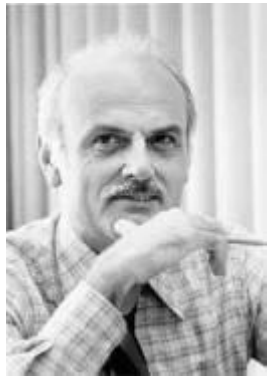
Introducción a DW

Diferencias OLAP vs OLTP

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

Ciclo de Vida de BI

Hechos disruptivos en data management



1970 – E. F. Codd – Modelo relacional de datos

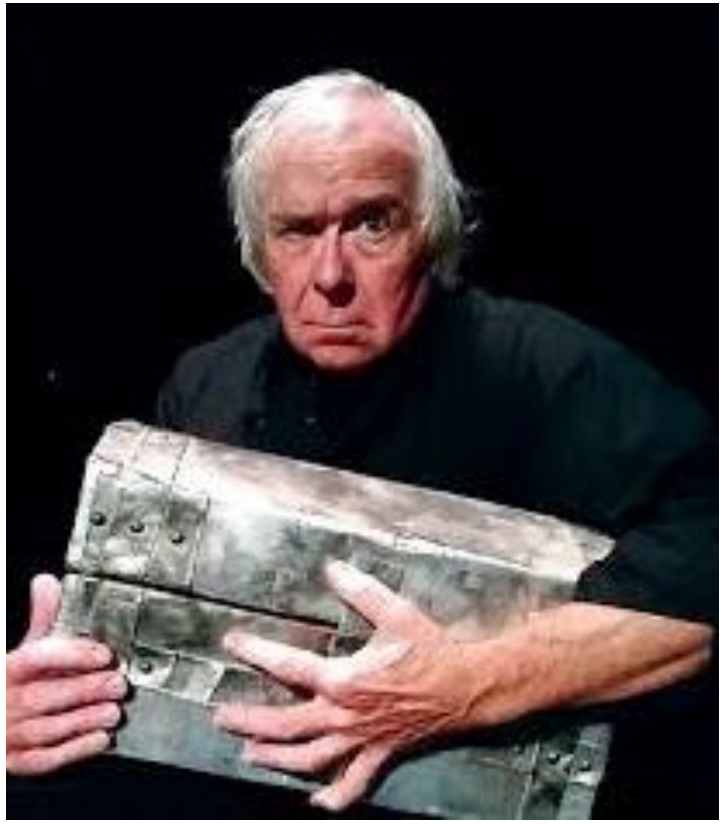
1991 – Tim Berners-Lee – La WWW

Escenario típico en los 90's

- Data Warehousing + Data Mining
- Tamaños grandes: GB…TB!!
- Datos estructurados
  - Mayoritariamente relacionales
  - Spreadsheets
  - Algo de texto
  - La web aún era incipiente
- Problema: integración de datos

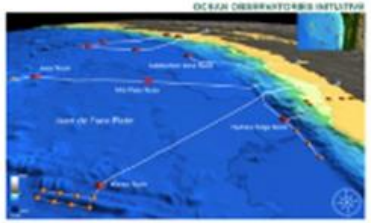
Un cambio de paradigma

Antes: los datos son míos, míos, míos…

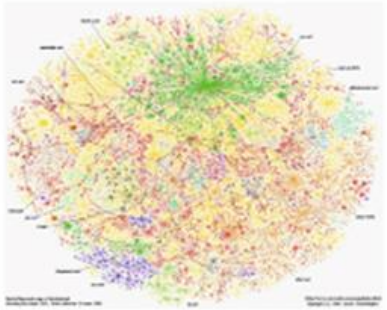


Astronomy: LSST

Physics: LHC



Oceanography



Sociology: The Web

Biology: Sequencing

Data-Driven Medicine

Economics: mobile, POS terminals

Neuroscience: EEG, fMRI

Sports

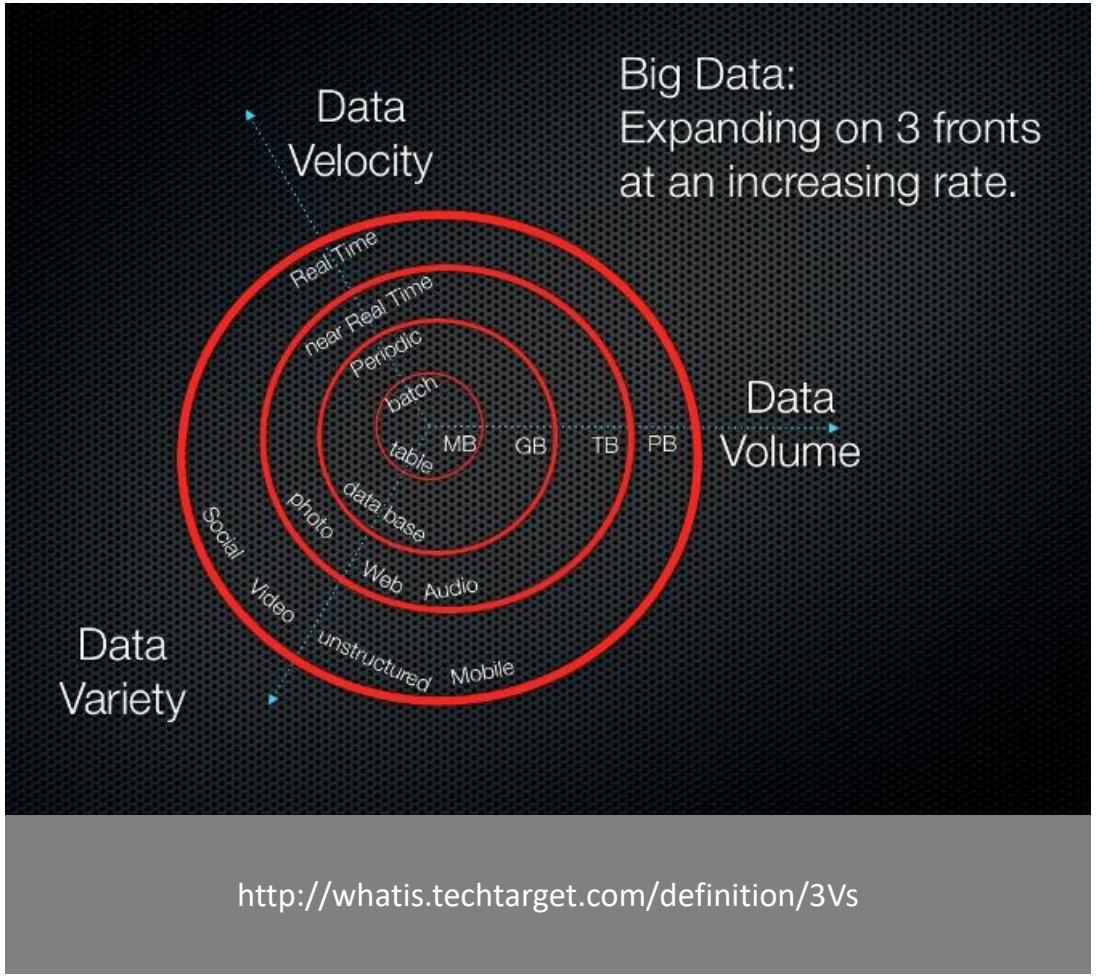
Ahora

¿Qué hacemos con todos estos datos?

Definición

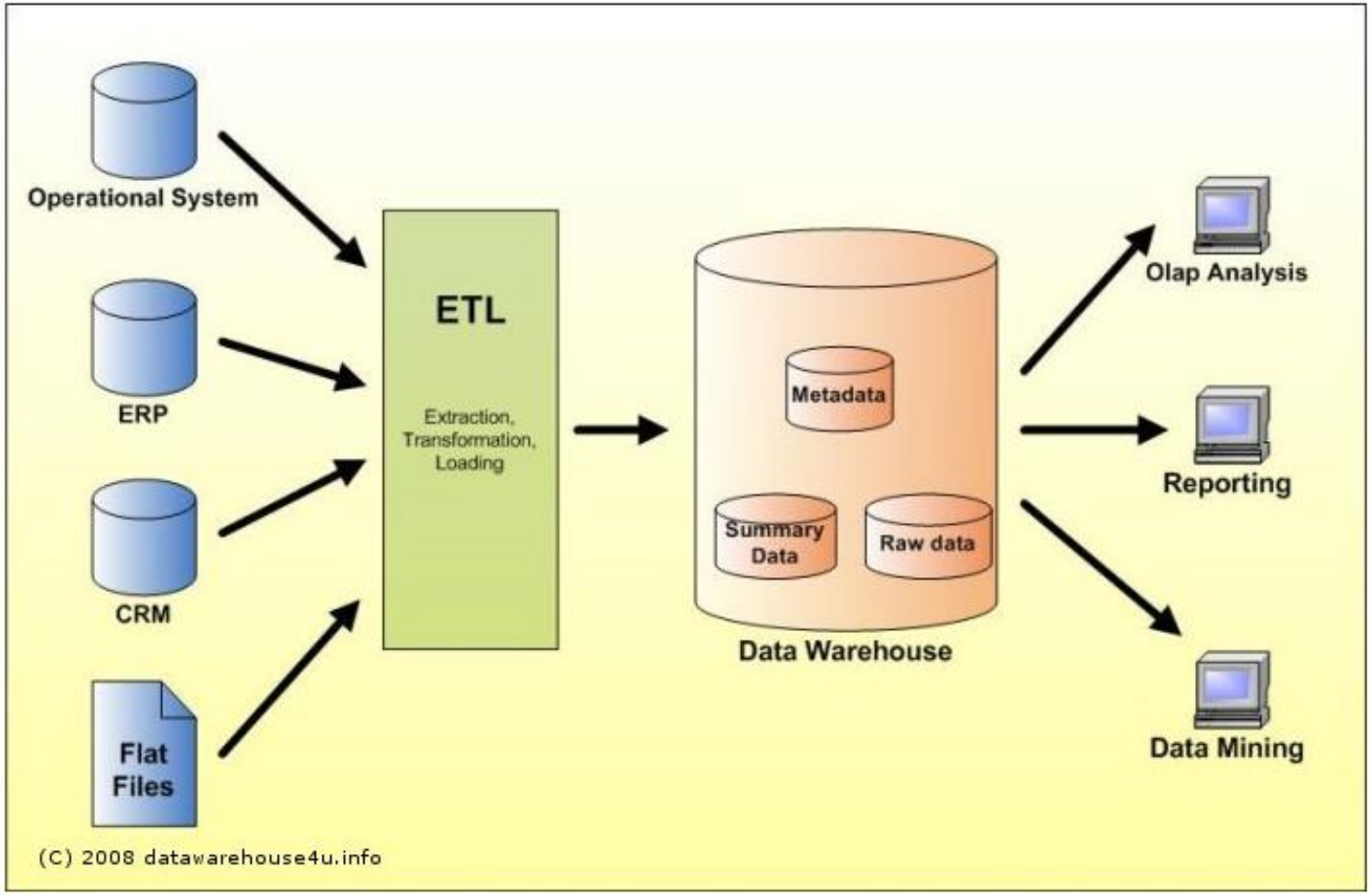
“Big Data” refiere a grandes volúmenes de datos, estructurados o no, que crecen a un ritmo tan grande que hace que su administración y explotación con las herramientas de bases de datos tradicionales no sea posible.

Las “3 Vs” (Doug Laney, 2001)

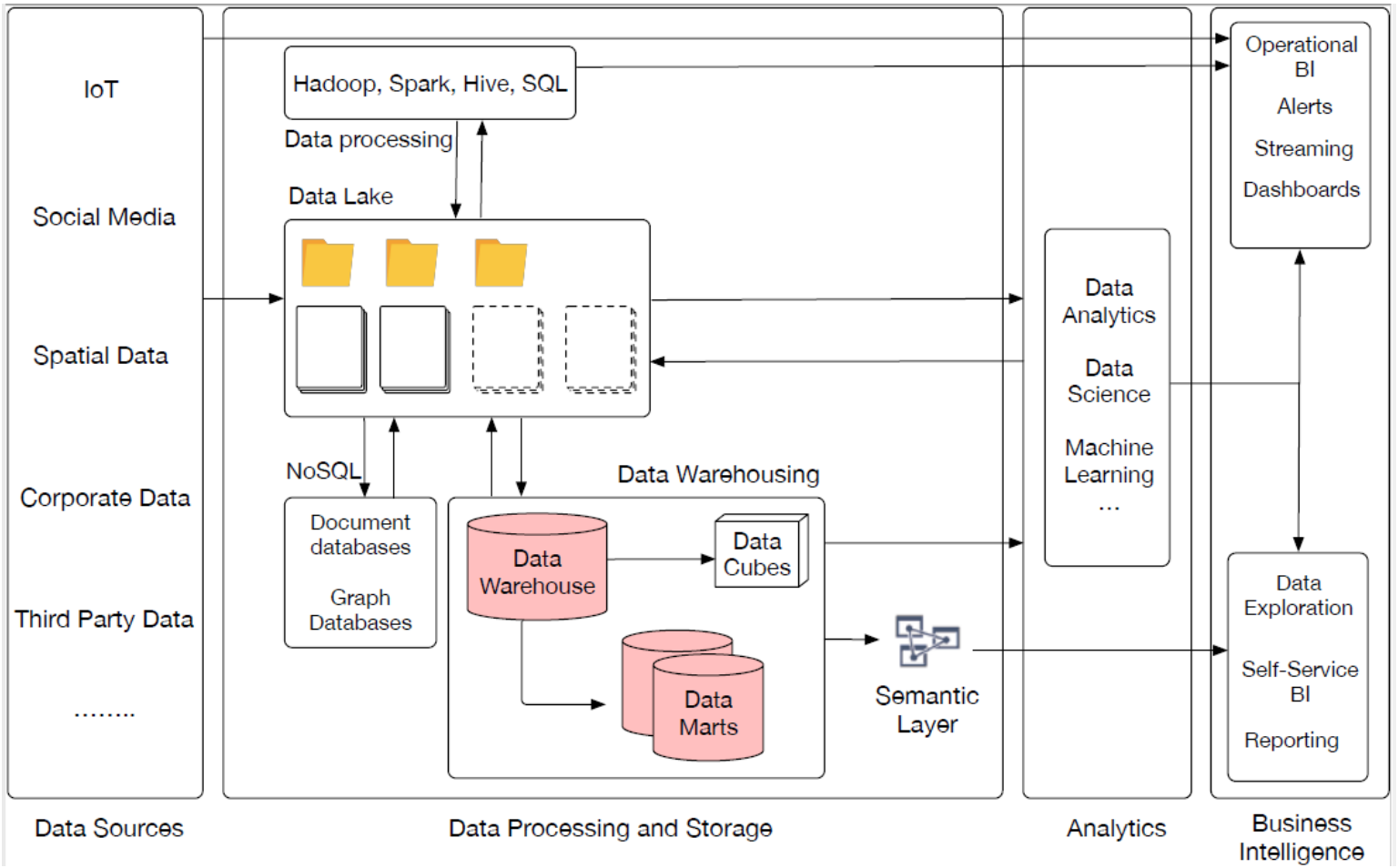


http://whatis.techtarget.com/definition/3Vs

Esquema Básico de DW



Arquitectura Moderna de BI



Hipótesis

El problema a resolver es el mismo:

Capturar, Integrar y Analizar datos!!

Lo que cambia con "Big Data" es el volumen y las características de esos datos, y las tecnologías para manipularlos

Dos Tecnologías Pilares de Big Data

- **Almacenamiento y procesamiento distribuido**
- **Tecnologías de bases de datos basadas en particiones y procesamiento paralelo**
  - NoSQL
  - New SQL/HTAP
  - Tecnologías de BD Distribuídas basadas en el Modelo Relacional

Técnicas

- Integración y Modelado de datos
- Online Analytical Processing (OLAP) & Data Warehousing
- Data Mining & Machine Learning
  - Clasificación
  - Arboles de decisión
  - Clustering
  - Análisis de sentimiento
  - Redes neuronales / Deep Learning
- Visualización de la información

+ EL CONOCIMIENTO DE LOS EXPERTOS EN CADA DOMINIO DE APLICACION

La Era de IA (A. Ng)

- En la Era de Internet

  - Compañía tradicional + Web Site NO es una compañía de Internet
  - Compañía de Internet:
    - A->B testing
    - Ciclos cortos
    - Niveles de decision más horizontales y más bajos, no solo el CEO

- En la Era de la IA

  - Compañía Tecnológica + IA NO es una compañía de IA
  - Compañía de IA:
    - Adquisición estratégica de datos
    - Data warehouse unificado
    - Nuevas job descriptions
    - Automatización pervasiva: empoderar a toda la compañía con estas técnicas

Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.

Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, Dan Dennison
{ebner,vchaudhary,mwyoung,jfcrespo,dennison}@google.com
Google, Inc.

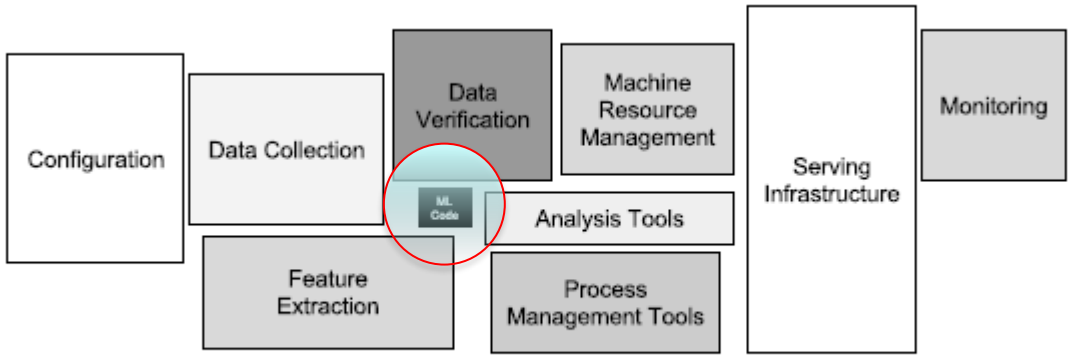


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Every single company I've worked with and talked to has the same problem without a single exception so far—poor data quality, especially tracking data

Ruslan Belkin vice president of engineering for Salesforce

Andrew Ng

If you can't build a data warehouse, you shouldn't do AI

Andrew Ng

Definición clásica (Immon)

Conjunto de datos históricos, integrados, no volátiles, y orientados a la resolución de un problema, para dar soporte a la toma de decisions.

- Orientados a la resolución de un problema (subject-oriented): el sujeto de análisis (ej: ausentismo laboral)
- Integrados: el contenido del DW resulta de la iontegración de datos de diversas fuentes
- No-volátiles: un DW acumula datos durante un período, normalmente no se elimina información
- Históricos: un DW mantiene la historia de la evolución de los datos a lo largo del tiempo

4 Etapas de diseño

Análisis de Requerimientos

Diseño conceptual

Diseño Lógico

Diseño Físico

4 Etapas de diseño: diferencias con OLTP

Análisis de Requerimientos:

- OLTP: basado en procesos, acceso mediante aplicaciones, usuarios de línea.
- OLAP: basado en queries, usuarios de tomas de decisión a distintos niveles.

4 Etapas de diseño

Diseño conceptual:

- OLTP: DER, el interés es modelar entidades y relaciones entre ellas. Deriva en relaciones normalizadas.
- OLAP: Modelo multidimensional, el interés es modelar hechos para ser analizados a través de dimensiones o perspectivas. Deriva en relaciones desnormalizadas.

4 Etapas de diseño

Diseño Lógico:

- OLTP: Típicamente modelo relacional.
- OLAP:
  - ROLAP: relational OLAP
  - MOLAP: Multidimensional OLAP
  - HOLAP: Hybrid OLAP

Diseño Físico

Diseño Físico:

- OLTP: Optimización mediante índices; transacciones online simultáneas, posibilidad de caídas….
- OLAP: Típicamente sólo lectura. Actualizaciones offline (refresh)

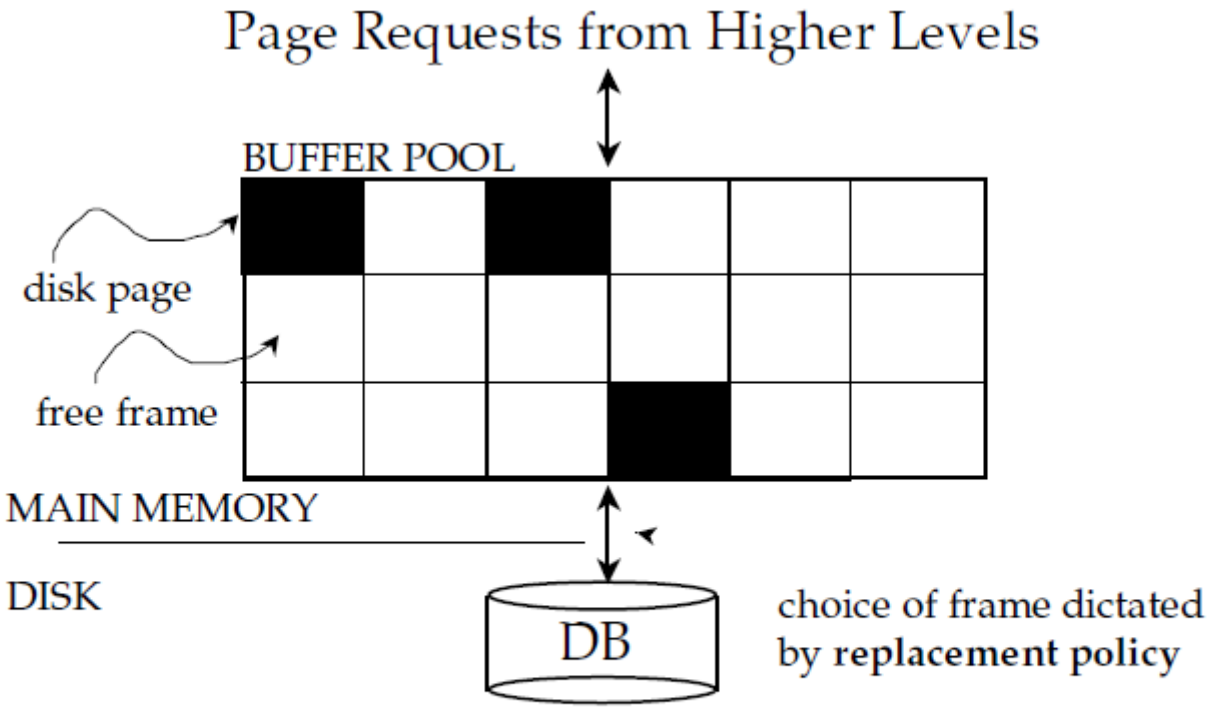
  Foco: Consultas.

Veamos primero algunas características de los DBMS

Almacenamiento en BDs

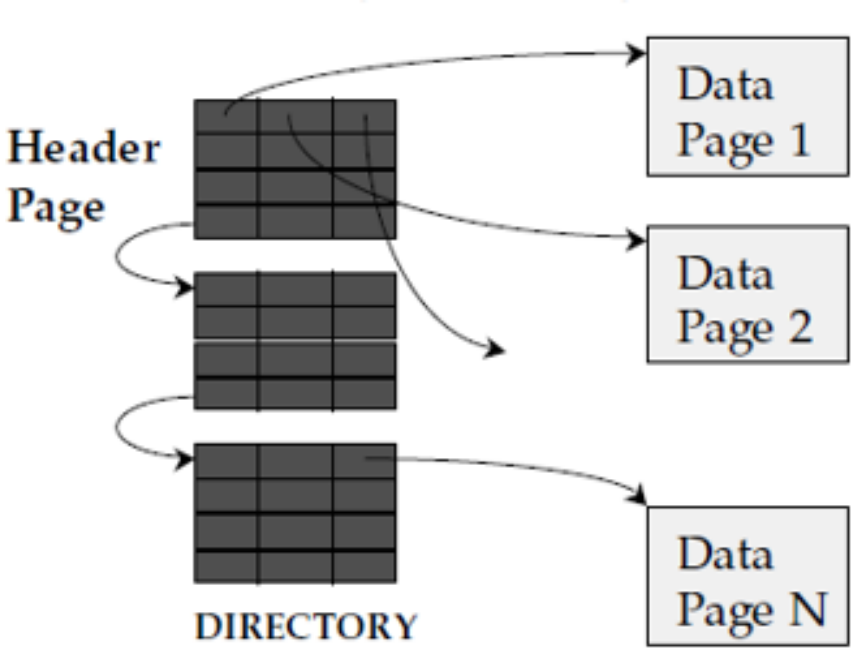
- Los niveles más bajos del DBMS manejan el espacio en disco
- Niveles superiores llaman a estas funciones para
  - Cargar y reemplazar una página
  - Leer/escribir una página
  - No conocen cómo esto se lleva a cabo

Manejo de Buffers

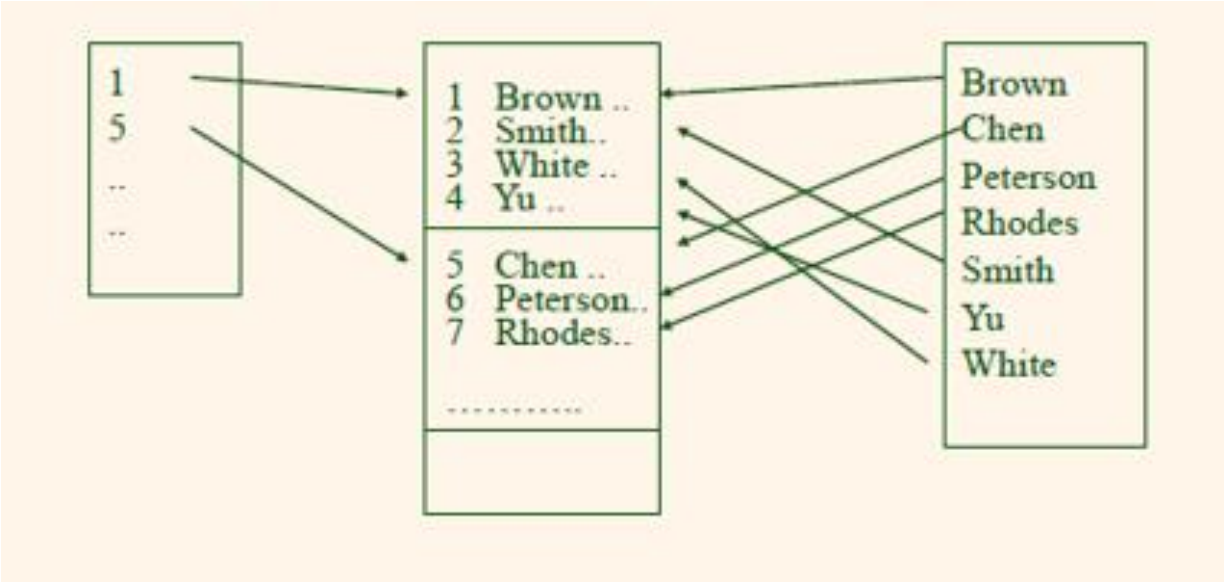


- *Data must be in RAM for DBMS to operate on it!*
- *Table of <frame#, pageid> pairs is maintained.*

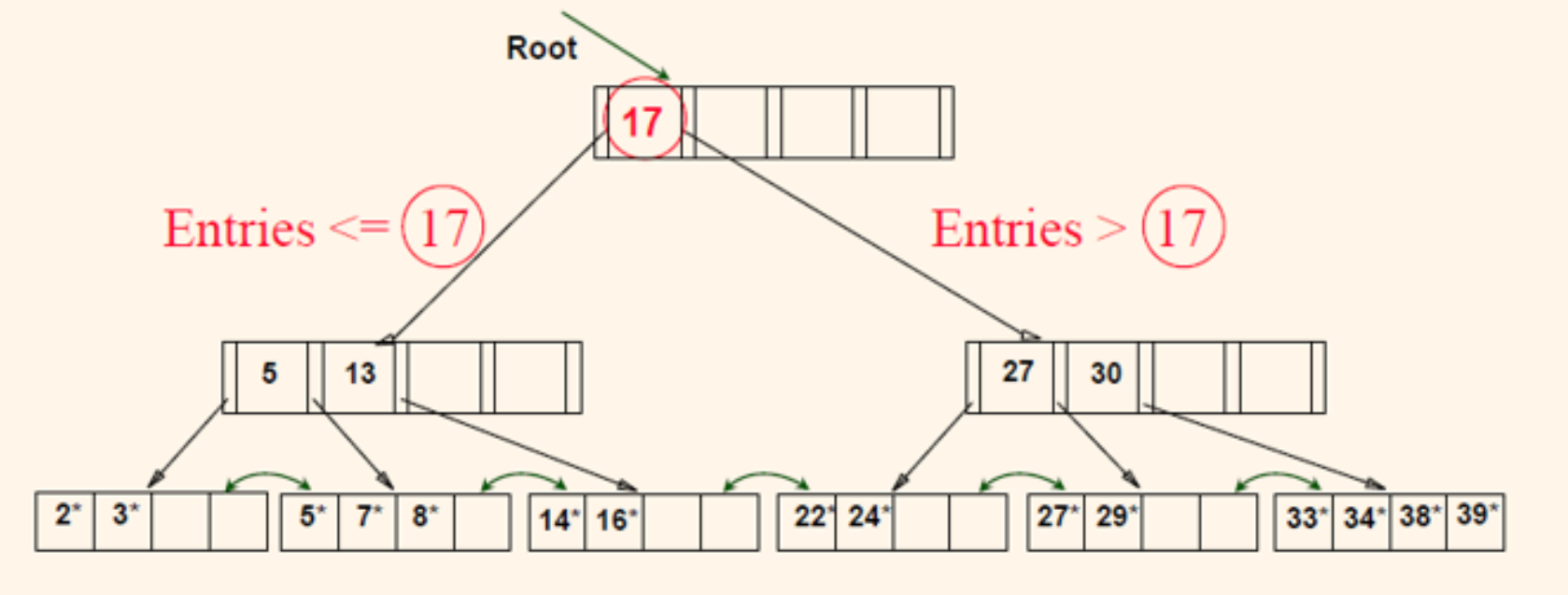
Organización básica: Heap Files



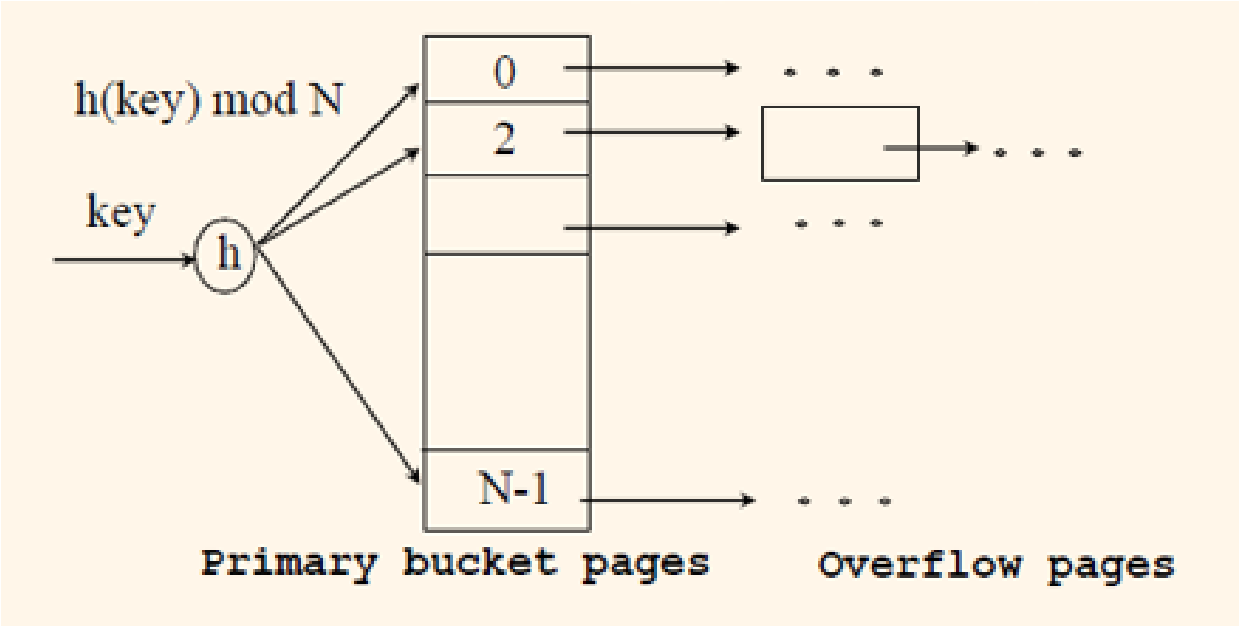
Organización Indexada



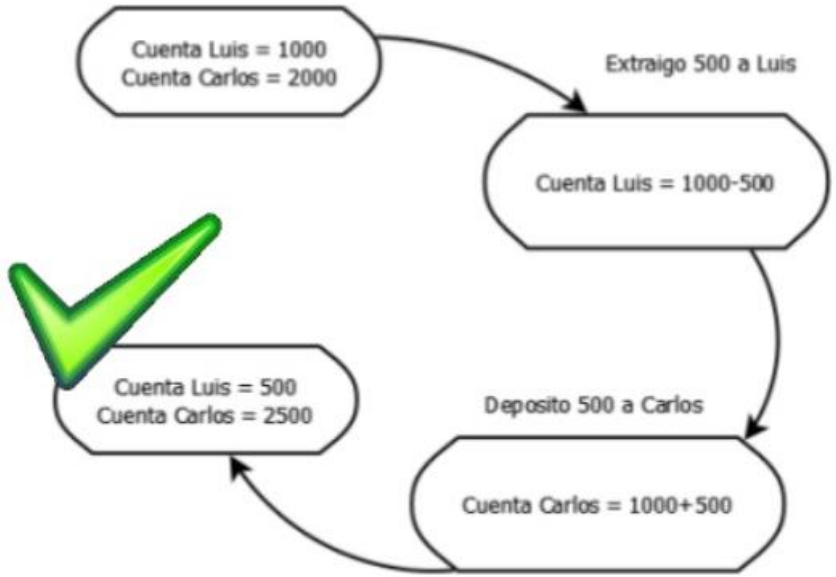
Organización Indexada: Árboles

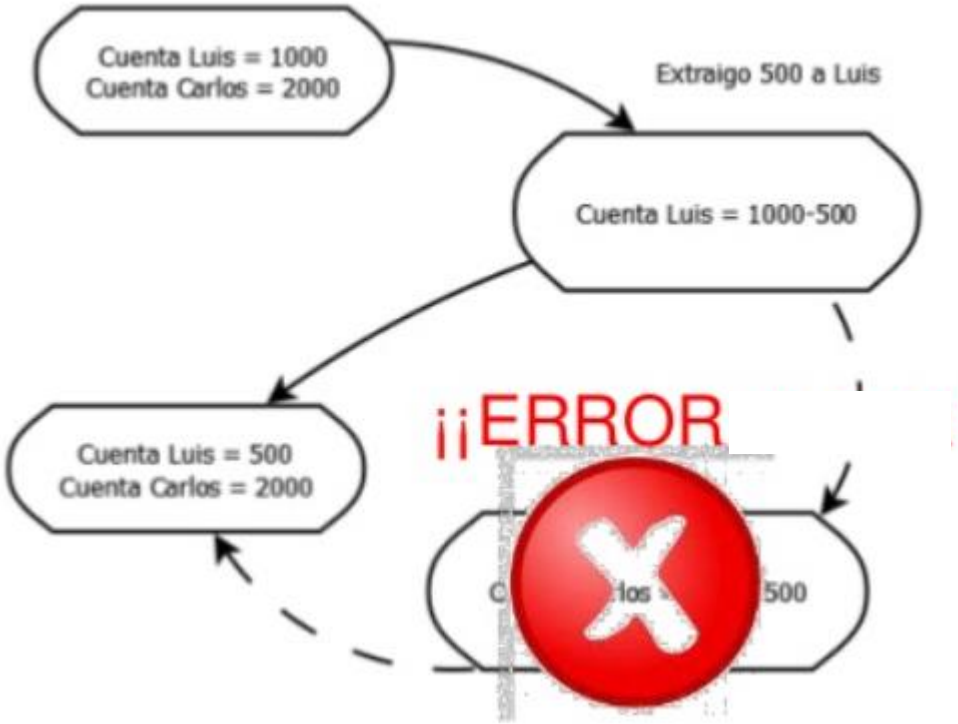


Organización Indexada: Hash



Transacciones

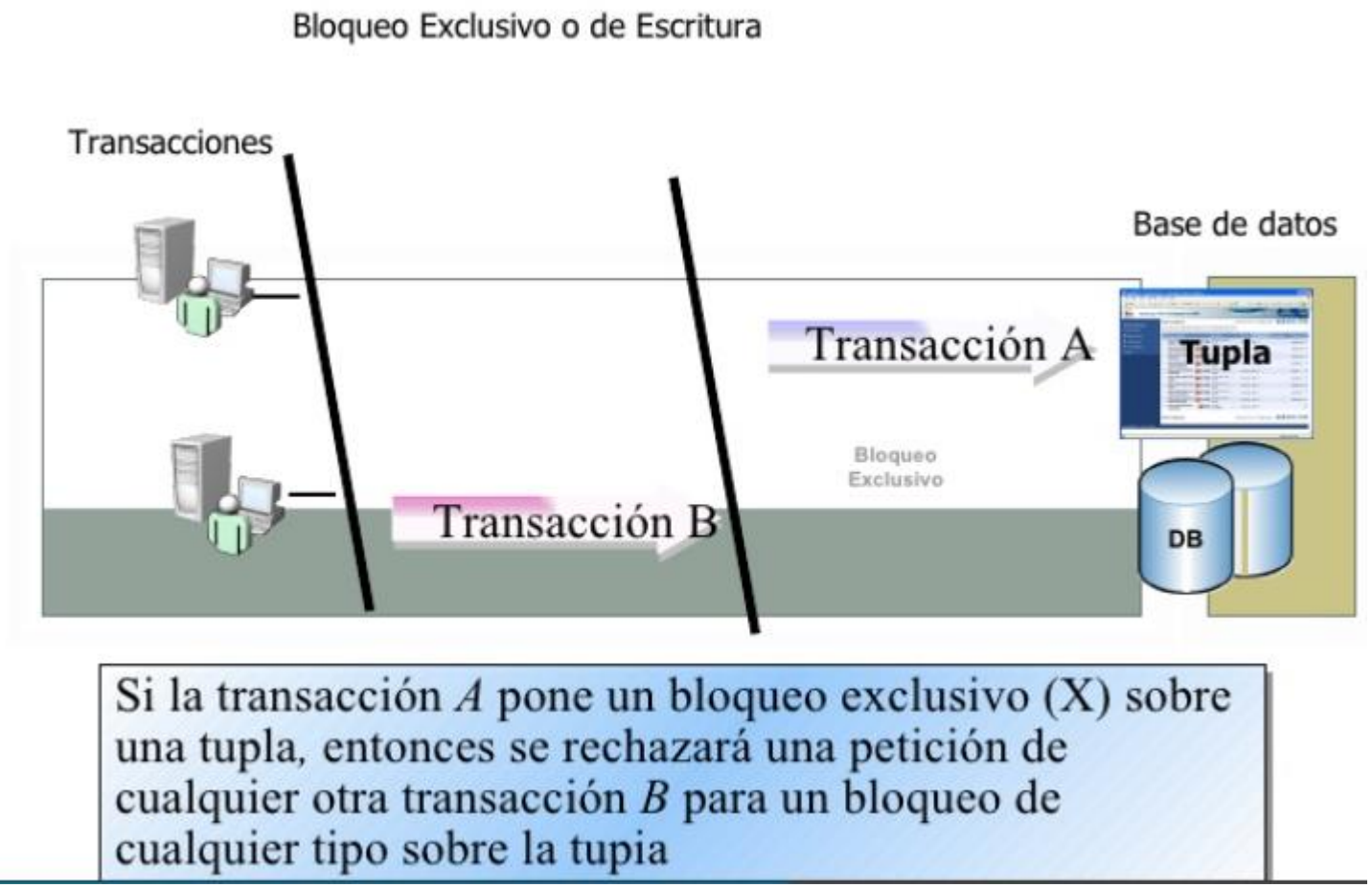




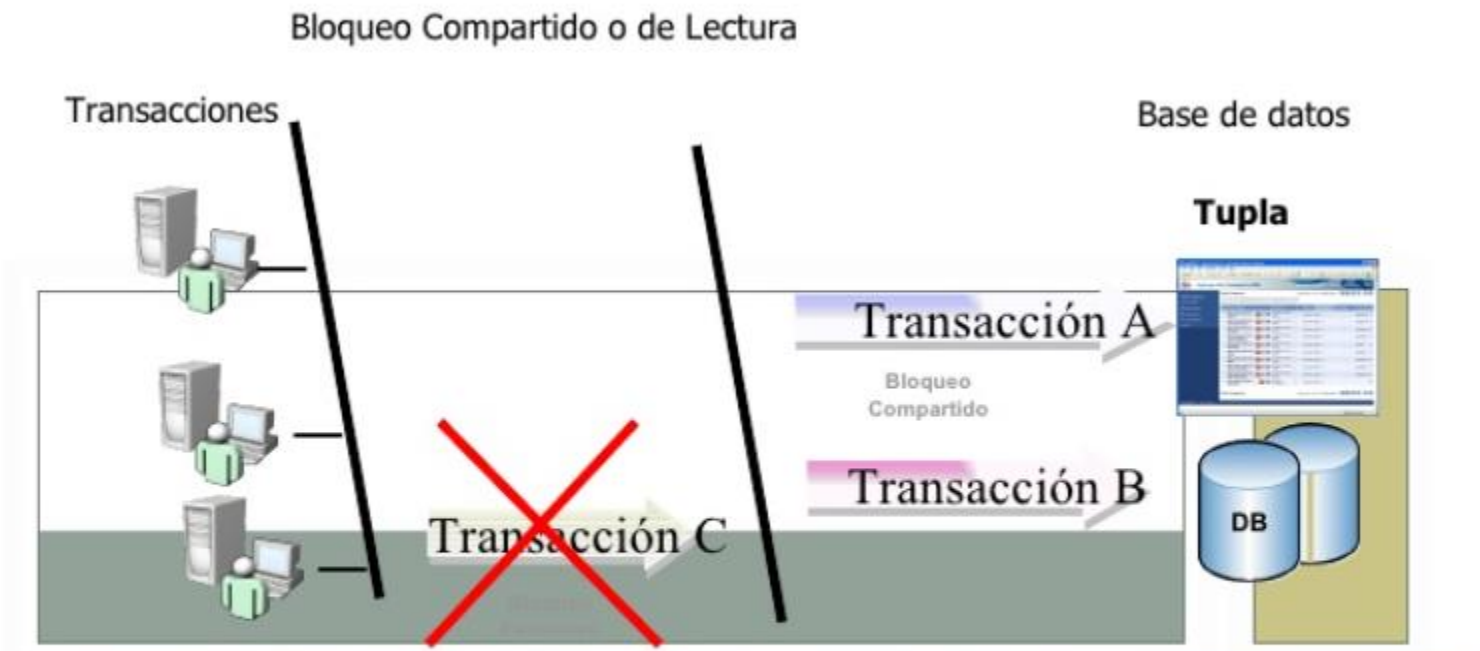
Transacciones

Transacción: secuencia de operaciones que se Ejecuta en forma atómica

Transacciones - Concurrencia



Transacciones - Concurrencia



Si la transacción *A* pone un bloqueo compartido (S) sobre la tupla entonces:
Se rechazará una petición de cualquier otra transacción *B* para un bloqueo Exclusivo sobre la tupla.
Se otorgará una petición de cualquier otra transacción *B* para un bloqueo S sobre la tupla (esto es, ahora también *B* tendrá un bloqueo S sobre la tupla

¿Cómo Afecta esto a un DW?

Diseño Físico:

- OLTP: Optimización mediante índices; necesidad de concurrencia, transacciones (ej., locking), recuperación (WAL). **Foco: transacciones (ej: transferir dinero de una cuenta a otra).**
  - **Optimización mediante índices**
  - **Consultas tipo SELECT * FROM T WHERE A = 'x'.**
- OLAP: Típicamente sólo lectura. No hay necesidad de control de concurrencia. **Foco: Consultas.**
  - **Optimización mediante índices + VISTAS MATERIALIZADAS + particiones**
  - **Consultas tipo SELECT Region, SUM (saldo) FROM Cuentas Group BY Region)**

Resumen

	Aspect	Operational databases	Data warehouses
1	User type	Operators, office employees	Managers, executives
2	Usage	Predictable, repetitive	Ad hoc, nonstructured
3	Data content	Current, detailed data	Historical, summarized data
4	Data organization	According to operational needs	According to analysis needs
5	Data structures	Optimized for small transactions	Optimized for complex queries
6	Access frequency	High	From medium to low
7	Access type	Read, insert, update, delete	Read, append only
8	Number of records per access	Few	Many
9	Response time	Short	Can be long
10	Concurrency level	High	Low
11	Lock utilization	Needed	Not needed
12	Update frequency	High	None
13	Data redundancy	Low (normalized tables)	High (denormalized tables)
14	Data modeling	UML, ER model	Multidimensional model