



DECSAI

Departamento de Ciencias de la Computación e I.A.

Universidad de Granada



Data Warehousing

© Fernando Berzal, berzal@acm.org

Acceso a los datos



- Bases de datos relacionales: SQL
- O/R Mapping
- Bases de datos distribuidas
- Bases de datos NoSQL
- Bases de datos multidimensionales: Data Warehousing



Data Warehousing



- OLAP vs. OLTP
- Data Warehousing
- El modelo multidimensional
- Implementación de un data warehouse
- Soluciones de data warehousing
- Apéndice: Business Intelligence



OLAP vs. OLTP



OLTP [On-Line Transaction Processing]

Aplicaciones típicas de gestión

- Tareas repetitivas.
- Tareas muy bien estructuradas.
- Transacciones cortas (actualizaciones generalmente).



OLAP vs. OLTP



OLTP [On-Line Transaction Processing]

Prioridad: **Gestión de transacciones**

- Las transacciones se realizan sobre grandes *bases de datos* a las cuales se puede acceder eficientemente mediante índices, ya que cada operación afecta sólo a unos pocos registros.
- Es de vital importancia garantizar la “acidez” de las transacciones (atomicidad, consistencia, aislamiento y durabilidad).



OLAP vs. OLTP



OLAP [On-Line Analytical Processing]

Sistemas de ayuda a la decisión (DSS)

- Consultas muy complejas (grandes volúmenes de datos y uso defunciones de agregación).
- Actualizaciones poco frecuentes.



OLAP vs. OLTP



OLAP [On-Line Analytical Processing]

Prioridad: **Procesamiento de consultas**

- Los data warehouses (DW) almacenan datos resumidos de tipo histórico.
- La optimización de las consultas y el tiempo de respuesta son primordiales.



OLAP vs. OLTP



	OLTP	OLAP
Usuarios	Operadores	"Trabajadores del conocimiento"
Función	Operaciones cotidianas	Soporte a la toma de decisiones
Diseño	Orientado a las aplicaciones	Orientado al usuario
Datos	Actuales Actualizados Detallados	Históricos Consolidados Resumidos
Uso	Repetitivo	Ad-hoc
Acceso	Consultas simples y actualizaciones	Consultas complejas
Rendimiento	Transacciones ACID	Consultas: Throughput & tiempo de respuesta
Volumen	GB - TB	TB - PB



Data Warehousing



Problema

Las organizaciones manejan enormes cantidades de datos...

- ... en distintos formatos.
- ... que residen en distintas bases de datos.
- ... organizados utilizando distintos tipos de gestores de bases de datos

Consecuencia

Resulta difícil acceder y utilizar todos los datos en aplicaciones de análisis (las cuales requieren extraer, preparar e integrar los datos)



Data Warehousing



Diseño de procesos e implementación de herramientas que proporcionen información completa, oportuna, correcta y entendible en la toma de decisiones.

"A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."

— **W. H. Inmon**



Data Warehousing



Características de un DW

- Orientado a un tema (clientes, productos, ventas): Vista de la BD operativa excluyendo los datos que no son útiles en la toma de decisiones.
- Integrado (ETL) a partir de múltiples fuentes de datos heterogéneas, p.ej. OLTP (RDBMS, NoSQL, ficheros...)
- Perspectiva histórica (mayor horizonte temporal que una base de datos OLTP).
- No volátil (no se realizan actualizaciones, por lo que no se requieren mecanismos de procesamiento de transacciones, control de concurrencia...).



Data Warehousing



- El DW se mantiene separado de las bases de datos operativas.
- El DW consolida datos históricos para su análisis.
- El DW accede a fuentes de datos heterogéneas, para lo que utiliza **herramientas ETL** [extract-transform-load]: limpieza, filtrado y transformación de los datos.
- Únicas operaciones: carga inicial de los datos y realización de consultas.



Data Warehousing



¿Por qué se mantiene separado el DW?

Distintos requisitos operativos:

- DBMS optimizado para OLTP: Métodos de acceso, indexación, control de concurrencia, transacciones...
- DW optimizado para OLAP: Consultas complejas, consolidación de datos, datos históricos...



Data Warehousing

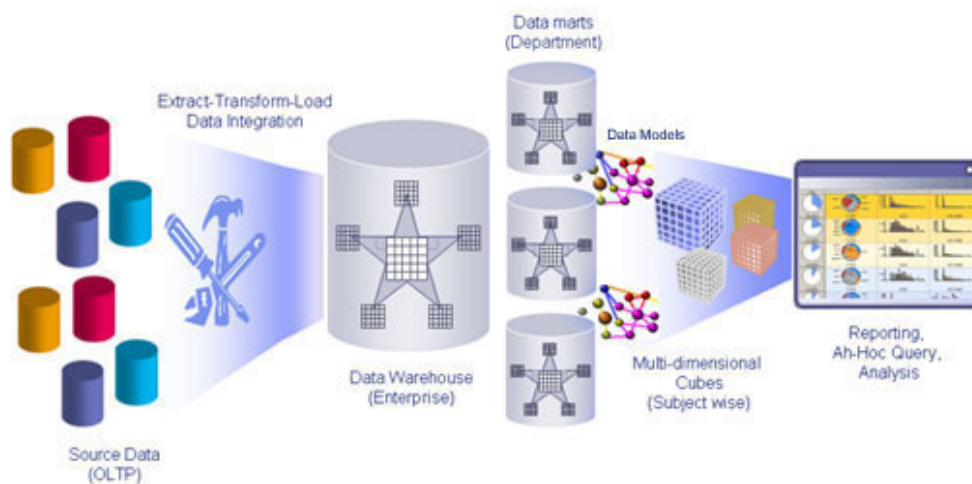


Modelos arquitectónicos

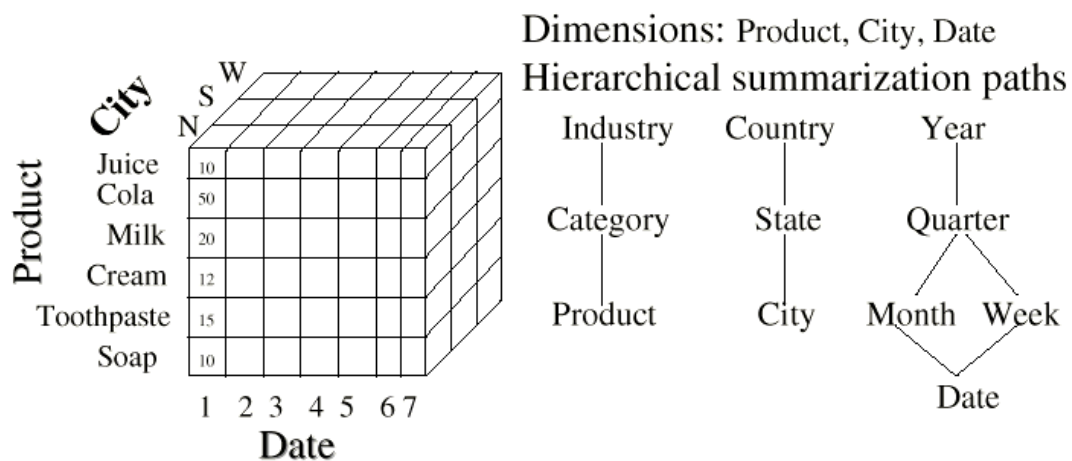
- **Enterprise warehouse**
(único DW para toda la organización).
- **Data marts**
(varios DW para grupos específicos de usuarios).
- **Virtual warehouse**
(vistas sobre las bases de datos operativas, de las cuales sólo se materializan algunas de ellas).



Data Warehousing



El modelo multidimensional



El modelo multidimensional



Los datos en un DW se modelan en cubos de datos [data cubes], estructuras multidimensionales (hipercubos, en concreto) cuyas operaciones más comunes son:

- **Roll up**
(incremento en el nivel de agregación de los datos).
- **Drill down**
(incremento en el nivel de detalle, opuesto a roll up).
- **Slice**
(reducción de dimensionalidad mediante selección).
- **Dice**
(reducción de dimensionalidad mediante proyección).
- **Pivotaje o rotación** (reorientación de la visión multidimensional de los datos).



El modelo multidimensional



Modelado multidimensional

Modelos de datos
como conjuntos de medidas descritas por dimensiones.

- Adecuado para resumir y organizar datos
(generalización de las hojas de cálculo).
- Enfocado para trabajar sobre datos de tipo numérico.
- Más simple, más fácil de visualizar
y de entender que el modelado E/R.



El modelo multidimensional

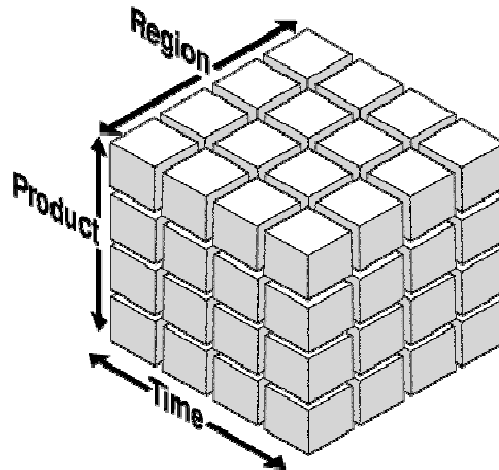


Dimensiones

Perspectivas o entidades respecto a las cuales una organización quiere mantener sus datos organizados.

Ejemplos:

- Tiempo
- Localización
- Clientes
- Proveedores
- Productos



El modelo multidimensional



Miembros

Nombres o identificadores que marcan una posición dentro de la dimensión.

Ejemplos:

- Meses, trimestres y años son miembros de la dimensión tiempo.
- Ciudades, regiones y países son miembros de la dimensión localización.

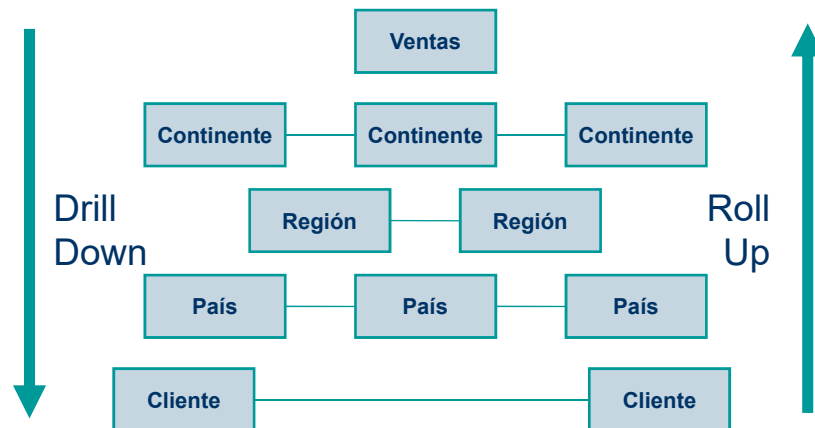


El modelo multidimensional



Jerarquías

Los miembros de las distintas dimensiones se suelen organizar en forma de jerarquías.



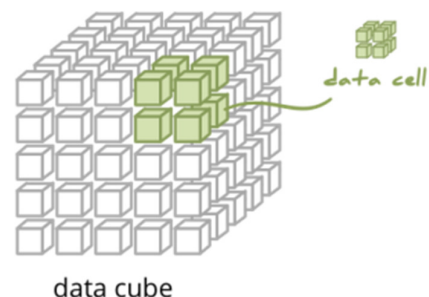
El modelo multidimensional



Hechos

Colecciones de datos relacionados compuestas por medidas y un contexto.

- Las dimensiones determinan el contexto de los hechos.
- Cada hecho particular está asociado a un miembro de cada dimensión.



Medidas

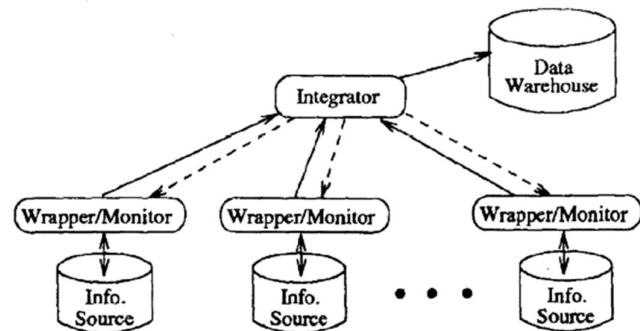
Atributos numéricos asociados a los hechos (lo que realmente se mide).



Implementación de un DW



Wrappers



- Los wrappers (encapsuladores) se encargan de extraer los datos de las distintas fuentes y transmitirlos al DW.
- Los monitores están en contacto directo con las fuentes de datos para detectar los cambios que se puedan producir en ellas.
- El integrador es el responsable de filtrar, resumir y unificar los datos de las distintas fuentes.



22

Implementación de un DW



Metadatos

- Estructura del DW: esquema, vistas, dimensiones, jerarquías, datos derivados, data marts (localización y contenidos)...
- Metadatos operativos: "linaje de los datos" [data lineage], actualidad de los datos (activos, archivados, purgados) e información de monitorización (estadísticas de uso, informes de errores y auditorías).
- Correspondencia entre el entorno operativo y el DW (p.ej. algoritmos utilizados para resumir los datos).
- Datos del negocio [business data]: Términos y definiciones, propiedad de los datos...



23

Implementación de un DW



Alternativas de implementación

■ MOLAP [Multidimensional OLAP]

Datos almacenados en estructuras de datos multidimensionales (matrices multidimensionales sobre las que se realizan directamente las operaciones OLAP).

■ ROLAP [Relational OLAP]

DW implementado como una base de datos relacional (las operaciones multidimensionales OLAP se traducen en operaciones relacionales estándar).



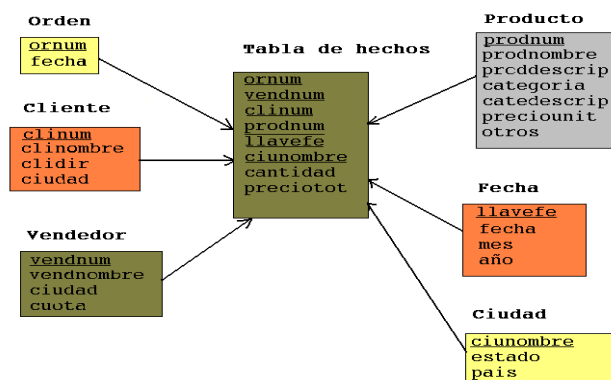
24

Implementación de un DW



ROLAP con esquema en estrella [star]

Una tabla de hechos
y una tabla adicional (denormalizada) por cada dimensión.

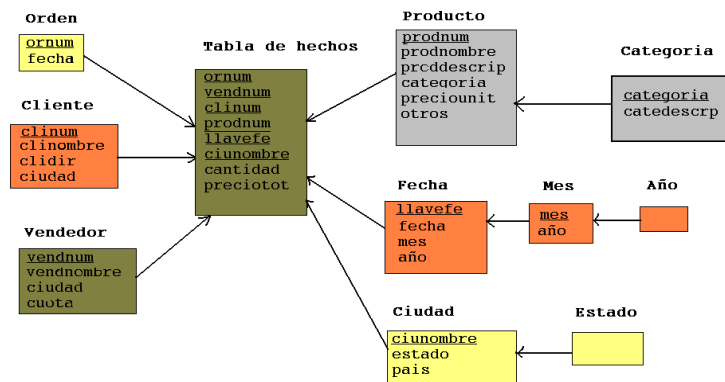


25

Implementación de un DW

ROLAP con esquema en copo de nieve [snowflake]

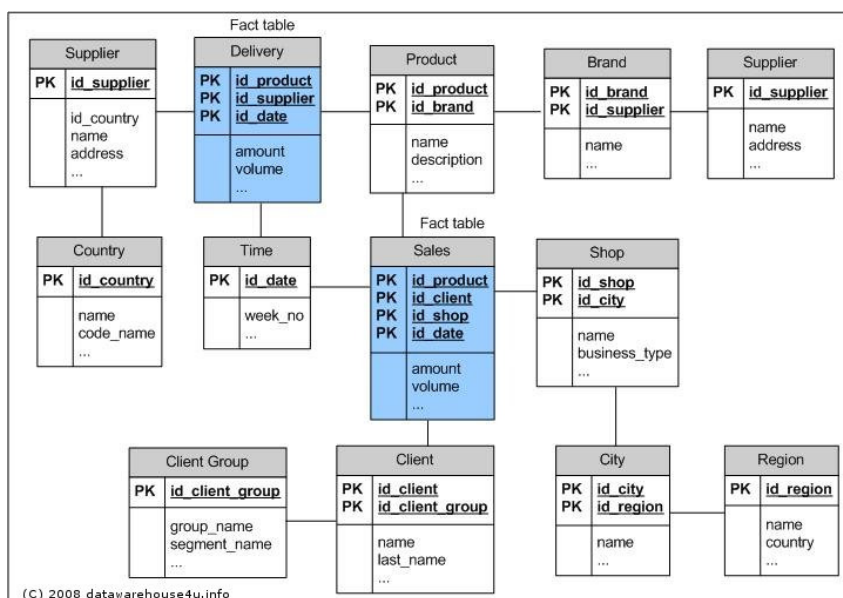
Refleja la organización jerárquica de las dimensiones...



Implementación de un DW

ROLAP con constelaciones de hechos

Múltiples tablas de hechos que comparten dimensiones



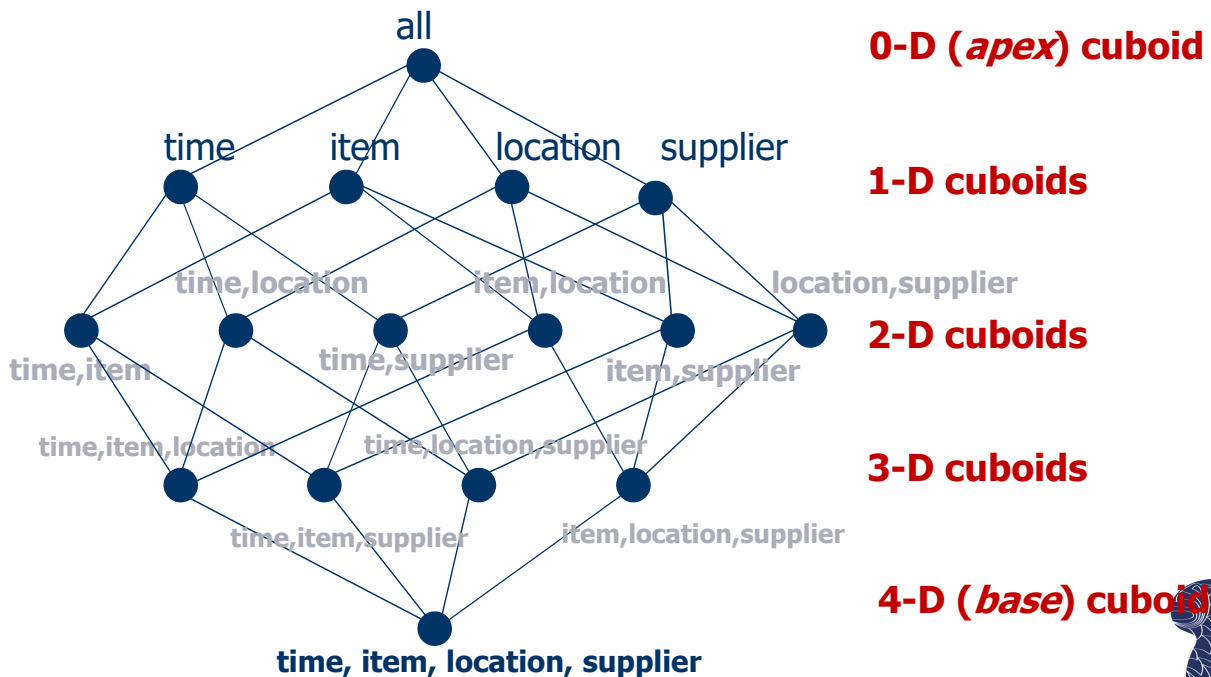
(C) 2008 datawarehouse4u.info



Implementación de un DW



Un cubo de datos como un retículo de cuboides:



28

Implementación de un DW



Un cubo de datos como un retículo de cuboides:

- El cuboide base tiene D dimensiones.
- El cuboide ápice tiene 0 dimensiones (1 celda).

Materialización del cubo de datos:

- Completa (todos los cuboides).
- Ninguna (ningún cuboide, i.e. sólo el cuboide base).
- Parcial (algunos cuboides materializados, que se seleccionan en función de su tamaño, uso en distintas consultas, frecuencia de acceso...)



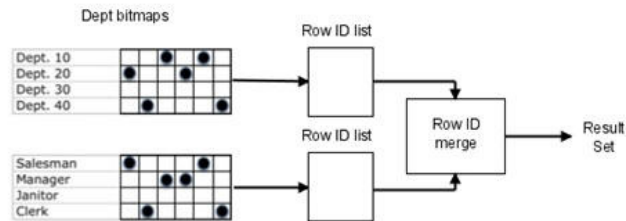
29

Implementación de un DW

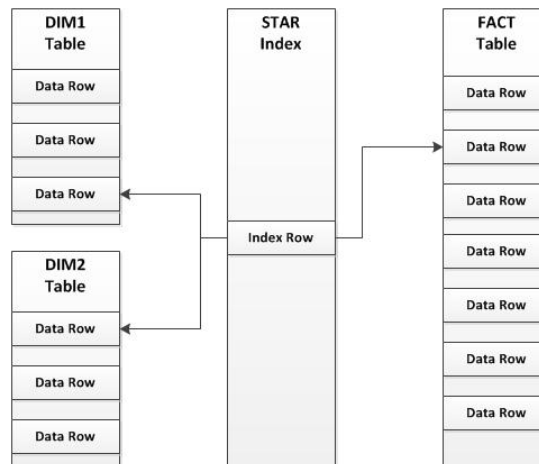


Indexación de datos

- Índice bitmap



- Índice de reunión [join index], a.k.a. star index



Implementación de un DW



Procesamiento de consultas OLAP

- Se determinan las operaciones que deben realizarse sobre cuboides: se transforman las operaciones sobre cubos de datos (roll up, drill down...) en operaciones relacionales

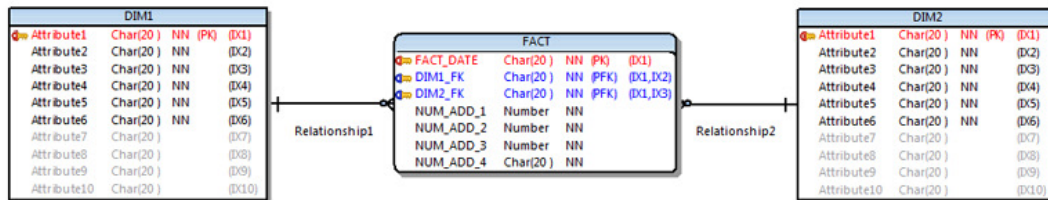
p.ej. dice = selección + proyección

- Se determinan los cuboides materializados que pueden utilizarse para resolver mejor la consulta.



Implementación de un DW

Ejemplo: ROLAP sobre Oracle



Índice bitmap tradicional

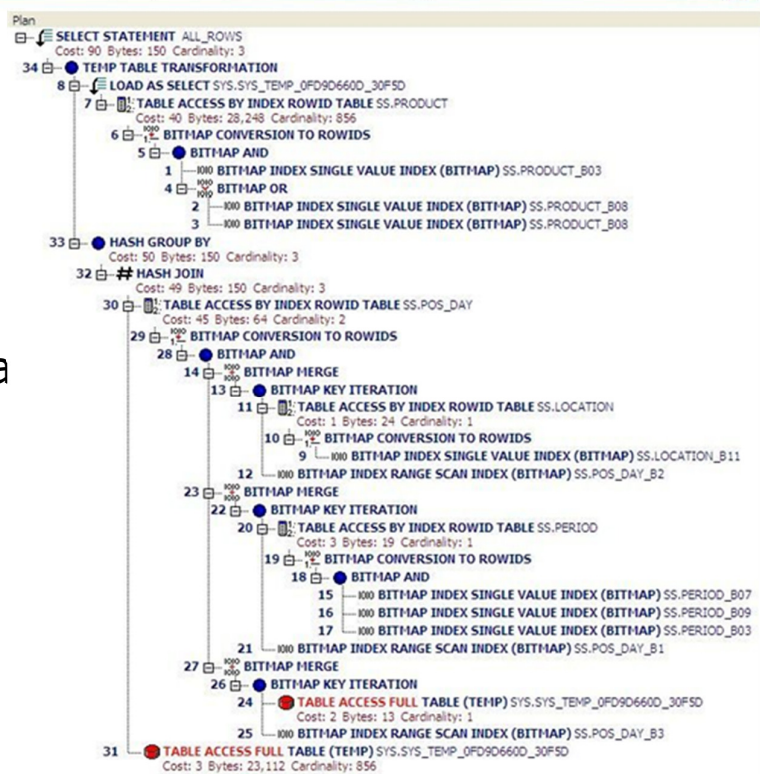
```
CREATE UNIQUE INDEX FACT_PK ON FACT (FACT_DATE, DIM1_FK, DIM2_FK);  
CREATE BITMAP INDEX DIM1_FK ON FACT ( DIM1_FK );  
CREATE BITMAP INDEX DIM2_FK ON FACT ( DIM2_FK );
```



Implementación de un DW

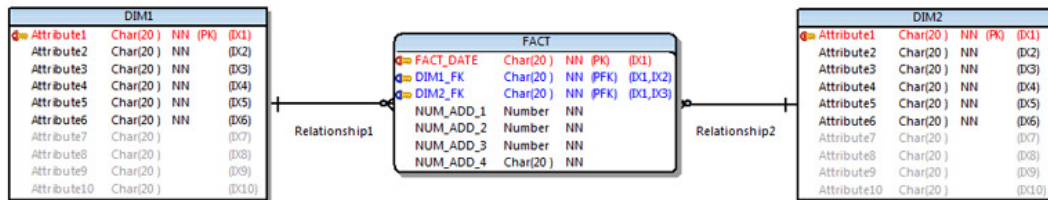
Ejemplo: ROLAP sobre Oracle

Plan de
ejecución
de una consulta
(bitmap index)



Implementación de un DW

Ejemplo: ROLAP sobre Oracle



Índice bitmap join (uno por cada dimensión)

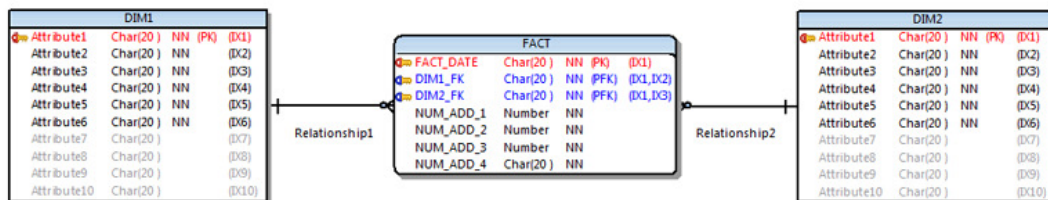
```
CREATE UNIQUE INDEX FACT_PK ON FACT (FACT_DATE, DIM1_FK, DIM2_FK);
CREATE BITMAP INDEX FACT_BJ1 ON FACT ( DIM1_FK )
FROM FACT, DIM1
WHERE FACT.DIM1_FK = DIM1.ATTRIBUTE1;
CREATE BITMAP INDEX FACT_BJ2 ON FACT ( DIM2_FK )
FROM FACT, DIM2
WHERE FACT.DIM2_FK = DIM2.ATTRIBUTE1;
```



34

Implementación de un DW

Ejemplo: ROLAP sobre Oracle



Índice bitmap join (único para las dos dimensiones)

```
CREATE UNIQUE INDEX FACT_PK ON FACT (FACT_DATE, DIM1_FK, DIM2_FK);
CREATE BITMAP INDEX FACT_BJ ON FACT ( DIM1_FK, DIM2_FK )
FROM FACT, DIM1, DIM2
WHERE FACT.DIM1_FK = DIM1.ATTRIBUTE1
AND FACT.DIM2_FK = DIM2.ATTRIBUTE1;
```

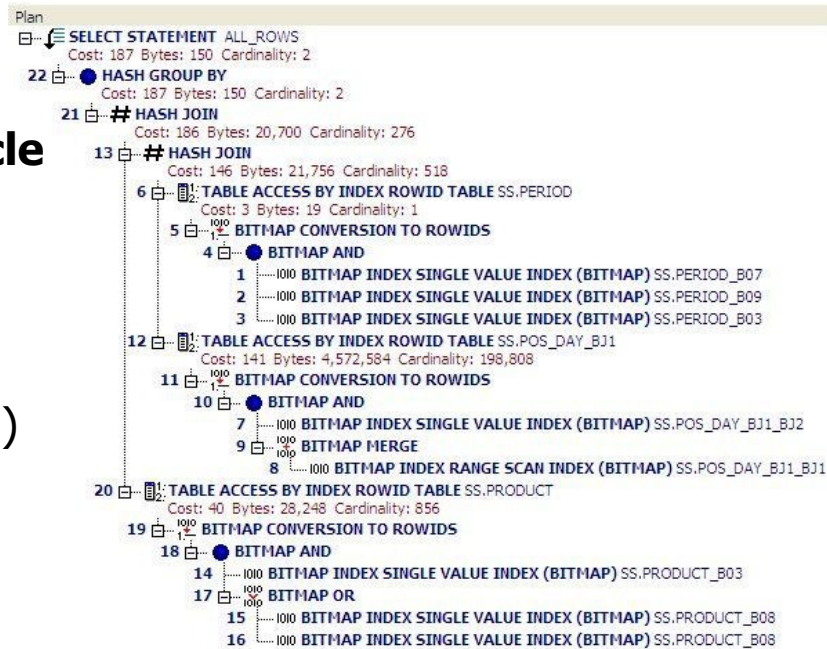


35

Implementación de un DW

Ejemplo: ROLAP sobre Oracle

Plan de
ejecución
de una
consulta
(bitmap join)



36

Soluciones DW

MPP [Massive Parallel Processing]

- Shared-nothing architectures vs. SMP [Symmetric Multiprocessing]
- Escalabilidad horizontal (añadiendo nodos).
- Descomposición de consultas (procesamiento paralelo en varios nodos).
- Mayor tasa de ingestión de datos (movimiento de datos en paralelo).

37

Soluciones DW



Mercado

- Proveedores especializados:
Teradata, Netezza, Vertica, Greenplum
- Proveedores de gestores de bases de datos:
Microsoft PDW [Parallel Data Warehouse],
IBM DB2 UDB with DPF [DB Partitioning Feature]
Oracle Exadata & Oracle Big Data Appliance
- Soluciones híbridas con Hadoop: Impala,
Stinger, Apache Drill, Shark, Hadapt, Teradata
SQL-H (Aster Data), EMC HAWK, IBM BigSQL...



Soluciones DW



Costes relativos de distintas plataformas

Teradata	Hardware and licenses the most expensive of all options. Staff costs can be expensive and it takes a great deal of effort to configure and administer.	
IBM Netezza	Hardware and licenses used to be much less than Teradata, but prices have been converging. Some of the highest staff cost due to scarcity, but that's tempered by lower effort for configuration and admin of single purpose appliance.	
Greenplum	Commodity hardware. Moderately priced licenses. Few Greenplum specialists, but can be staffed by PostgreSQL DBAs and developers.	
Vertica	Commodity hardware. Moderately priced licenses, but special purpose orientation limits usefulness. Few specialists, but can be staffed by traditional DBAs and developers.	
Hadoop HBase	Commodity hardware and no license cost, resulting in lowest up-front cost. Likely to buy more hardware for redundancy and load. But requires highly technical staff and implementation is less productive than more mature options.	

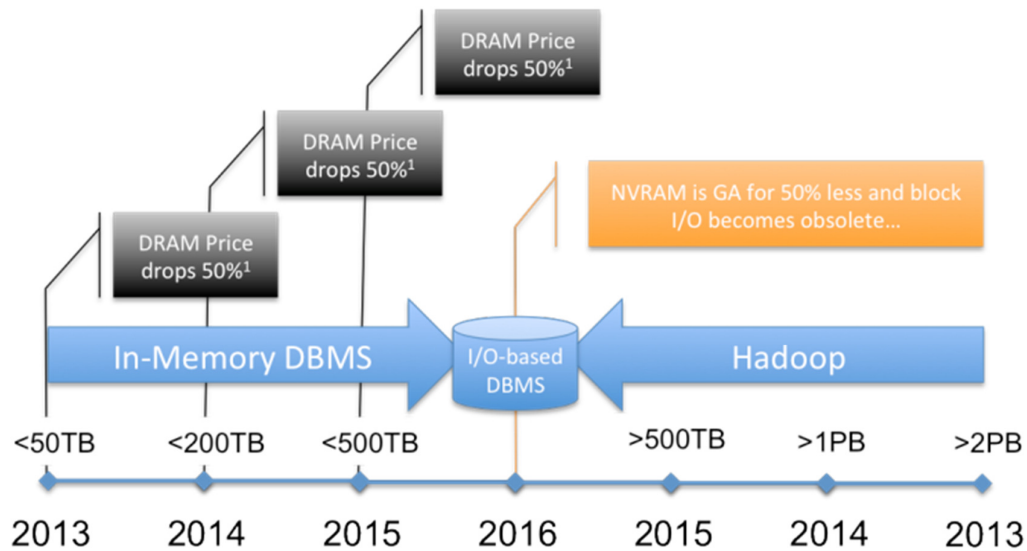
Hardware
a medida
(p.ej. FPGA)

Open
Source



Soluciones DW

Tendencias...



¹ Gartner Forecast: Memory, Worldwide, 2006-2016, 2Q12 Update

Fig. 1 – The Squeeze

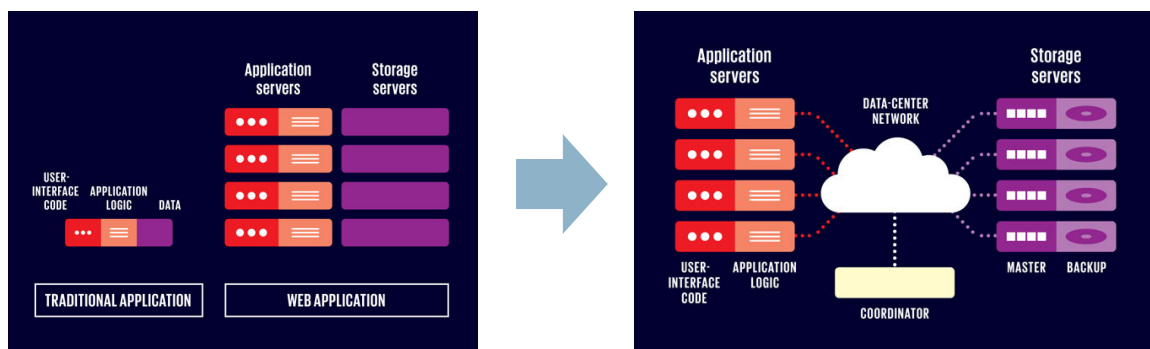


En el futuro...

RAMCloud

Stanford University

"a general-purpose storage system...
which keeps all of its data in DRAM at all times."



A Radical Proposal: Replace Hard Disks With DRAM

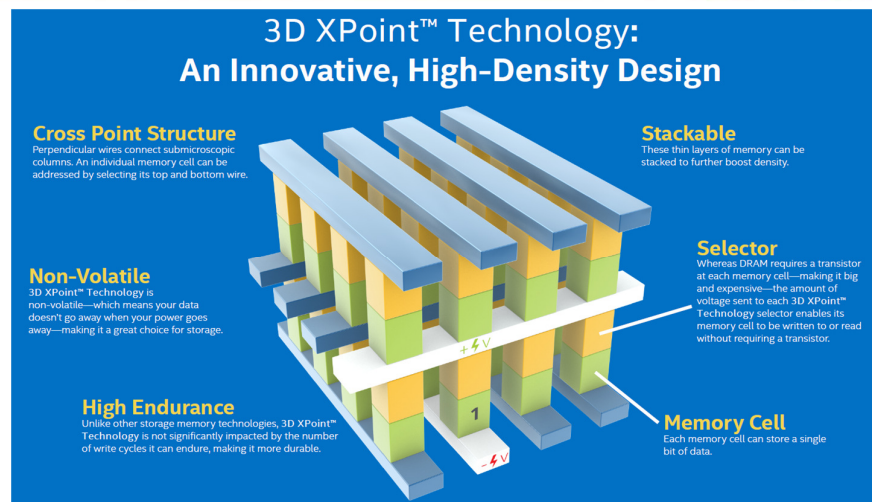
IEEE Spectrum, October 2015

<http://spectrum.ieee.org/computing/hardware/a-radical-proposal-replace-hard-disks-with-dram>



En el futuro...

3D XPoint Intel & Micron



- Memoria no volátil.
- Más rápida que la memoria flash, más lenta que la memoria DRAM.

https://en.wikipedia.org/wiki/3D_XPoint

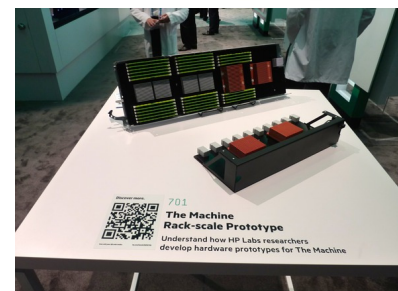


En el futuro...

"The Machine" Hewlett-Packard Labs

"Universal Memory" Say goodbye to disk drives

- The goal: "... with this architecture we can ingest, store, manipulate truly massive datasets while simultaneously achieving multiple orders of magnitude less energy per bit."
- The biggest missing piece: "the memristor - a resistor that stores information after losing power and that would allow computers to store and retrieve large datasets far more rapidly than is possible today."

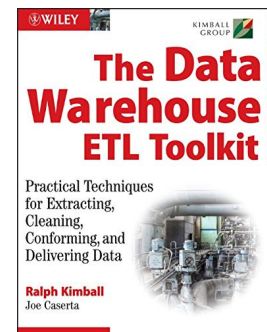
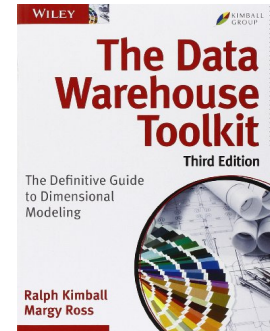


Inside The Machine: Hewlett Packard Labs mission to remake computing
<http://www.techrepublic.com/article/inside-the-machine-hewlett-packard-labs-mission-to-remake-computing/>



Bibliografía recomendada

- Ralph Kimball & Margy Ross:
**The Data Warehouse Toolkit:
The Definitive Guide
to Dimensional Modeling.**
Wiley, 3rd edition, 2013.
ISBN 1118530802
- Ralph Kimball & Joe Caserta:
The Data Warehouse ETL Toolkit.
Wiley, 2004.
ISBN 0764567578



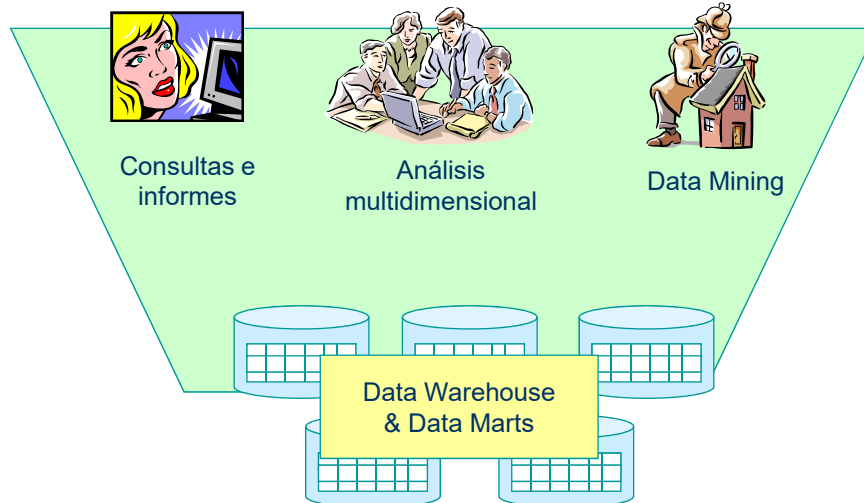
44

Business Intelligence



45

Business Intelligence

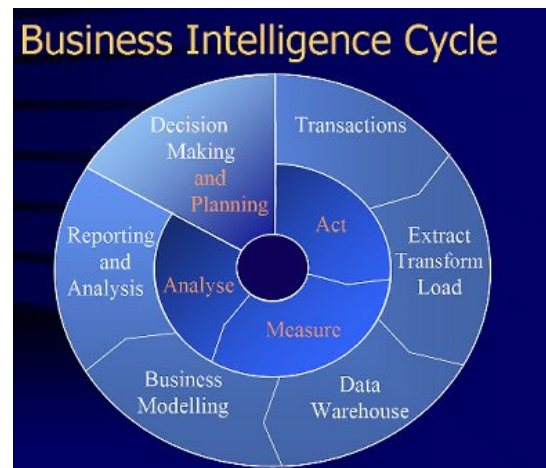


Business Intelligence

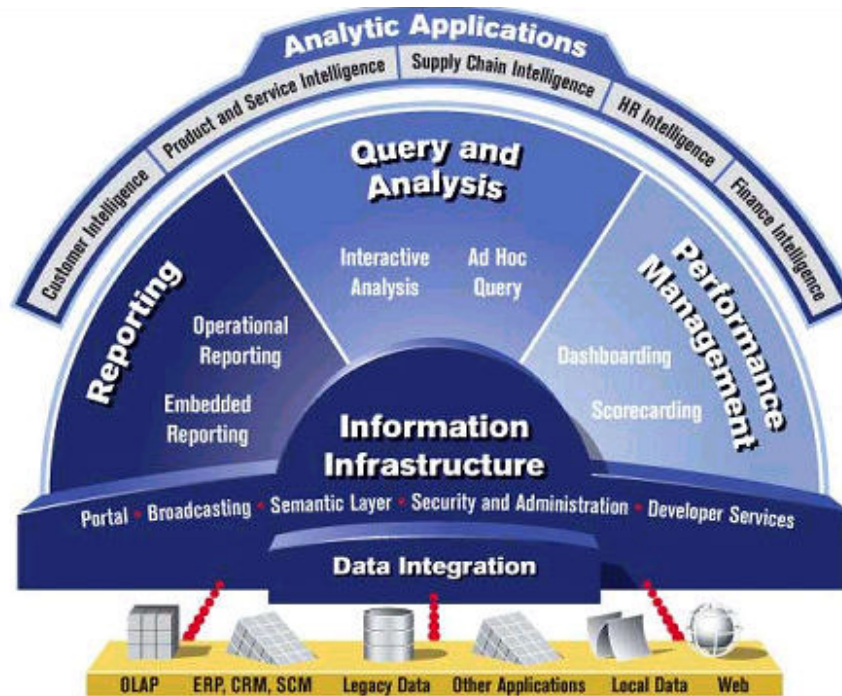


BI [Business Intelligence]

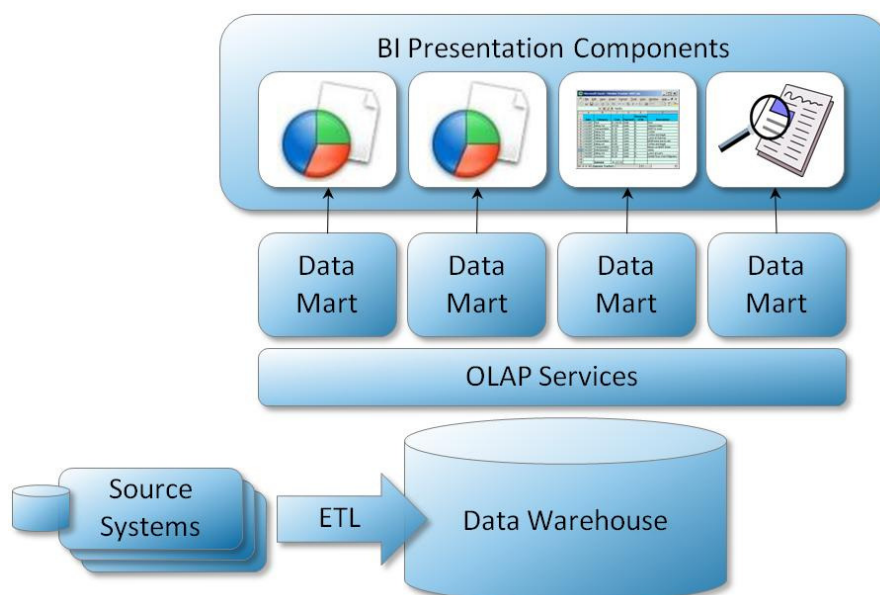
- Recopilación de datos: ETL
- Almacenamiento: DW
- Análisis: Data Mining
- Evaluación
- Diseminación: Informes



Business Intelligence



Business Intelligence



Business Intelligence

