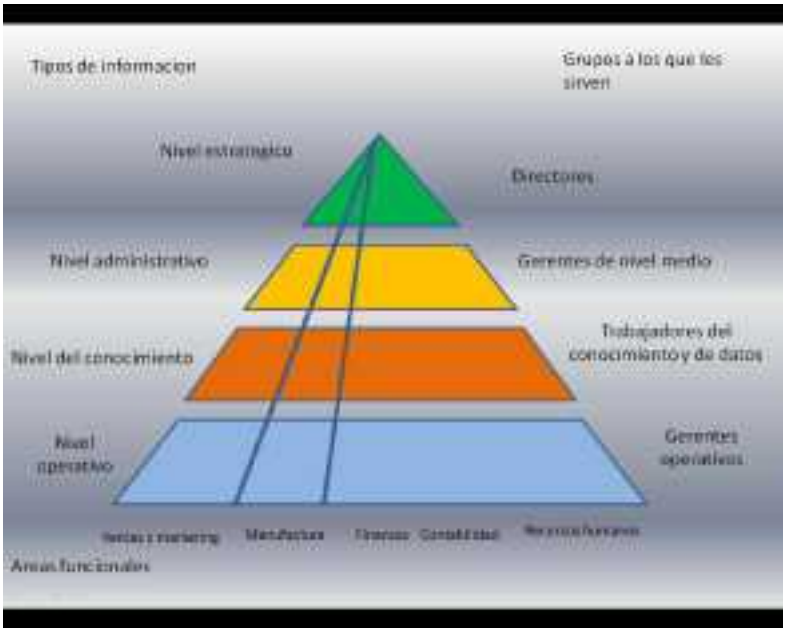
# Tema 9: Paradigmas BD – Big Data

- **Introducción. Paradigmas de Almacenamiento.**
  - características ACID
  - teorema CAP
- **NoSQL & NoRel - DB taxonomía**
  - par clave-valor (key-value)
  - basado en columnas
  - orientado a documento
  - en grafo
- **Big Data**
  - Concepto
  - Arquitectura

# Tema 9: Paradigmas - Evolución

- Con la evolución de la tecnología, se amplían los usos y necesidades, y funcionalidades que dan soporte a todos los niveles de la organización




- **Ficheros de Datos:** Distintas aplicaciones, específicas para cada tarea (nivel operativo)
- **Bases de Datos** ('60)**:** Menor redundancia, mayor integridad, independencia dato-uso, ...

OLTP

- **Data Warehouse:** Sintetizar información que pueda soportar la toma de decisiones

OLAP

# Tema 9: Paradigmas - OLTP

- El paradigma ***On-Line Transaction Processing*** es un tipo de arquitectura de sistemas de información centrado en dar servicio al procesamiento de operaciones (de BD).
- Diferencia entre procesos interactivos (on-line) y *por lotes* (batch)
- La transacción es una operación que cumple ciertas propiedades:

  **A** **Atomicidad**: toda operación, o se completa o no se realiza en absoluto.

  **C** **Consistencia**: debe preservarse las reglas de integridad de la BD.

  **I** **Aislamiento**: las operaciones no se afectan entre sí (son independientes).

  las concurrentes siempre deben dar lo mismo.

  **D** **Persistencia**: una vez realizada, su efecto es permanente.

- Soportan la actividad de la organización: *robustez* y *disponibilidad*.
- Generalmente, se construyen sobre arquitectura cliente-servidor.
- Modelos de datos estructurados: *relacional*, jerárquico, red, OO, ...

# Tema 9: Paradigmas - OLAP

Diogenes, acumula datos.

- El paradigma ***On-Line Analytic Processing*** parte de los sistemas OLTP para crear grandes almacenes (*data-warehouse*), generalmente estructurados, cuyo objeto es sintetizar información y generar informes.

- Mientras que *OLTP* busca eficiencia en las transacciones (limitado contenido histórico, estructuras complejas), *OLAP* busca eficiencia en consultas analíticas (gran **contenido histórico**, estructuras simples).

- Presenta diversos tipos:
  - **ROLAP**: construidas sobre motores relacionales (tab. desnormalizadas)
  - **MOLAP**: bases de datos multidimensionales (varios niveles abstracción)
  - HOLAP: BD híbridas (parte relacional, parte multidimensional)
  - Otras: WOLAP (web), SOLAP (spatial), DOLAP (escritorio), **RTOLAP** (real-time)

Distribución Multiclave

- Objetivos habituales: **agregación**, **comparación**, **correlación**, **clasificación** (rank), **previsión**, **simulación** (what-if), etc.

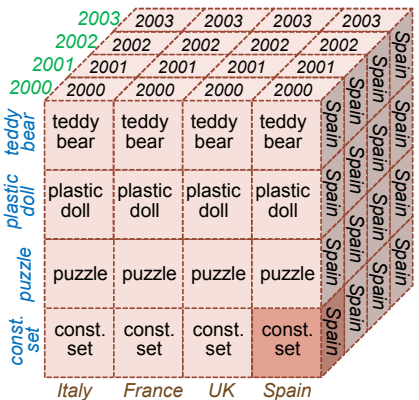
- Estos almacenes suelen basarse en datos precalculados sobre algunas dimensiones prefijadas (cubos OLAP) → requieren análisis y diseño.

# Tema 9: *Back-End* para Analítica

- **Consultas Analíticas:** *¿Buscas algo concreto o solo quieres echar un vistazo?*

  Toma un libro. Con el índice puedes encontrar fácilmente cierto capítulo, pero…
  ¿podrías saber el número medio de figuras de cada capítulo?
  ¿puedes contar cuántos verbos hay en los capítulos impares?

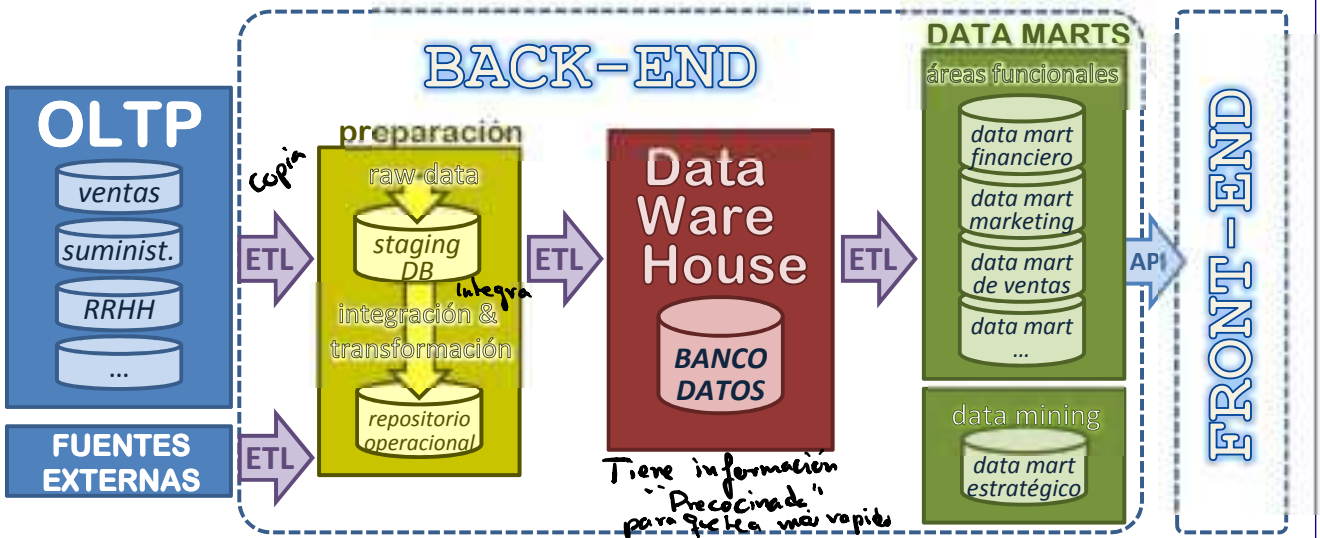
- **Estructuras Específicas:**
  - rejillas n-dimensionales con datos extractados, resumidos o agregados
  - el análisis de atributos aislados puede hacerse procesando solo índices.
  - en otro caso, tendrás que recorrer (*fullscan*) el fichero completo…



# Tema 9: Data Warehouse

*Enormes almacenes de datos de alto contenido histórico e información resumida usados para "generar informes"* (asientan la *inteligencia de negocio*).

- **Almacenamiento**: de *micro-datos* eventuales a *macro-datos consolidados*
- **Proceso** básico: ***ETL*** ( extracción → transformación → carga ).
- Metodología **Diseño**: top-down (*¿qué tengo?*) vs. bottom-up (*¿qué quiero?*)



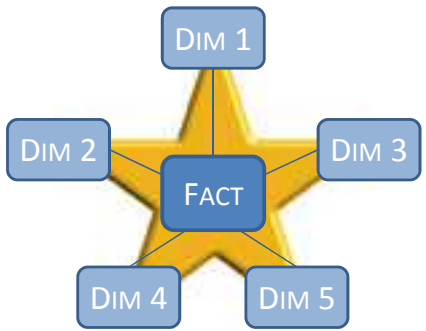
# Tema 9: Diseño en ROLAP

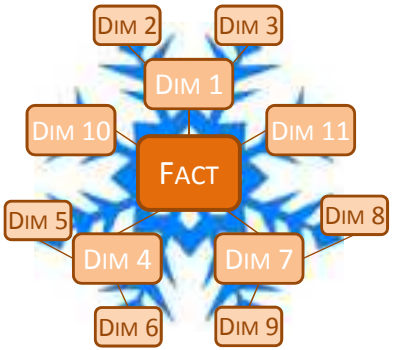
- **ROLAP**: persigue escalabilidad, ↓coste, mantenimiento (a costa de la eficiencia)
  - esquema en **estrella**: una tabla central (hechos) relacionada con varias tablas satélite (normalizadas) referidas a las dimensiones de los hechos (por ejemplo, para 'ventas' las dimensiones son cliente, producto, tienda, ...)

    → ventajas: simplicidad (diseño), baja redundancia, mantenimiento sencillo.
  - esquema en **copo de nieve**: una o más tablas centrales (hechos) articulando varias tablas satélite referentes a cada dimensión, y éstas a su vez se vinculan a varias sub-dimensiones... (por ejemplo, *tienda* tiene *empleados* y vehículos, y éstos a su vez tienen *perfil_vehículo* y *hangar*, ...

    → ventajas: mayor fragmentación (tablas menos grandes), agilidad y simplicidad

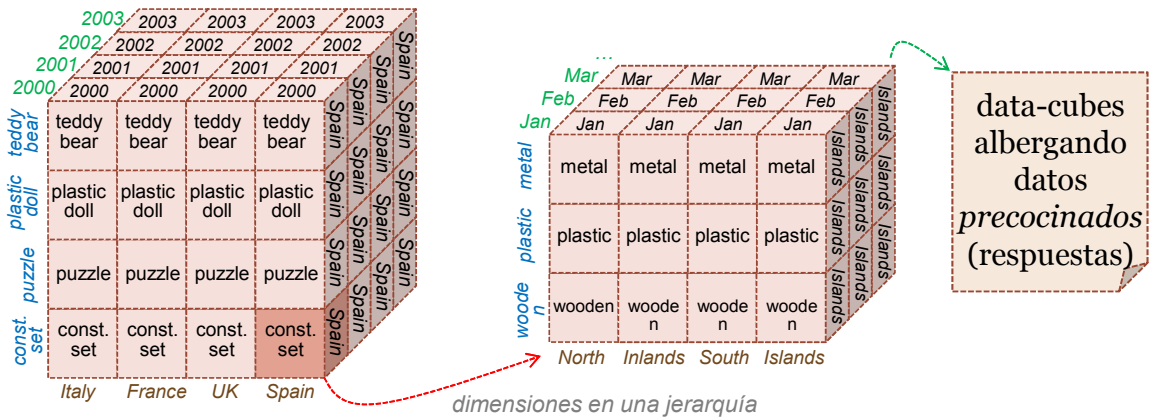
Mucho más compleja



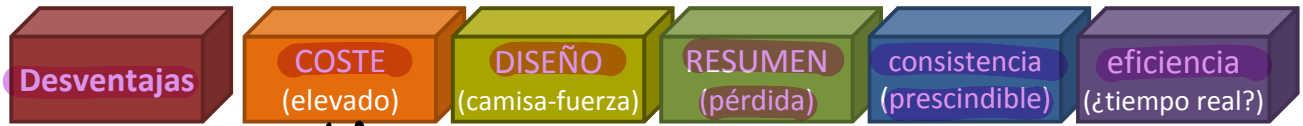
# Tema 9: características MOLAP (vs. ROLAP)

- **MOLAP**:
  - compacto
    - eficiente: rápido, pero en riesgo de *data burst*
    - versátil: supports complex analysis
  - constreñido
    - rígido: sujeto a un diseño (dimensiones, ETL, …)
    - resúmenes: pérdida de información y de precisión
    - latencia de datos: procesos ETL pesados



# Tema 9: Evolución OLAP

- Aunque hubo algunos precursores desde mediados de los 70', el impulso real vino en los 90'. Los Data Warehouses se desarrollaron para soportar la expansión de las compañías hacia posiciones dominantes del mercado. Las crecientes exigencias de los clientes desbordan la tecnología…




Muchos en paralelo.

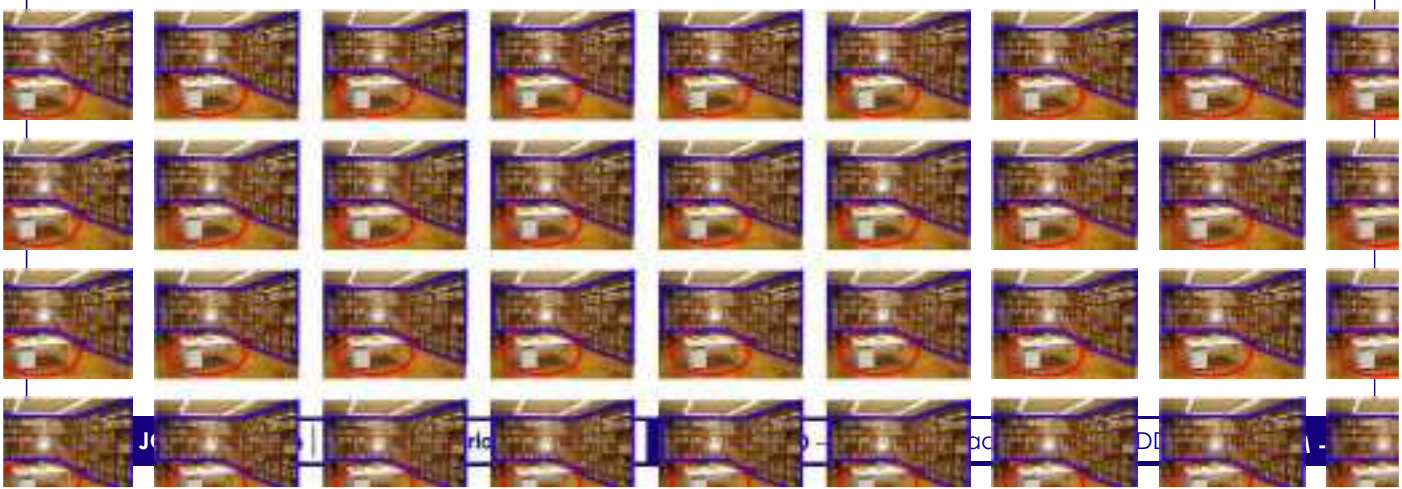
- *Al evolucionar los recursos Hw y las necesidades, es necesario adaptarse…*

# Tema 9: *La unión hace la fuerza…*

**Tarea pesada:** Contar los verbos del *'Quijote'* implica recorrer serialmente 1128 páginas

**A menos que…**
llames a unos amigos, y repartas el trabajo

**Sólo 141 bibliotecas federadas…**
podrían procesar 8 páginas (cada una)



# Tema 9: Paradigmas - RTAP

- El paradigma ***Real-Time Analytic Processing*** parte del enfoque OLAP buscando respuestas estratégicas, pero no en informes sino en tiempo real (permitiendo dirigir decisiones a corto plazo).
- Esta eficiencia sobre grandes volúmenes de datos (crecientes) se busca mediante la escalabilidad horizontal y el ***paralelismo masivo***.
- Datos y procesos deben ser distribuidos, con los problemas que conlleve.
- Los sistemas se ajustan al **teorema CAP**, que establece que se ajustan a dos características prescindiendo necesariamente de la tercera:

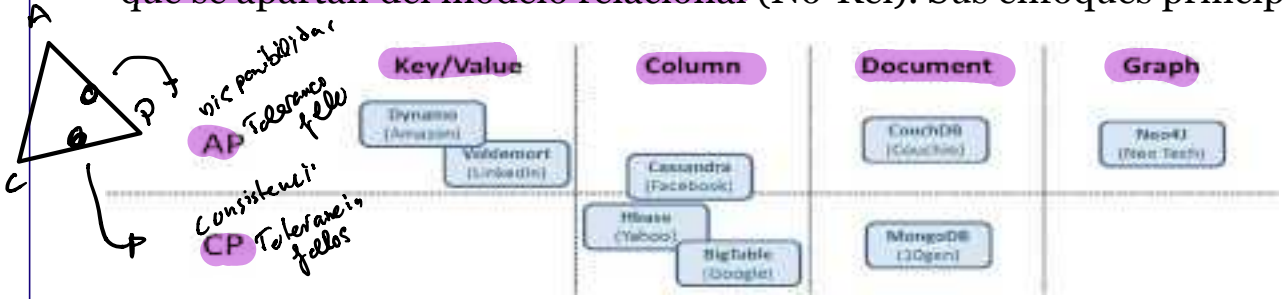
**C** **Consistencia**: todos los nodos ven la misma información (eq. con *atomicidad*).

**A** **Disponibilidad**: toda solicitud es atendida (y respondida + ó -). Siempre disponible

**P** **Tolerancia a fallos**: el sistema responde correctamente aún con fallos de red.

# Tema 9: NoSQL y NoRel

- Los sistemas relacionales distribuidos no son robustos ante fallos de red (pérdida de mensajes) o si se caen nodos (son CA en el teorema).
- Para posibilitar tal escalabilidad y distribución sin sacrificar la tolerancia a fallos, se precisa un almacenamiento flexible, prescindiendo (o cuando menos relajando) de la estructuración de los datos.
- El concepto ***NoSQL*** (not only SQL) inicialmente hacía referencia a extensiones sobre la gestión de datos relacional (orientación a objetos, anidamiento, gestión XML, federación, fragmentación, rejilla, etc.).
- Sin embargo, No-SQL ha acabado referenciando a tecnologías y SGBD que se apartan del modelo relacional (No-Rel). Sus enfoques principales:



# Tema 9: Key-Value



- Cada elemento tiene una clave (que lo identifica) y un valor en un dominio. El dominio puede abarcar varios atributos, pero no está esquematizado (no presenta un diseño fijo).
- Cada ocurrencia (fila o registro) es un elemento con un valor de clave y un valor de registro (formada por uno o más atributos y sus valores).
- Ventajas:
  - almacenamiento de gran flexibilidad e independencia.
  - alta escalabilidad
- Desventajas: eficiencia según qué proceso

| RowID | 14637 |
|---|---|
| Name | Jack |
| Order | 85468 |
| Total | 300€ |

| RowID | 14637 |
|---|---|
| Name | John |
| Order | 74378 |
| Total | 199€ |

| RowID | 14637 |
|---|---|
| Name | Jane |
| Order | 73568 |
| Total | 95.9€ |

clave

valor

dominio: *transacciones*

# Tema 9: Column Based


- En este tipo de almacenes, cada columna se almacena por separado.
- Ventajas:
  - si los datos se agrupan (cluster), se puede ahorrar espacio (en algunas columnas).
  - permite calcular agregaciones de datos con mayor eficiencia
  - No se recuperan las columnas que no se necesitan.
- Desventajas:
  - recomposición de grandes filas

| RowID | Name | Surname |
|---|---|---|
| 123 | John | Doe |
| 656 | Jack | Smith |
| 724 | Jane | García |

| RowID | Name |
|---|---|
| 123 | John |
| 656 | Jack |
| 724 | Jane |

| RowID | Surname |
|---|---|
| 123 | Doe |
| 656 | Smith |
| 724 | García |

# Tema 9: Document Oriented


- **Orientada a almacenar documentos (JSON, BSON, XML, …). Ejemplo JSON:**

```
{
   "Nombre":"Alexandre",
   "Apellido":"Dumas",
   "Títulos": [ { "Nombre":"Los tres Mosqueteros",
                  "Año": 1844  },
                { "Nombre":"El Conde de Montecristo",
                  "Año":"1844" }
              ]
},
{
    "Author":"Alexandre Dumas, hijo",
    "Titles": [ { "Nombre":"La Dama de las Camelias",
                  "Año": 1848  },
                { "Nombre":"El caso Clemenceau",
                  "Año": 1867
        }
    ]
}
```

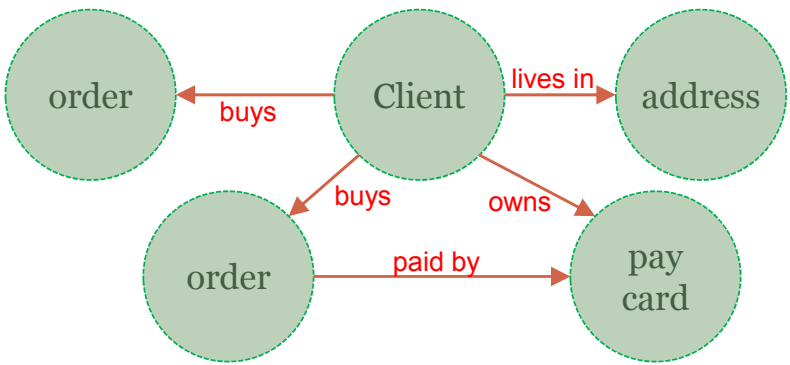
Ventajas: máxima flexibilidad; intuitivo y de fácil manejo; escalable

Desventajas: mayores volúmenes; perder estructura complica la consulta

# Tema 9: Graph shaped



- Tan flexible como los documentos, pero con relaciones entre nodos (docs).
- Las relaciones pueden ser muy numerosas y tienen semántica.
- Puede dotarse de estructura, y prescindir de esta en las consultas.



# Tema 9: Concepto Big Data

- Se refiere al conjunto de técnicas y tecnologías necesarias para procesar volúmenes de datos tan elevados que las herramientas habituales no pueden dar el soporte adecuado.

- Tal volumen de datos puede tener orígenes diversos: transacciones, web, sensores, datos generados por personas, biometría, instrumentos científicos, dispositivos, …



- *Big Data* es un área multidisciplinar. Los conocimientos y herramientas que precisa están relacionados con las fases del *ciclo de vida de la información*.

# Tema 9: Las 'V' de Big Data

# Tema 9: La Nube (y la *tormenta*)

- El concepto de 'nube' se refiere a un conjunto de recursos Hw-Sw **compartidos** por una multiplicidad de usuarios. Los recursos están bien comunicados y son **accesibles** desde prácticamente cualquier sitio (**ubicuos**).
- Los recursos pueden alquilarse a un *proveedor de servicios*, pudiendo alquilar una capa (infrastructura), dos (plataforma) o el lote completo (con software).
- Específicamente, el término *nube de datos* habitualmente se refiere al alquiler de servicios de base de datos (sin datos, cuya transferencia puede ser ilegal).

Según los retos del área, en unos años estaremos nadando en ríos de información privada y océanos de información pública.

Sin embargo, las herramientas (y tecnología) deben evolucionar para poder navegar por tal caos de datos con agilidad, eficacia y eficiencia.