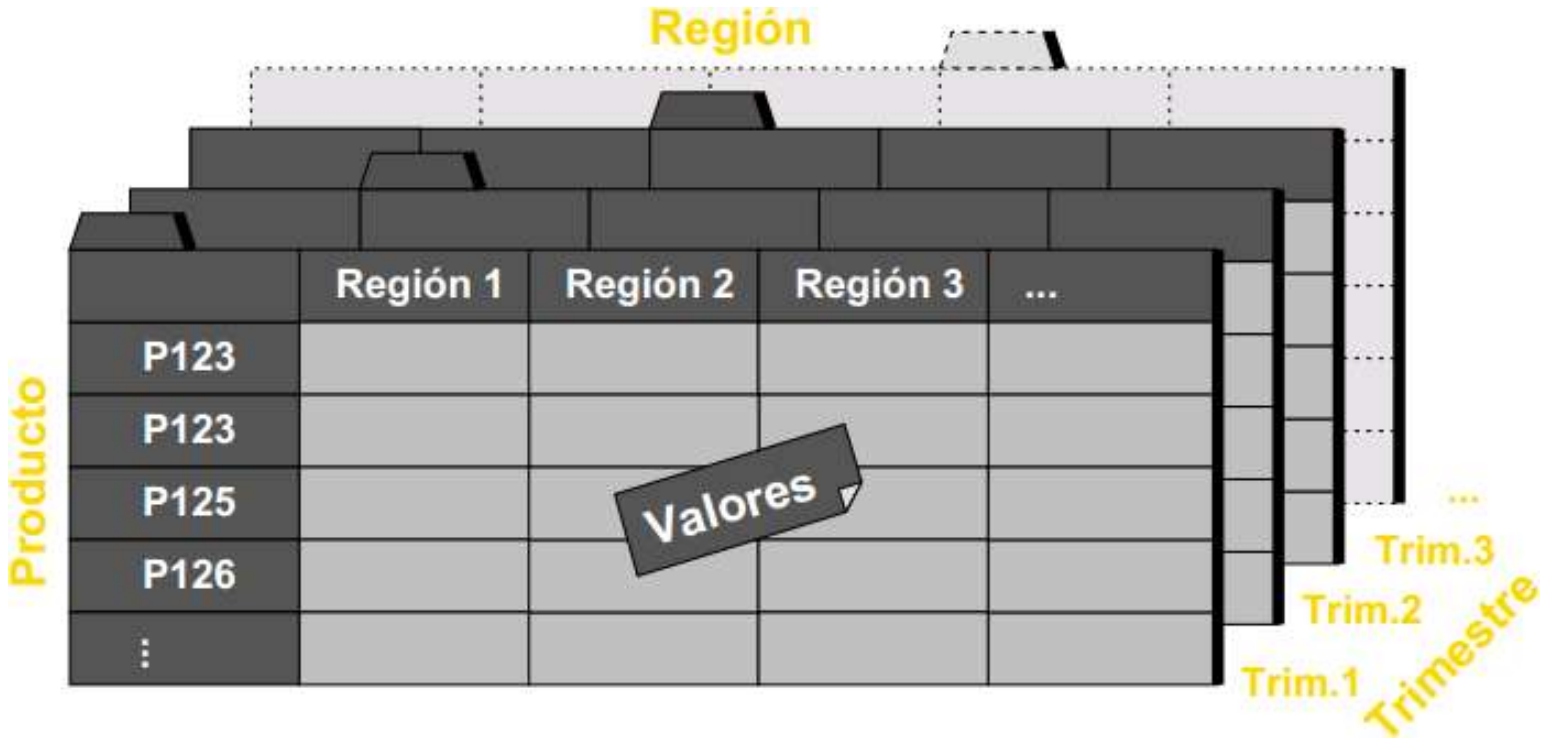


Modelo de datos

- Es un conjunto de conceptos que pueden usarse para describir la estructura de un **DWH**.
- 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

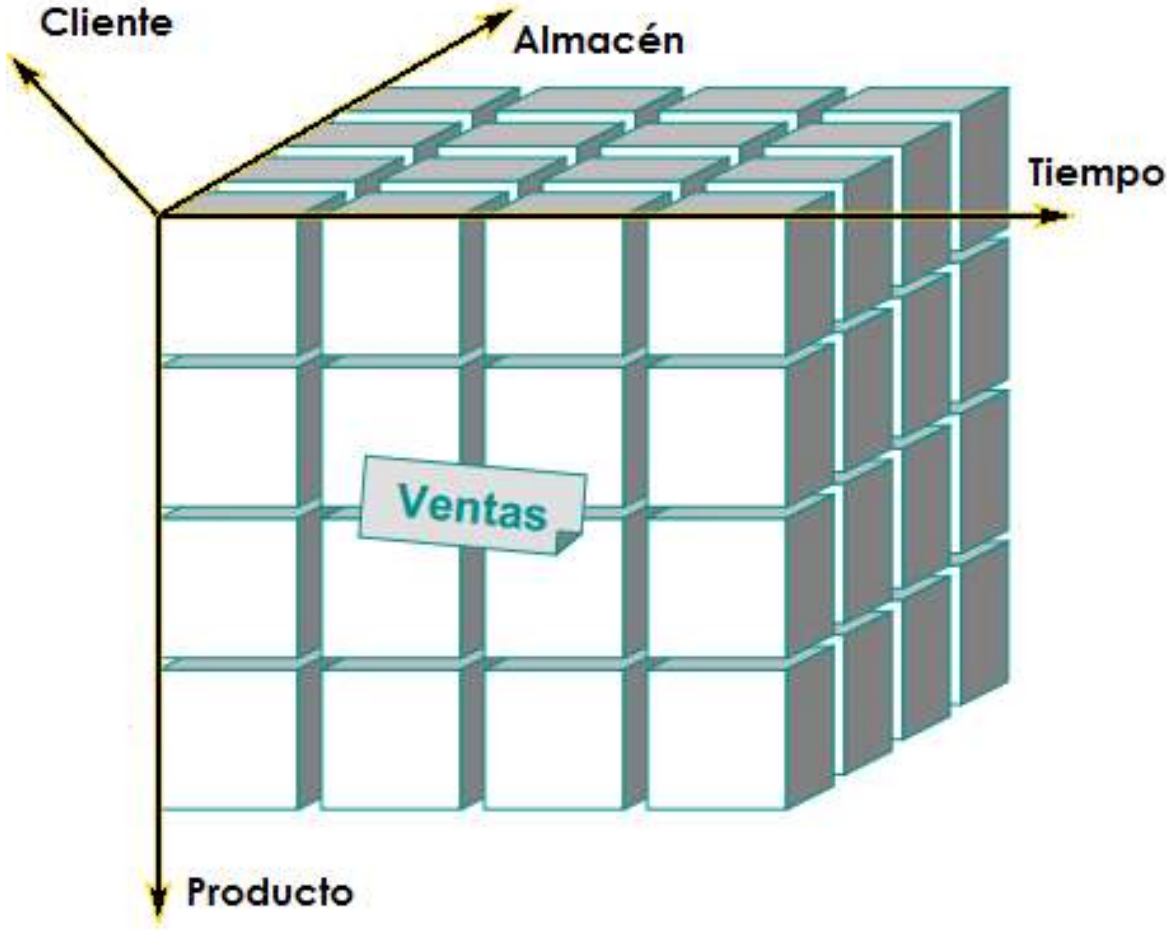
- Siguiendo con el mismo ejemplo, si añadimos una dimensión más, tendríamos una matriz de tres dimensiones:



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

…Modelo de datos

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

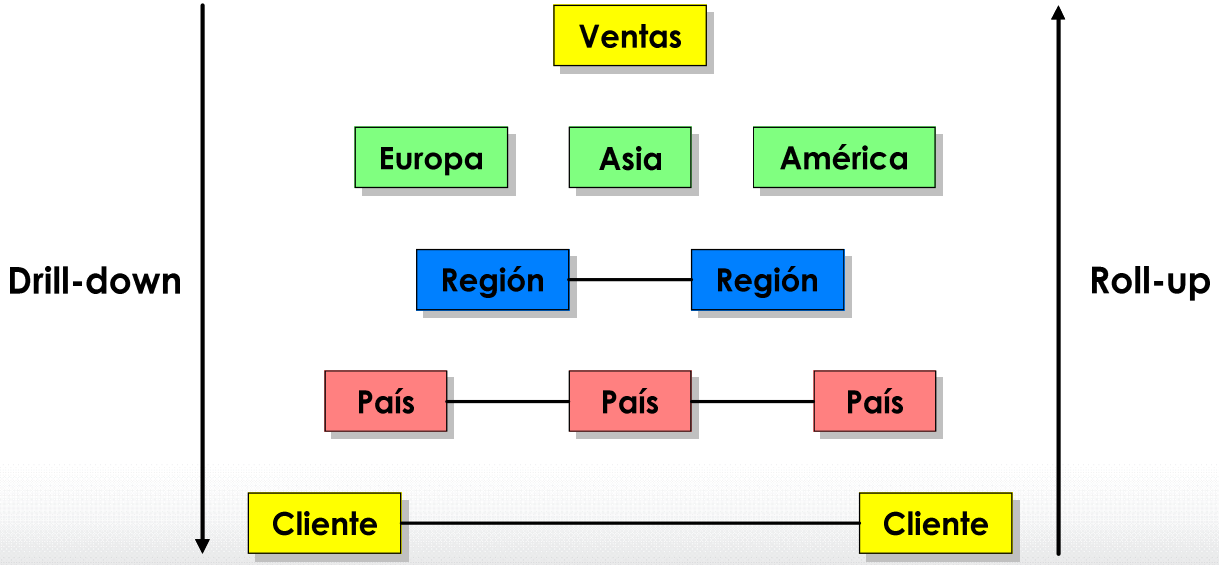


…Modelo de datos

- 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.

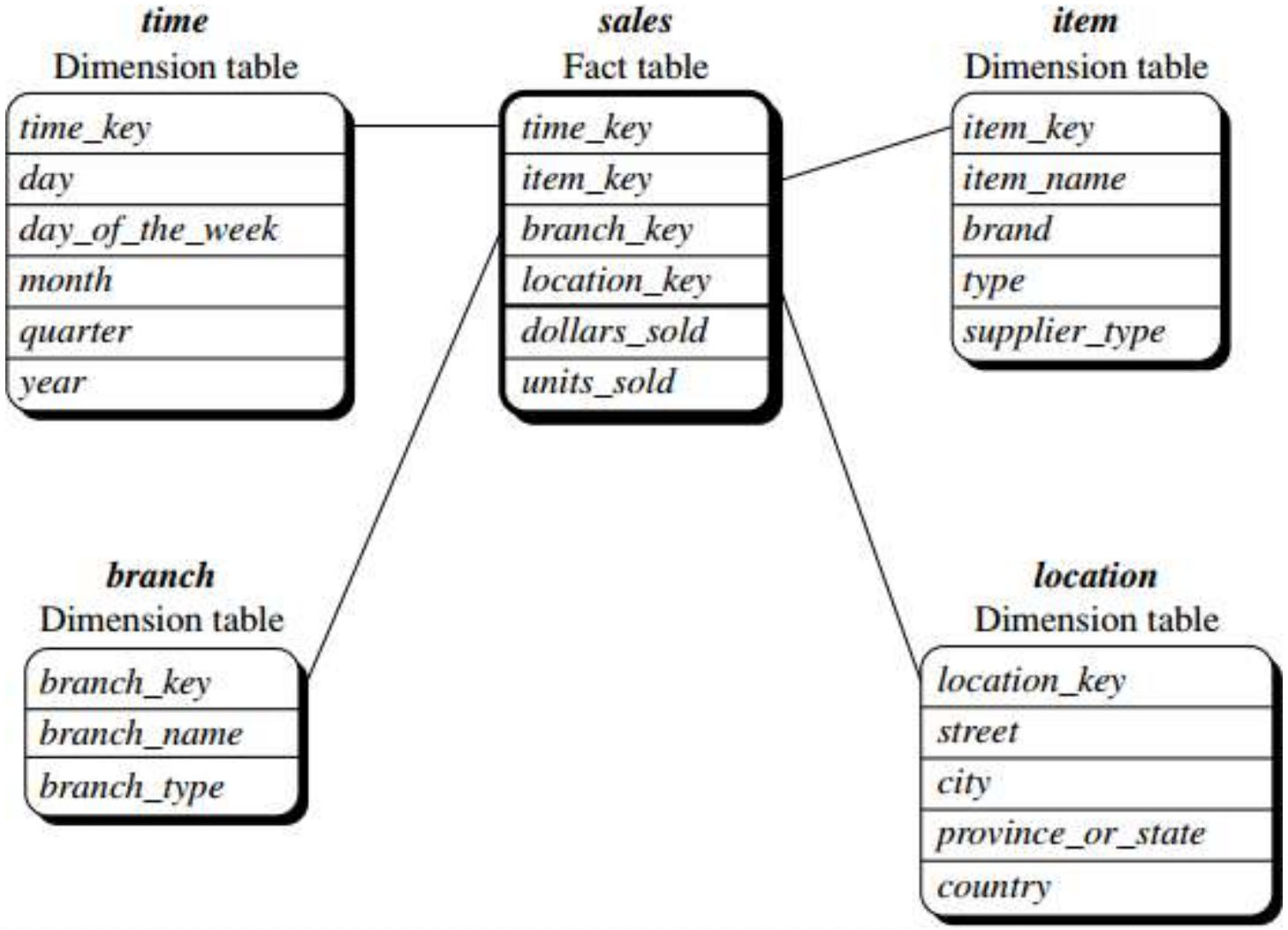


…Modelo de datos

- 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 de dimensión.
  - ✓ Una **tabla de hecho**, 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.
- De esta forma, las tablas de hecho contiene los datos y las de dimensiones identifican a esos datos por medio de tuplas.

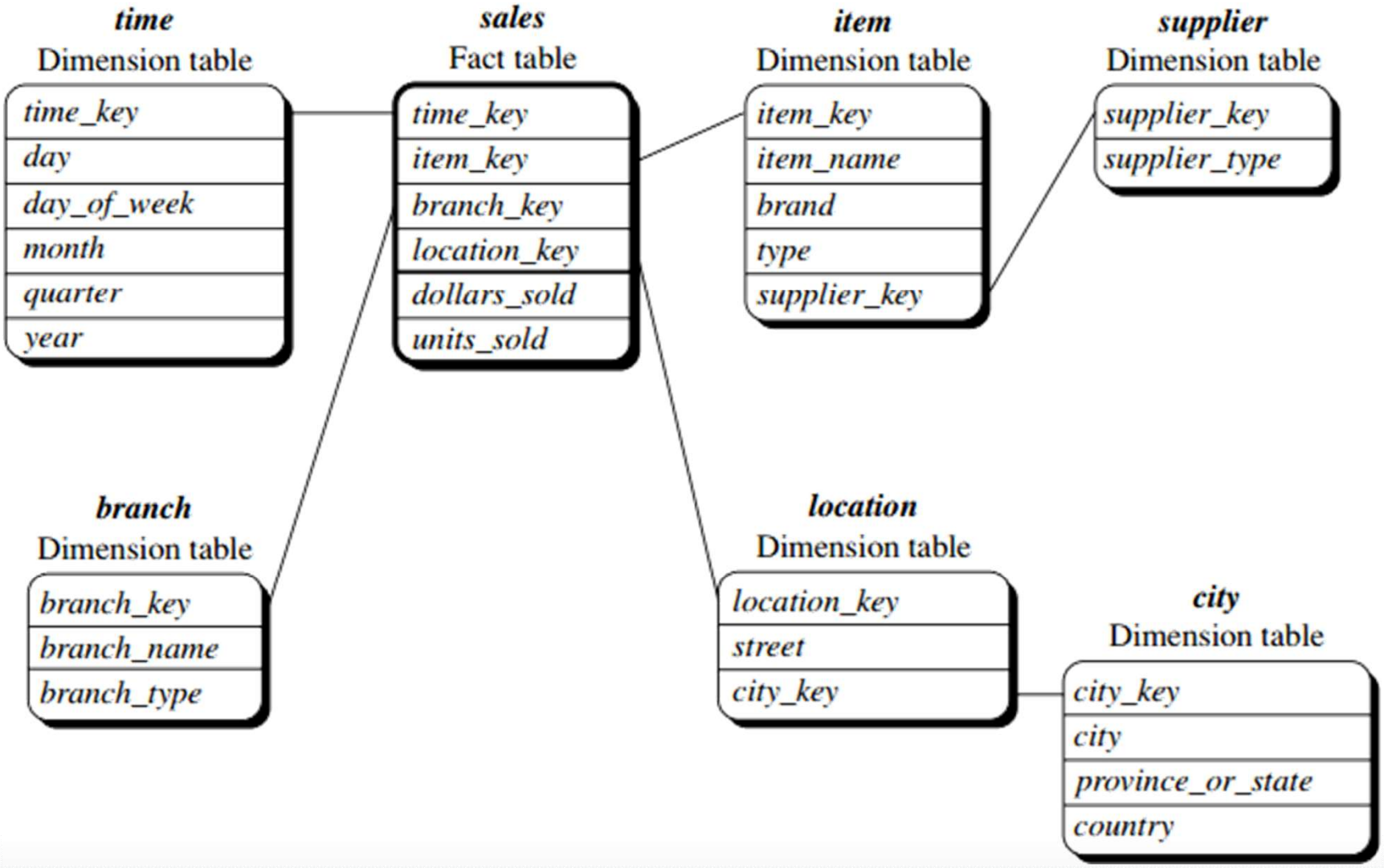
…Modelo de datos

- Esquema **Estrella**



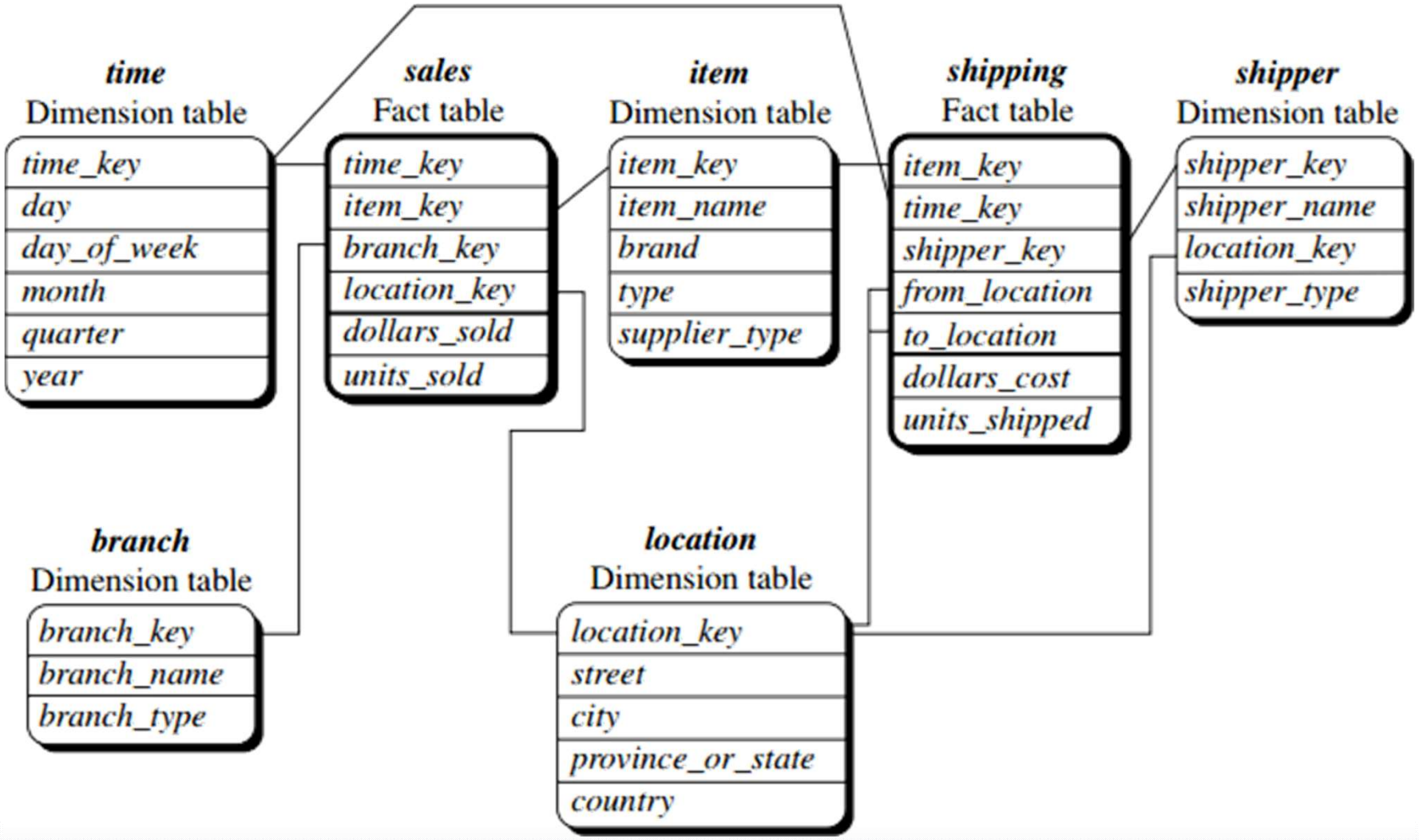
…Modelo de datos

- Esquema **Copo de Nieve**



…Modelo de datos

- Esquema **Constelación de hechos**



Procesamiento analítico

- Como se puede observar, los datos en el modelo multidimensional son percibidos y manejados como si estuvieran almacenados en una matriz de varias dimensiones.
- El procesamiento analítico 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

- 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 claúsula **GROUP BY**.

Cubos OLAP

- Se trata de la forma más popular para analizar información (*bases de datos multidimensionales*).
- Básicamente, un **cubo** es una estructura de datos organizada mediante **jerarquías**. Cada **medida** se puede evaluar en cualquiera de los niveles de las jerarquías: *por ejemplo, se podrían analizar las* ***ventas*** *diaria, mensual o anualmente, para un cliente, una región o un país*.
- Tienen la capacidad de **analizar** y **explorar** los datos: *Nos permiten cambiar el enfoque del* ***¿qué esta pasando?*** *(enfoque relacional) al* ***¿por qué esta pasando?*** *(enfoque multidimensional)***.**
- Las herramientas con capacidades OLAP nos proporcionan **análisis interactivo** por las diferentes dimensiones de los datos.

…Definición


Aguascalientes
Chihuahua
Distrito Federal
Clientes
Oaxaca
304
Veracruz
2009
2008
2007
2006
Tiempo
Ciencia Ficción
Deportes
Drama
Terror
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**, eso supone un gasto extra de recursos (*servidores, discos, procesos de carga, etc.*)
- El modelo de negocio no siempre se adapta bien en un modelo jerárquico, 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.*

Cubos de datos

- Como ya se vio, un cubo de datos se modela y visualiza a partir de múltiples dimensiones (**dimensiones y hechos**).
- Usualmente pensamos en los cubos, como una estructura geométrica de **3 dimensiones**, sin embargo, en los almacenes de datos, los cubos son ***n-dimensionales***.
- Por ejemplo, un cubo en **dos dimensiones** es una **tabla** (*o una hoja de cálculo*):

	location = "Vancouver"			
	item (type)			
time (quarter)	home entertainment	computer	phone	security
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q4	927	1038	38	580

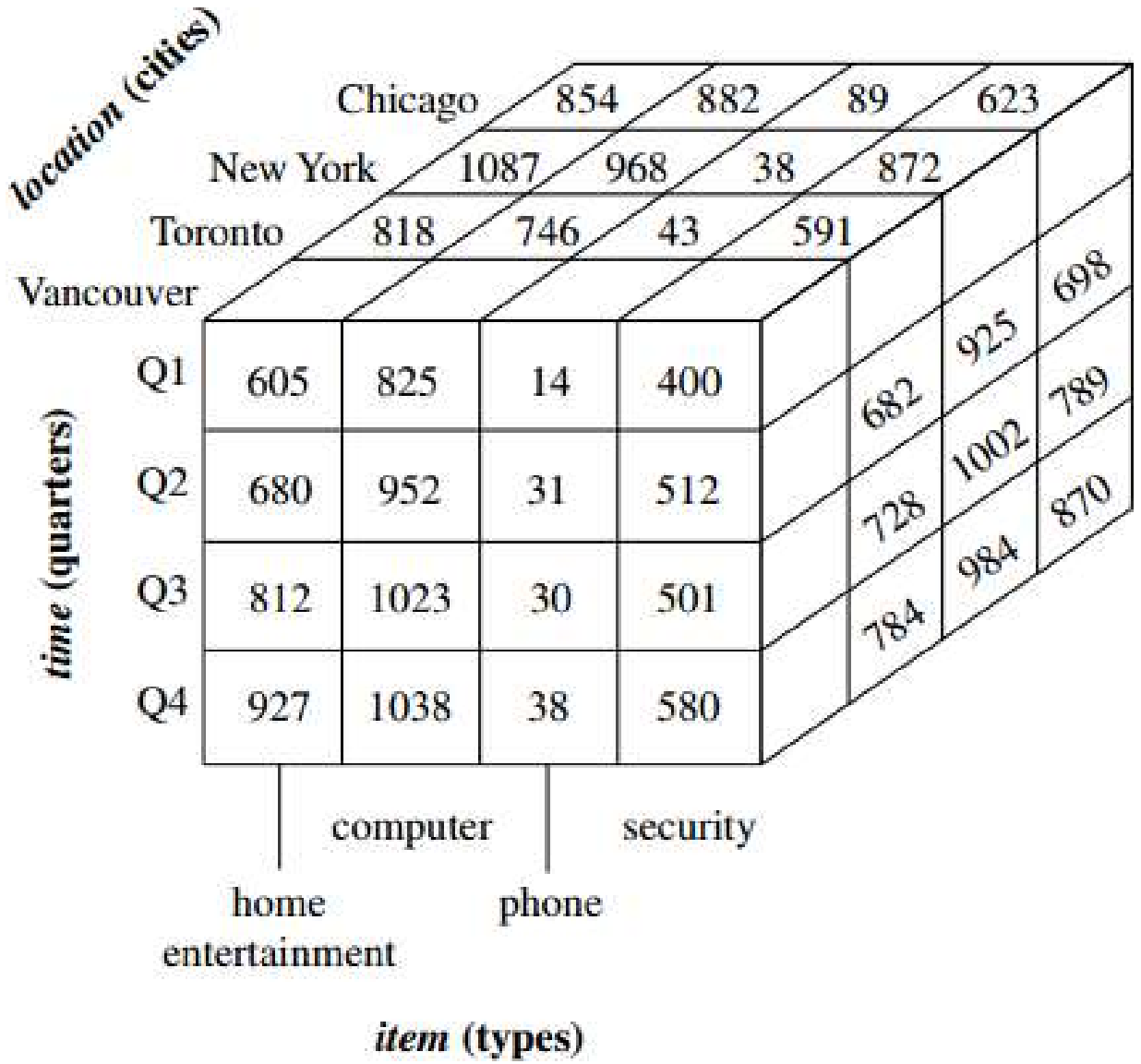
…Cubos de datos

- Vamos a suponer que queremos ver la perspectiva de ventas con una tercera dimensión:

	location = "Chicago"				**location** = "New York"				**location** = "Toronto"				**location** = "Vancouver"			
	item				**item**				**item**				**item**			
	home				*home*				*home*				*home*			
time	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*	*ent.*	*comp.*	*phone*	*sec.*
Q1	854	882	89	623	1087	968	38	872	818	746	43	591	605	825	14	400
Q2	943	890	64	698	1130	1024	41	925	894	769	52	682	680	952	31	512
Q3	1032	924	59	789	1034	1048	45	1002	940	795	58	728	812	1023	30	501
Q4	1129	992	63	870	1142	1091	54	984	978	864	59	784	927	1038	38	580

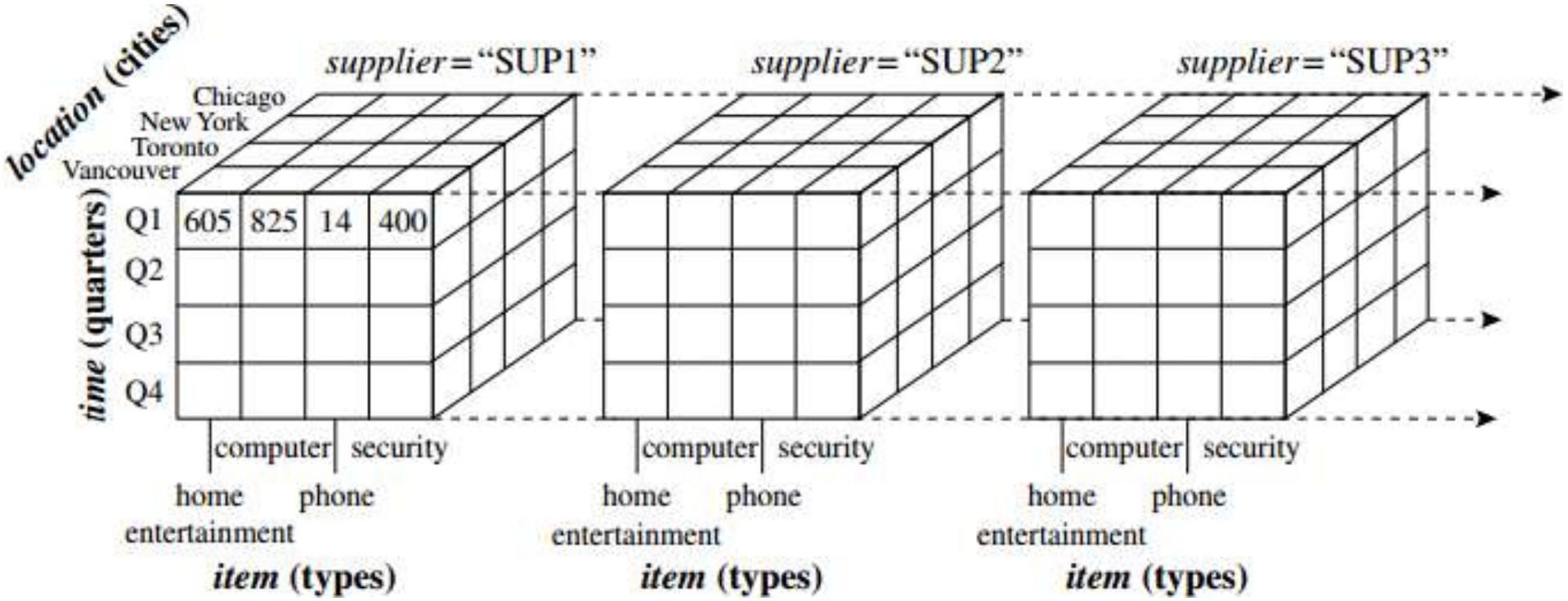
- Como se puede observar, los datos se presentan en series de tablas de 2 dimensiones, lo cual conceptualmente se suele representar a través de un cubo.

...Cubos de datos



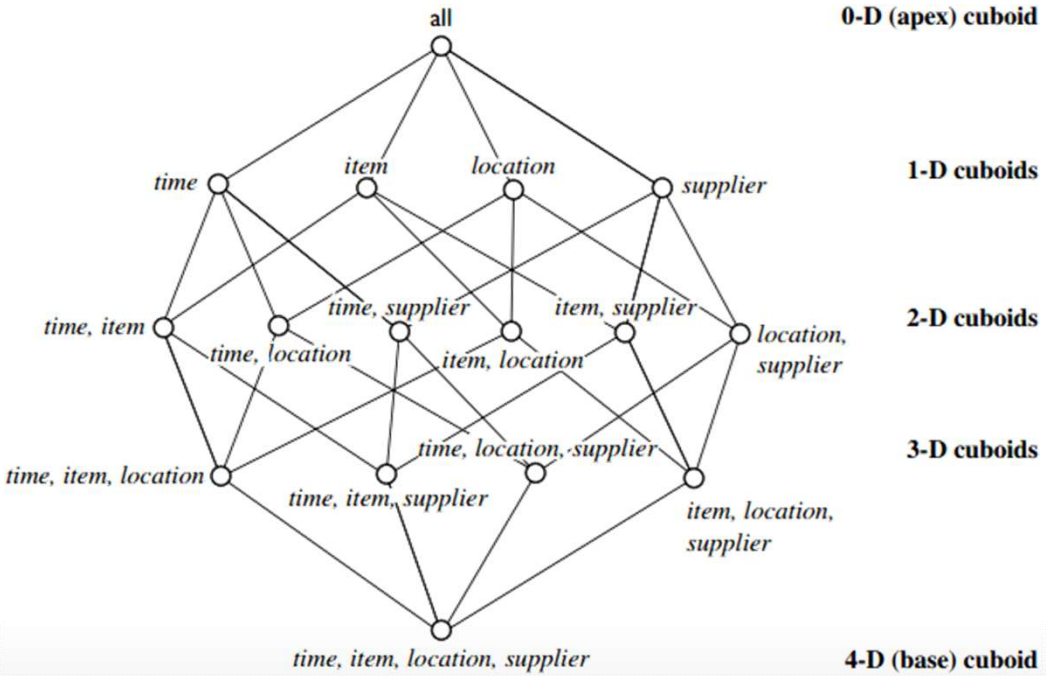
…Cubos de datos

- Si ahora queremos visualizar la perspectiva de ventas a través de la dimensión **distribuidor**, tendría que ser a través de una perspectiva de cuatro dimensiones, sin embargo, podemos visualizarlo a través de series de cubos de 3-D:



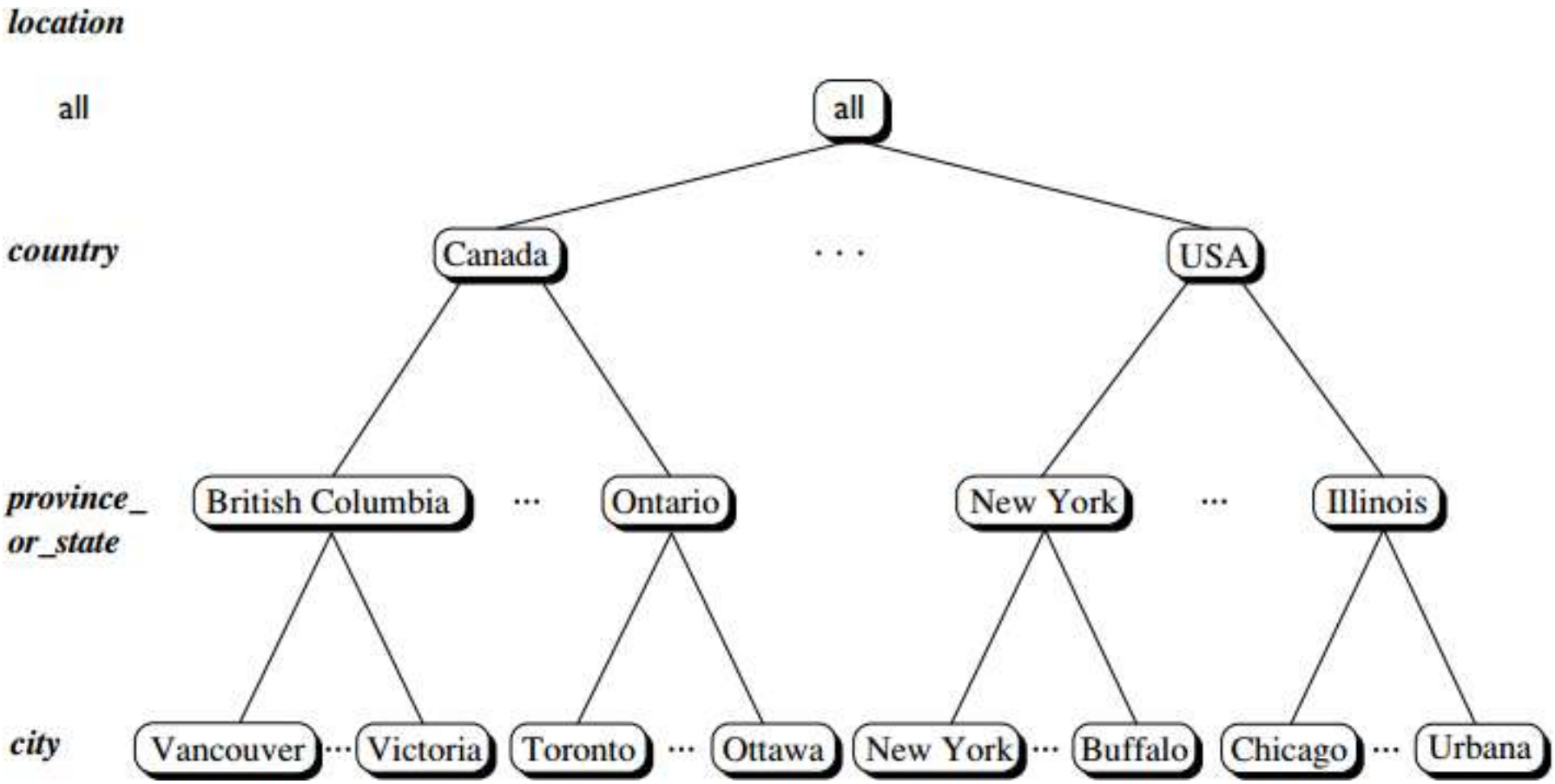
…Cubos de datos

- Los cubos de más allá de 3 dimensiones se denominan en la literatura como **cuboides**. El cubo que tiene el menor grado de agregación se denomina **cuboide base.** El cuboide con el más alto grado de agregación se denomina **cuboide ápex (*all*)**.
- Dado un conjunto de dimensiones, podemos generar un cuboide para cada posible subconjunto de dimensiones, lo cual se conoce como **enrejado de cuboides**:



Jerarquía de conceptos

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



THE END

Don't be too proud of having passed this subject. The ability to approve the Databases' course is insignificant next to the power of the Force…