


On-Line Analytical Processing

Javier Bonet

Joel Catacora

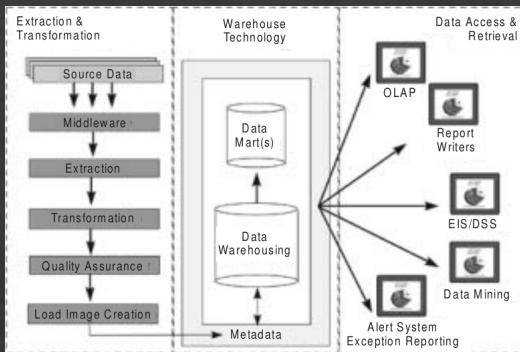
- Base de datos avanzada • 22 de abril del 2015

12 QUANTUM



BREVE REPASO

Recordamos los distintos tipo de tecnologías que mencionamos en la charla *Data Warehouse*.



Arquitectura del Data Warehouse



¿QUÉ ES OLAP?

El **procesamiento analítico en línea** (On-Line Analytical Processing, OLAP) es una solución utilizada en el campo de la inteligencia de negocios, cuyo objetivo es permitir la consulta de grandes cantidades de datos de forma eficiente y sencilla.

El concepto OLAP puede definirse mediante 5 palabras:
Análisis Rápido de Información Compartida Multidimensional,
(Fast Analysis of Shared Multidimensional Information, o
FASMI).

- **Rápida:** el sistema está dirigido a proporcionar la mayoría de las respuestas a los usuarios en pocos segundos.
- **Análisis:** el sistema puede hacer frente a cualquier lógica de negocio y análisis estadístico que sea relevante para el usuario, en forma relativamente sencilla.
- **Compartida:** significa que el sistema debe implementar todos los requisitos de seguridad, para la confidencialidad de los datos.
- **Multidimensionalidad:** el sistema debe proveer una vista conceptual multidimensional de los datos.
- **Información:** se refiere a todos los datos, relevantes para la aplicación.

REGLAS DE CODD

- **Visión multidimensional**

- **Visión multidimensional**
- **Manipulación intuitiva de los datos**

- **Visión multidimensional**
- **Manipulación intuitiva de los datos**
- **Accesibilidad**

- **Visión multidimensional**
- **Manipulación intuitiva de los datos**
- **Accesibilidad**
- **Soporte multi-usuario**

- **Visión multidimensional**
- **Manipulación intuitiva de los datos**
- **Accesibilidad**
- **Soporte multi-usuario**
- **Información separada del origen de datos**

- **Visión multidimensional**
- **Manipulación intuitiva de los datos**
- **Accesibilidad**
- **Soporte multi-usuario**
- **Información separada del origen de datos**
- **Flexibilidad ante valores nulos**

- **Visión multidimensional**
- **Manipulación intuitiva de los datos**
- **Accesibilidad**
- **Soporte multi-usuario**
- **Información separada del origen de datos**
- **Flexibilidad ante valores nulos**
- **Rendimiento uniforme**

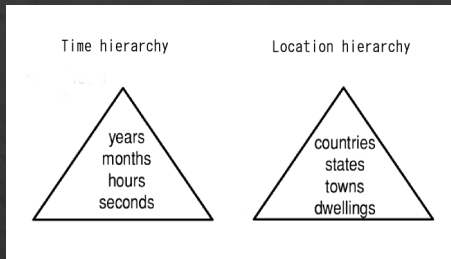
- **Visión multidimensional**
- **Manipulación intuitiva de los datos**
- **Accesibilidad**
- **Soporte multi-usuario**
- **Información separada del origen de datos**
- **Flexibilidad ante valores nulos**
- **Rendimiento uniforme**
- **Dimensiones y niveles de agregación ilimitados**



JERARQUÍAS

Las **jerarquías** se utilizan para especificar en menor o mayor detalle los datos contenidos en la dimensión. Las jerarquías son de mucha utilidad a la hora de calcular valores "agregados"

Ejemplos:



A large, solid black circle is centered on a background of crumpled, light-colored paper. The word "AGREGACIONES" is written in white, uppercase letters across the middle of the black circle.

AGREGACIONES

Un **agregado** es un resumen precalculado, almacenado en un DW, por lo general en un esquema separado. Los agregados típicamente se calculan en base a los registros en el nivel más detallado (o base) de su jerarquía. Se utilizan para mejorar el rendimiento de aquellas consultas que requieren sólo de datos sumariados, o de alto nivel.

Ejemplo:



Esquema de nivel base que usa la parte inferior del nivel de jerarquías dimensionales.



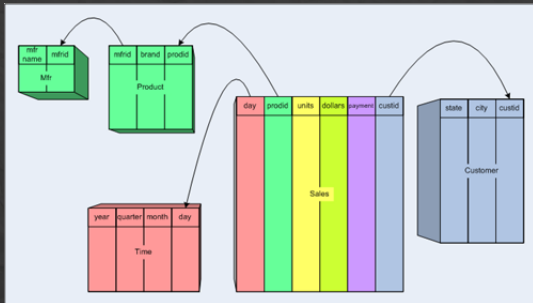
Esquema agregado, que resulta en datos en un nivel más alto en la jerarquía dimensional.

Los **esquemas agregados** proporcionan mejoras en el rendimiento, ya que tienen un número significativamente menor de registros, en relación al esquema que se quiere resumir.

El uso más común de los agregados es tomar una dimensión y cambiar su granularidad. Al cambiar la granularidad de la dimensión, la tabla de hechos tiene que ser parcialmente resumida para adaptarse a la nueva dimensión. Se crean así, nuevas tablas de dimensiones y tablas de hechos, que encajan en este nuevo nivel de granularidad.

Ejemplo:

Este esquema copo de nieve, tiene una sola tabla de hechos *Sales*, dos métricas (*units* and *dollars*) y cuatro tablas de dimensiones (*Product*, *Mfr*, *Customer*, *Time*, y *Customer*).



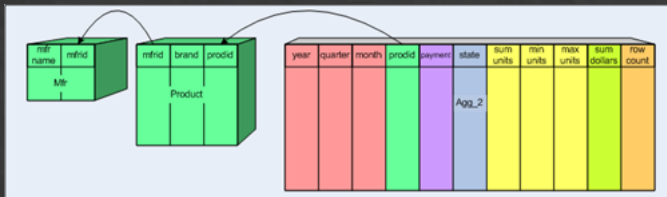
A partir del esquema anterior, creamos una **tabla agregada**, que llamamos **Agg_1**:

year	quarter	mfrid	brand	prodid	sum units	min units	max units	sum dollars	row count
				Agg_1					

Veamos como se combinaron las columnas del esquema original, en la tabla agregada **Agg_1**:

- La dimensión *Time* "colapsó" en la tabla de agregación, omitiendo las columnas *month* y *day*.
- Las dos tablas de la dimensión *Product* "colapsaron" en la tabla de agregación.
- La dimensión *Customer* se "perdió".
- Para cada métrica en la tabla de hechos (*units*, *dollars*), hay uno o más métricas en la tabla de agregación (*sum units*, *min units*, *max units*, *sum dollars*).
- También hay una nueva métrica, *row count*, que representa la métrica "conteo".

Veamos otra posible tabla agregada, **Agg_2**:



Varias dimensiones colapsaron: *Time* en el nivel *month*; *Customer* en el nivel *state*; y *Payment Method* a nivel *payment method*. Mientras que la dimensión *Product* se mantuvo tal como estaba en el esquema original.

Tener datos agregados en el modelo dimensional hace que el entorno sea más complejo. Para que esta complejidad adicional sea transparente al usuario, se utiliza una funcionalidad conocida como *navegación de agregados*, la cual es implementada por el motor OLAP, para consultar las tablas dimensionales y de hecho, con el nivel de granularidad correcto.

A large, solid black circle is centered on a background of crumpled, light-colored paper. The circle contains the text "HIPERCUBO DE DATOS" in white, uppercase letters.

HIPERCUBO DE DATOS

OLAP utiliza **hipercubos** o **cubos**, de la misma manera que las bases de datos utilizan tablas. Toda la navegación, informes y análisis se hacen en términos de hipercubos. Por lo tanto, un "cubo" de datos se refiere a una representación multidimensional de los datos en dicha forma (obviamente, sólo en el espacio tridimensional).

Construiremos un cubo, tomando como ejemplo el siguiente conjunto de datos:

Month	Sales	Direct Costs	Indirect Costs	Total Costs	Margin
January	790	480	110	590	200
February	850	520	130	650	200
March	900	530	140	670	230
April	910	590	150	740	170
May	860	600	120	720	140
June	830	490	100	590	240
July	880	500	110	610	270
August	900	620	130	750	150
September	790	300	90	390	400
October	820	540	100	640	180
November	840	570	150	720	120
December	810	600	120	720	90
Total	10,180	6,340	1,450	7,790	2,390

Decimos que *sales*, *costs*, y *margin* representan **variables**.

Si alguien pregunta, "¿Qué estás midiendo?".

Le respondemos, "ventas, costos y márgenes".

Ante la pregunta, "¿De dónde consigue sus datos?", o "¿con qué frecuencia está haciendo mediciones?".

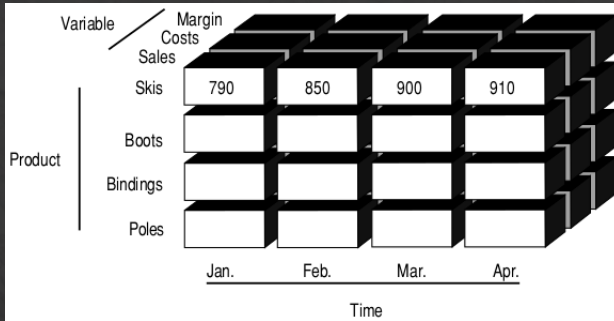
Responderíamos, "Estamos siguiendo las ventas mensuales."

Los meses representan la organización de los datos.

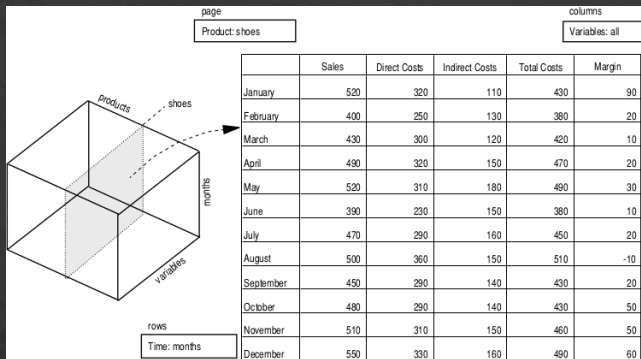
Entonces, hay dos dimensiones, el *tiempo* y las *variables*.

Este enfoque de **dimensiones genéricas**, utiliza invariablemente una **dimensión de hechos**, o **variables**. Tratar a los **hechos** como miembros de una dimensión, implica la creación de una **dimensión de datos**.

¿Qué sucede cuando añadimos una tercera dimensión llamada *products* (productos)?. Tenemos un cubo.



El conjunto de datos tridimensionales, que consiste de *variables*, *time*, y *products*, se puede mostrar en una pantalla en términos de: **fila**, **columna** y **página**.



¿Qué pasa si intentamos añadir una cuarta dimensión *tienda*, al cubo?. El cubo como metáfora visual se rompe.

Presentamos una nueva estructura para representar datos, o eventos generadores de datos, que es capaz de reproducir cualquier número de dimensiones.

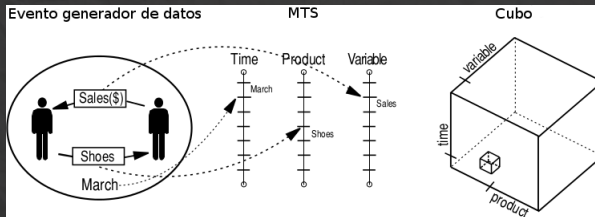
La llamamos, **estructura de tipo multidimensional** (*Multidimensional Type Structures*, o MTS).

Cada dimensión está representada por una línea. Cada miembro dentro de una dimensión está representado por un intervalo, dentro del segmento correspondiente.

Siguiendo el ejemplo, tenemos tres líneas:

Una para el *tiempo*, una de los *productos*, y por último, una para las *variables*.

Cualquier unión de los intervalos de cada una de los tres líneas, está conectada a un evento y a un elemento del cubo.

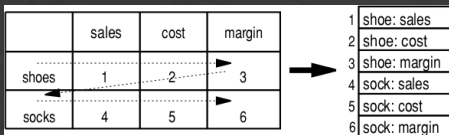


A large, solid black circle is centered on a background of crumpled, light-colored paper. The text "HIPERCUBOS EN UNA PANTALLA" is written in white, uppercase letters across the middle of the black circle.

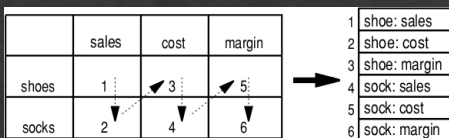
HIPERCUBOS EN UNA PANTALLA

Para ver los datos en la pantalla, tenemos que mapear múltiples dimensiones lógicas en dos dimensiones físicas (la pantalla).

Mapear dos dimensiones en una dimensión, implica crear una versión unidimensional de las dos dimensiones. El método típico consiste en anidar una dimensión dentro de la otra.



Variables anidadas en productos.



Productos anidados en variables

Efectos de combinar dimensiones:

- Cambia la forma de los datos visibles. La longitud de una lista unidimensional es igual al producto de las longitudes de cada una de las dos dimensiones.
- Cambia el conjunto de vecinos que rodean cualquier punto.

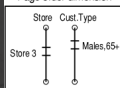
	Jan	Feb	Mar
Ridgewood	555	611	677
Newbury	490	539	598
Avon	220	242	271

Jan	Ridgewood	555
	Newbury	490
	Avon	220
Feb	Ridgewood	611
	Newbury	539
	Avon	242
Mar	Ridgewood	677
	Newbury	598
	Avon	271

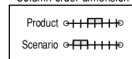
Siguiendo el ejemplo, que ahora consiste de: *products, times, stores, customers, variables, y escenarios*; vemos una posible representación, en base a filas, columnas y páginas:

Store	Cust.Type	Variable	Time	Scenario	Product
Store 1	18 males (1)	Margin	Jan.	Actual (1)	Tables
Store 2			Feb.		Desks
Store 3	18 females (2)	Total sales	Mar.		Chairs
Store 4			Apr.	Planned (2)	Lamps
Store 5	Adults (3)	Direct sales	May		Shirts
Store 6			Jun.		Shoes
Store 7	65+ males (4)	Indirect sales	Jul.	Variance (3)	Socks
Store 8			Aug.		Caviar
			Sep.		Coffee
			Oct.		
			Nov.		
			Dec.		Wine

Page order dimension



Column order dimension



Row order dimension



Store 3	Males, 65+		Desks		Lamps
		Actual	Planned	Actual	Planned
	Total sales	375	450	400	480
January	Direct sales	250	300	267	320
	Indirect sales	125	150	133	160
	Total sales	500	600	425	510
February	Direct sales	333	400	283	540
	Indirect sales	167	200	142	170
	Total sales	525	630	375	450
March	Direct sales	350	420	250	300
	Indirect sales	175	210	125	150

La capacidad de cambiar fácilmente las vistas de los mismos datos, mediante la reconfiguración de cómo se muestran las dimensiones, es uno de los grandes beneficios que proveen los sistemas multidimensionales, a la navegación de los usuarios finales. Esto se debe a la separación de la estructura de datos (MTS), de su visualización (grilla multidimensional).

Hay algunas reglas básicas que hay que tener en cuenta en el análisis de datos multidimensionales:

- Debe tratar de utilizar páginas. Esto ayuda a maximizar el grado en que todo en la pantalla es relevante.
- Cuando necesita anidar múltiples dimensiones a través de las filas y columnas, generalmente es mejor anidar más dimensiones a través de las columnas que a través de las filas, ya que suele haber mas espacio vertical en la pantalla, que horizontal.
- Antes de decidir cómo mostrar la información en la pantalla, pregúntese "¿Qué quiero mirar?", o "¿Qué estoy tratando de comparar?".

Vista clásica OLAP:

Store.Paris								
	Actual				Plan			
	Toys		Clothes		Toys		Clothes	
	<i>Sales</i>	<i>Costs</i>	<i>Sales</i>	<i>Costs</i>	<i>Sales</i>	<i>Costs</i>	<i>Sales</i>	<i>Costs</i>
Q1	320	200	825	750	525	603	750	629
Q2	225	220	390	250	554	600	365	400
Q3	700	600	425	630	653	725	720	530
Q4	880	850	875	700	893	875	890	889

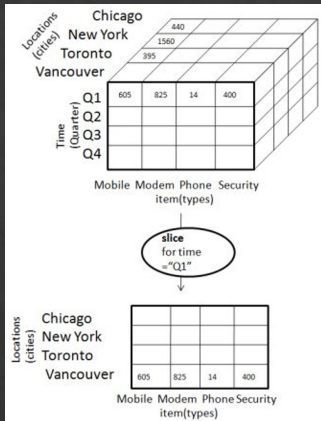


OPERACIONES

Slice:

En esta operación seleccionamos un valor particular de una de las dimensiones del cubo, buscando así quedarnos con una “rebanada” del cubo.

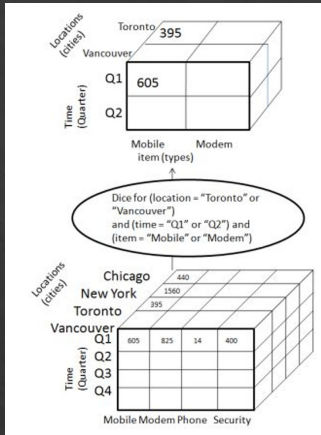
Slice:



Dice:

Seleccionar valores específicos para 2 o más dimensiones de las que se visualizan y con esto obtener un subcubo del cubo original.

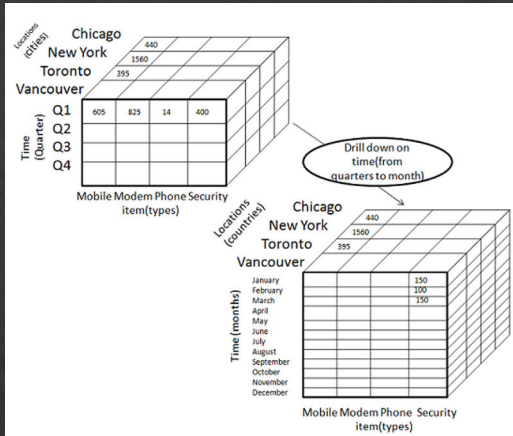
Dice:



Drill Down:

Esta operación lo que busca es ofrecer más detalle respecto de la información que actualmente se puede visualizar. Esto se puede lograr de dos formas, bajando un nivel en la jerarquía de una dimensión particular o agregando una dimensión para dar aún mas información sobre el contexto.

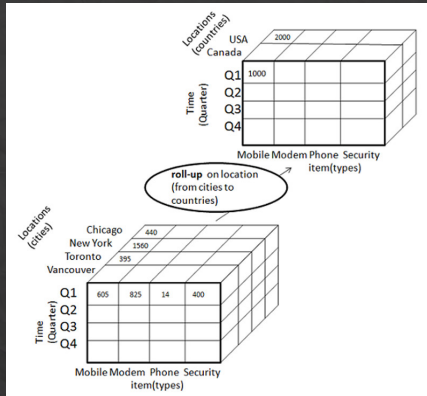
Drill Down:



Drill Up:

A diferencia de la anterior operación lo que queremos lograr es visualizar menos información, abstraernos un poco más, lo cual puede realizarse subiendo un nivel en la jerarquía de una dimensión o eliminando alguna de las dimensiones.

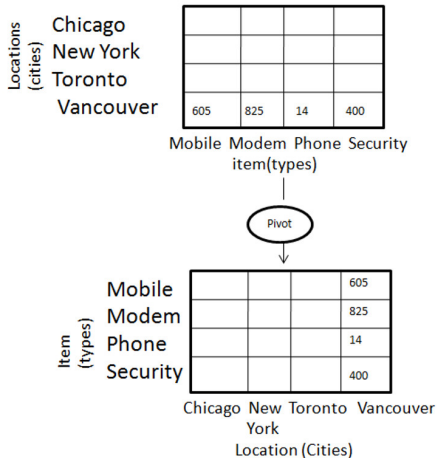
Drill Up:



Pivot:

Aplicando esta operación ofrecemos al usuario una presentación alternativa de la información rotando los ejes de datos.

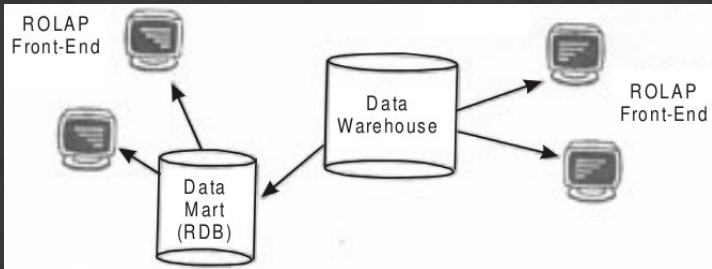
Pivot:



A large, solid black circle is centered on a background of crumpled, light gray paper. The circle contains the word "ROLAP" in white, uppercase, sans-serif font.

ROLAP

El **procesamiento analítico relacional en línea**, (relational online analytical processing, ROLAP) se trata de sistemas y herramientas OLAP contruidos sobre una base de datos relacional (RDB). Las ventajas de este modelo es que puede manejar una gran cantidad de datos y puede aprovechar todas las funcionalidades de la base de datos relacionales. Tiene como desventaja, un rendimiento lento, y además cada informe ROLAP es una consulta SQL, con lo cual está limitado por las funcionalidades de SQL.

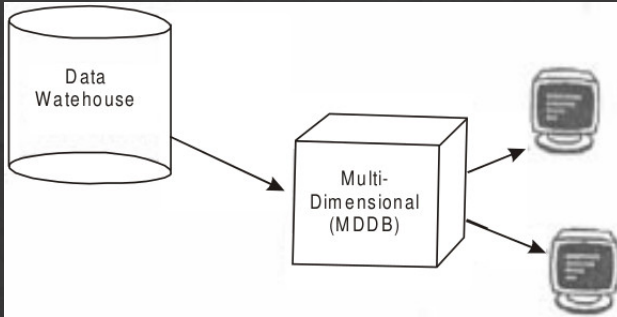


Bases de datos relacionales.

A large, solid black circle is centered on a background of crumpled, light gray paper. The circle contains the word "MOLAP" in white, uppercase, sans-serif font.

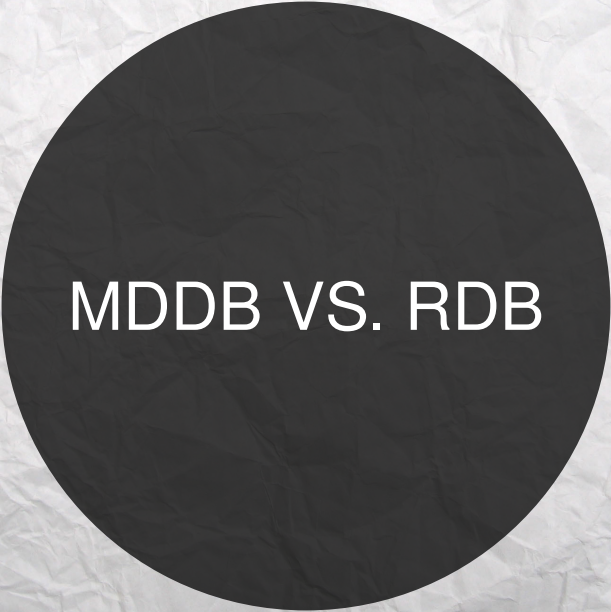
MOLAP

El **procesamiento analítico multidimensional en línea**, (multidimensional online analytical processing, MOLAP) consiste en herramientas que se ejecutan sobre una base de datos multidimensional (MDDDB). La ventaja principal de este modelo es que proporciona un gran rendimiento de consultas, dado que se pre-computan los datos en el cubo, cuando éste es creado. Sin embargo este modelo solo puede manejar una cantidad limitada de datos, ya que el cubo no se puede derivar de un gran volumen de datos.



Bases de datos multidimensionales.

ROLAP	MOLAP
Los cubos son generados dinámicamente al momento de ejecutar la consulta	Cada vez que se requiere realizar cambios sobre algún cubo, se debe recalcularlo totalmente, para que se reflejen las modificaciones llevadas a cabo
Los datos de los cubos se deben calcular cada vez que se ejecuta una consulta sobre ellos	Los datos de los cubos no son calculados en tiempo de ejecución



MDDB VS. RDB

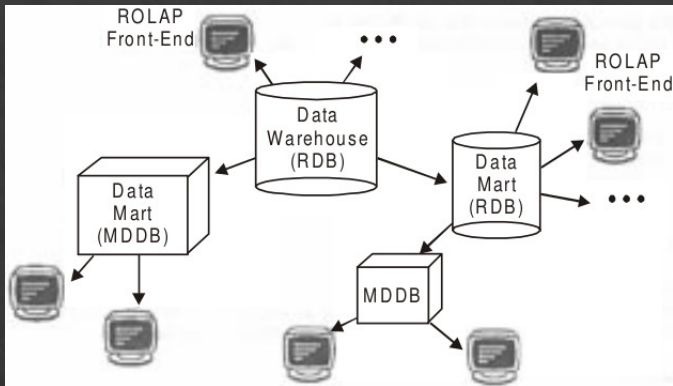
- **Bases de datos multidimensionales:** las MDDBs almacenan los datos en un "hipercubo", es decir, una matriz de almacenamiento multidimensional optimizada.
- **Bases de datos relacionales:** las RDBs guardan los datos como tablas, con filas y columnas que no mapean directamente con la vista multidimensional que tienen los usuarios de los datos.

- **Tamaño:** las MDDBs están, generalmente limitados por el tamaño de los datos, aunque el límite en el tamaño ha ido aumentando gradualmente a lo largo de los años.
- **Volatilidad de los datos de origen:** las RDBs tratan mejor la alta volatilidad de los datos. Los datos multidimensionales en hipercubos generalmente tardan mucho en cargarse y actualizarse.
- **Protección de la inversión:** La mayoría de las empresas ya han realizado importantes inversiones en tecnología relacional. El uso continuado de estas herramientas y habilidades para otro propósito proporciona un rendimiento adicional de la inversión.

A large, solid black circle is centered on a background of crumpled, light-colored paper. The word "HOLAP" is written in white, uppercase, sans-serif font in the center of the circle.

HOLAP

El **procesamiento analítico en línea híbrido** (Hybrid Online Analytical Process, HOLAP) es una combinación de ROLAP y MOLAP. HOLAP permite almacenar una parte de los datos como en un sistema MOLAP y el resto como en uno ROLAP.



Bases de datos relacionales, y multidimensionales.

A large, solid black circle is centered on a background of crumpled, light-colored paper. The text "CHARLA CONCLUIDA" is written in white, uppercase letters across the middle of the black circle.

CHARLA CONCLUIDA