

Guía de Estudio Análisis de la Información y la Decisión

Parte I (DW)

UP
**Universidad
de Palermo**

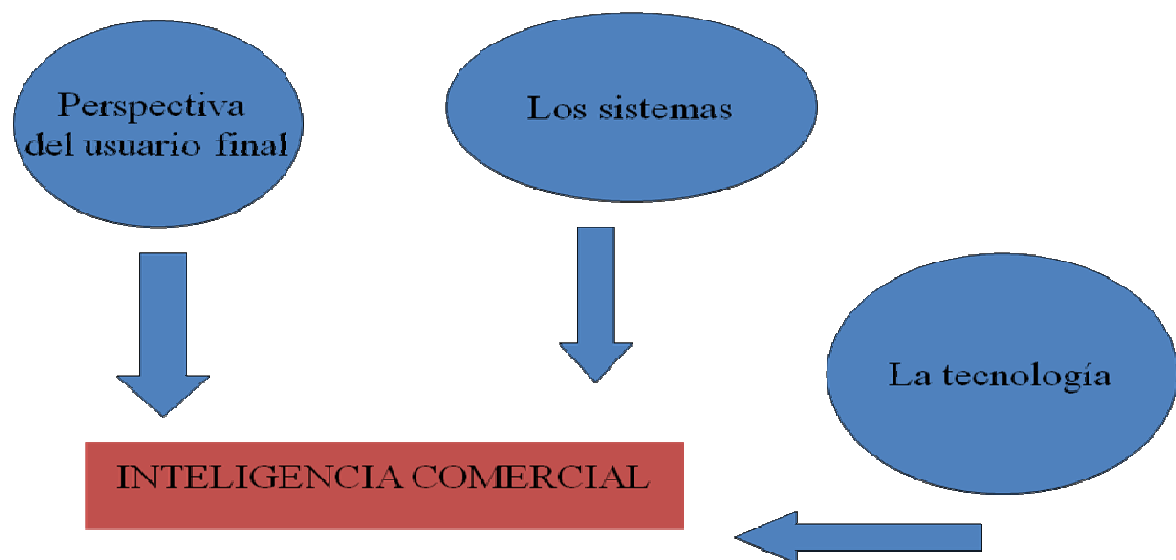
Esta guía contiene solamente un resumen de los contenidos de la materia, para mas información, acudir a la bibliografía obligatoria que figura en el syllabus.

Análisis de la Información y la Decisión - Parte I (DW)

La inteligencia comercial:

Es un conjunto de productos y servicios que permiten acceder a datos, analizarlos y convertirlos en información. Representa una iniciativa corporativa amplia que incluye DataWarehouse y DataMining. Involucra los usuarios, los sistemas y la tecnología.

Explota los datos a favor de la propia empresa. Analiza la información para tomar decisiones, apoya a la gerencia de la empresa. Permite entender las necesidades de los clientes.



Las demandas del usuario final:

Situación:

- Competitividad creciente
- Valoración de la atención al cliente
- Organizaciones más delgadas delegan más responsabilidad en el usuario final

NECESIDAD DE CONTAR CON INFORMACIÓN INTEGRADA

Problemas:

- Inexistencia de datos corporativos
- Islas (información de un mismo cliente separada en sistemas de distintas áreas)
- Falta de datos históricos

LOS SSD (Sistemas para el soporte de decisiones): HISTORIA:

Ciclos manuales:

- Recepción demorada del informe
- Pérdida de performance de los sistemas operacionales

Extracción y vuelco de datos operacionales en las PC's:

- Existencia de una maraña de *Programas de extracción*
- *Falta de Credibilidad: información de distintos momentos, con distintos algoritmos....*
- *Falta de productividad: comprender distintos sistemas, soportar sus cambios*
- *Falta de flexibilidad: para ver los datos de distintas maneras, para responder a distintas preguntas*

Nuevo enfoque de los DSS:

- Consolidar los datos en un nuevo entorno, integrado, con perspectiva histórica y con facilidad de manipulación (e-data = datos reunidos y sincronizados electrónicamente)
- La solución se analiza desde dos enfoques:
 - a. *Desde la Tecnología y su evolución.*
 - b. *Desde los tipos de sistemas existentes.*

Desde la tecnología: Autonomía versus Centralización.

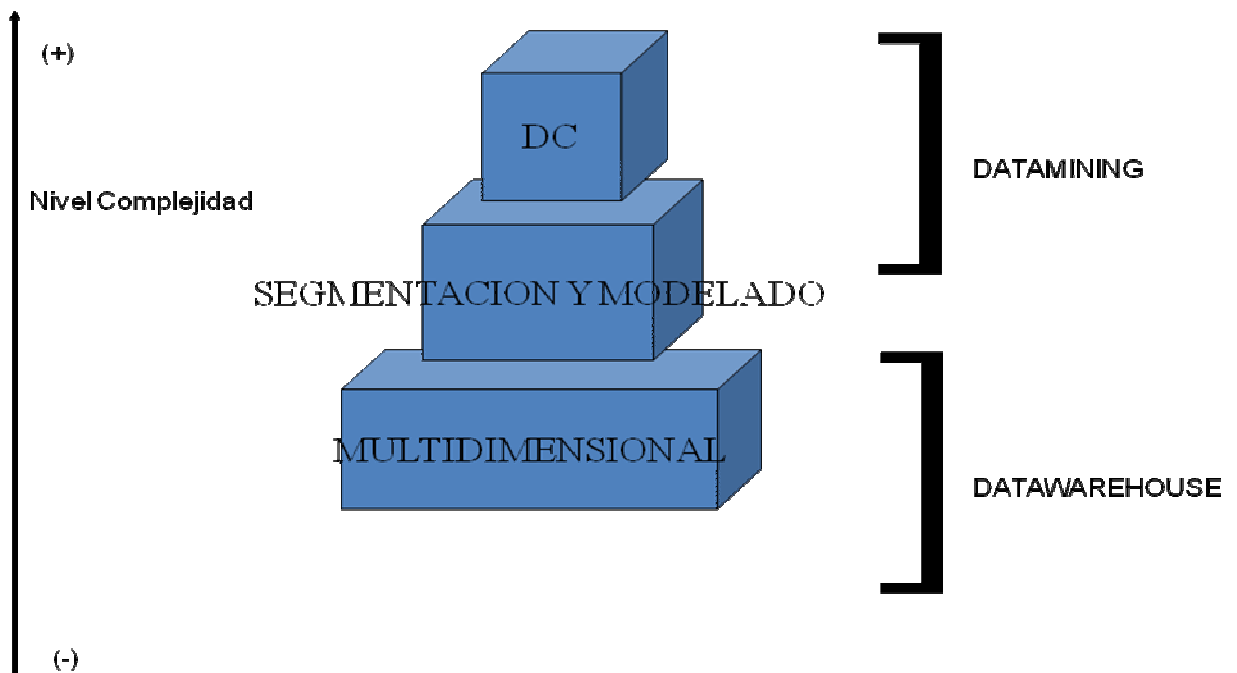
- 60: Master files, programas Cobol y reportes: redundancia y desincronización, mucho hard y programas complejos
- 70: Mainframe y papel ; Direct Access Storage Device (DASD) y Bases de Datos
- 75: sistemas OLTP, con alta performance (sin escalabilidad ni GUI)
- 80: Aparece la PC y las redes de PC (file server o print server pero no aplicaciones compartidas); programas extract
- 90: Cliente-servidor, aplicaciones centralizadas y para acceder a bases heterogéneas: ODBC o similares
- Cómo satisfacer a distintos usuarios, acostumbrados a distintos front-ends y con distintas necesidades de información?
- Cómo interrogar a la base?
 - *SQL*
 - *Vistas: Se deben construir permanentemente nuevas*
 - *Evitar SQL con herramientas gráficas*

- Necesidad de nuevas herramientas con workstation heterogéneas y usuarios autónomos
 - Que trabajen en términos de negocios
 - Con polimorfismo
 - Con facilidades para que se pueda administrar a los usuarios (crear repositorio, desalentar los queries que harían caer al sistema, controlar accesos de usuarios a queries y reportes)
 - Que sean tan expresivas como SQL y generen SQL
 - Con inteligencia para usar el poder de los Server

Desde los sistemas:

- Las actividades:
 - Operacionales (comprar, vender, producir, transportar)
 - De toma de decisión (presupuestar, evaluar, planificar)
- Los sistemas:
 - Operacionales (OLTP)
 - De análisis orientado a la toma de decisión (OLAP)

Pirámide de Explotación y Análisis de la Información



Tipos de Consultas:

Consultas estándar:

- Listar préstamos hechos al cliente X y fechas de pago en que se demoró
- Mostrar todos los clientes que compraron el producto X el año pasado
- Listar clientes para los que el consumo en horas pico disminuyó un 20%

Análisis multidimensional:

- Mostrar ingresos trimestrales por ventas a grandes clientes por zona (drill) en los años 2003 y 2004 (Slice)

Modelado:

- Valor de vida del cliente
- Desgaste del cliente
- Impacto del mal tiempo en las ventas

Segmentación:

- Clientes que responden a descuentos
- Clientes que no responden a promociones

DC (Descubrimiento de Conocimiento):

- Análisis de afinidad

DATAWAREHOUSE:

- Repositorio de datos históricos referidos a un tema en particular
- Colección de "data marts" más pequeños.
- Plataforma de hardware, software y datos separada – *mainframe* o *PC*- que permita a un hombre de negocios tomar decisiones.
- Colección de datos derivados (*Según Bill Inmon*)
 - Orientados a un tema (<> A una transacción)
 - Integrados
 - Variables en el tiempo
 - No volátiles

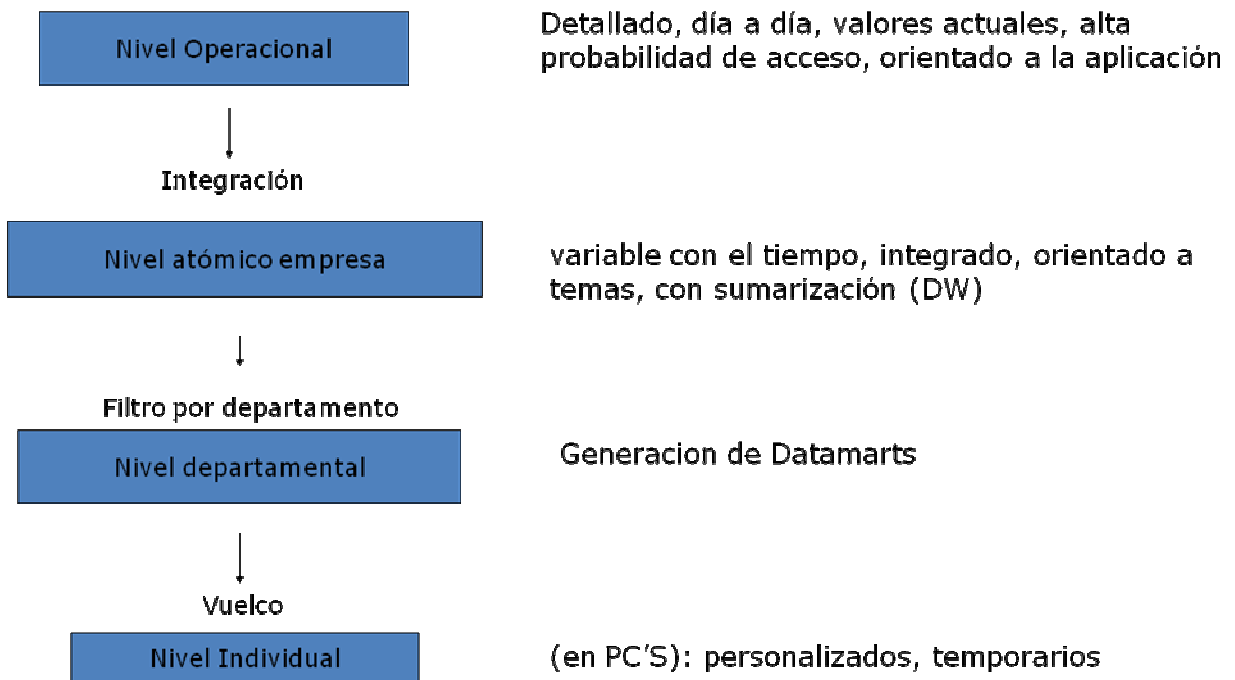
características:

- En general, una plataforma de hardware aislada;
- Integra datos de diferentes fuentes u orígenes (Sistemas OLTP, archivos planos, información externa, etc.)
- Sus datos se usan para la toma de decisiones;
- Los datawarehouses desnormalizan información;
- Es una combinación de hardware, software especializado y datos.

Por qué?

- Aliviar la carga de los servidores
- Acabar con datos sucios
- Seguridad en el acceso a los datos corporativos + democratización del acceso a los datos
- Una única verdad
- Mejor relación con el cliente

Niveles de la información:



Sistemas OLTP (On-Line Transactional Processing):

- Las actividades incluyen comprar, vender, producir, transportar.
- Son puramente operacionales. Preguntas: dirección del cliente, saldo, etc.
- Orientado a responder a eventos.
- Verifica consistencia de datos.
- Usa normalización.

Sistemas OLAP (On-Line Analytical Processing):

Sistemas especialmente diseñados para el análisis de la información en apoyo a la toma de decisiones.

- Las actividades incluyen presupuestar, evaluar, planificar. Son orientados a la toma de decisiones. Preguntas: Que quieren los clientes? Cual es el producto más rentable? Como cambio lo que quieren los clientes en los últimos 5 años?
- Refleja lo que no hay en la normalización.
- No necesita consistencia
- No preocupa el espacio
- No focaliza en cada evento
- La unidad es la consulta
- Se apoya en información histórica y proyectada
- Utiliza hechos, medidas y dimensiones con las cuales crea el modelo estrella.
- Agregan valor a los datos por su organización y agregación

Refleja como es visto todo el proceso

Se ve en términos de Hechos o Medidas, Parámetros o Dimensiones.

Requerimientos de un Sistema OLAP:

- Rápido, flexible y con acceso a grandes volúmenes de datos;
- Rápido acceso a datos y rápidos cálculos
- Fuertes capacidades analíticas (formulaciones estadísticas complejas)
- Interfaces amigables
- Vistas flexibles: para realizar cálculos impensados y ofrecer modalidades de exposición (gráficos, tablas, etc.)
- Soporte a múltiples usuarios: la cantidad de usuarios crece día a día.

OLAP versus DW

Surgieron en forma independiente.

- OLAP: hace énfasis en proceso de satisfacción al usuario final y de explotación de la información
- DW: hace hincapié en la obtención y almacenamiento de los datos; proceso para obtener datos seguros, consistentes, integrados y disponibles

Solución robusta: Utilización de sistemas OLAP explotando un DW.

Operacional vs. Soporte de decisiones:

- OLTP: que dirección tiene el cliente? Tiene saldo? Qué productos de una orden no fueron entregados?
- OLAP: Qué quieren los clientes? Cómo cambió eso en los últimos 5 años? Cuál es el producto más rentable en cada región?

Desventajas de Hojas de Cálculo y SQL:

- Las hojas de cálculo guardan la información como la veo.
- Las hojas de cálculo si bien permiten realizar cálculos condicionales, permiten realizar libros multiniveles que no separan la estructura de las vistas.
- En la esencia de SQL no esta previsto el análisis, para algunos análisis se necesita crear muchas tablas.

Tabla comparativa de Sistema Operacional vs Soporte Toma de Decisiones

	OPERACIONAL	SOPORTE DECISIONES
<i>Usuario</i>	Empleado	Profesional
<i>Uso</i>	Ejecución	Análisis
<i>Soporte</i>	Día a día	Estrategias
<i>Interacción</i>	Predeterminada, repetitiva	Ad hoc, no estructurada
<i>Tiempo Respuesta</i>	> 2",3"	De seg. a minutos
<i>Pantallas</i>	Fijas	Variables
<i>Unidad</i>	Transacción	Consulta
<i>Características</i>	Read /Write	Read
<i>Foco</i>	Ingreso de Datos	Información
<i>Acceso a Datos</i>	Decenas	Millones
<i>Valores</i>	Corrientes	Históricos y proyectados
<i>Naturaleza</i>	Dinámico	Estático hasta Refresh
<i>Organización</i>	Por aplicación seguros de vida, salud, auto....	Por tema siniestros, pólizas, clientes
<i>Estructura</i>	Normalizada	Desnormalizada
<i>Granularidad</i>	Detallada	Con cierto nivel de sumariación

El Modelo Multidimensional

Modelo Multidimensional:

- Preparado para facilitar la definición y el manejo de datos sumariados y análisis a múltiples niveles.
- Se puede definir un conjunto de datos en términos de múltiples dimensiones (un hipercubo)
- Las dimensiones pueden ser jerárquicas y permiten distintos niveles de agregación
- Se puede mostrar tridimensionalmente (páginas, filas, columnas), combinando dimensiones en filas y columnas

Metodología Multidimensional:

Aparece para responder a nuevos objetivos:

- Debe seguir los requerimientos del análisis del negocio.
- Tiene que ser para el usuario final fácil y obvia.
- Tiene que ser flexible a los cambios del negocio.
- Debe describir exactamente el pasado.
- Debe reunir datos sumariados y tambien de bajo nivel, soportando una implementación incremental (crecer).
- No se debe buscar solamente una herramienta de sumariación, más flexible que un reporte manual.
- Para diseñar no vale el DER.
- Es fundamental comprender el negocio para la selección correcta de las dimensiones.
- Es muy importante disponer de herramientas software adecuadas para explotar el modelo.

Vinculo entre DW, MM y OLAP:

Lo ideal es diseñar especialmente un Datawarehouse para explotar un Modelo Multidimensional a través de una herramienta OLAP.

Componentes de una solución OLAP:

Alcance del concepto

DW + MM + Herramientas OLAP

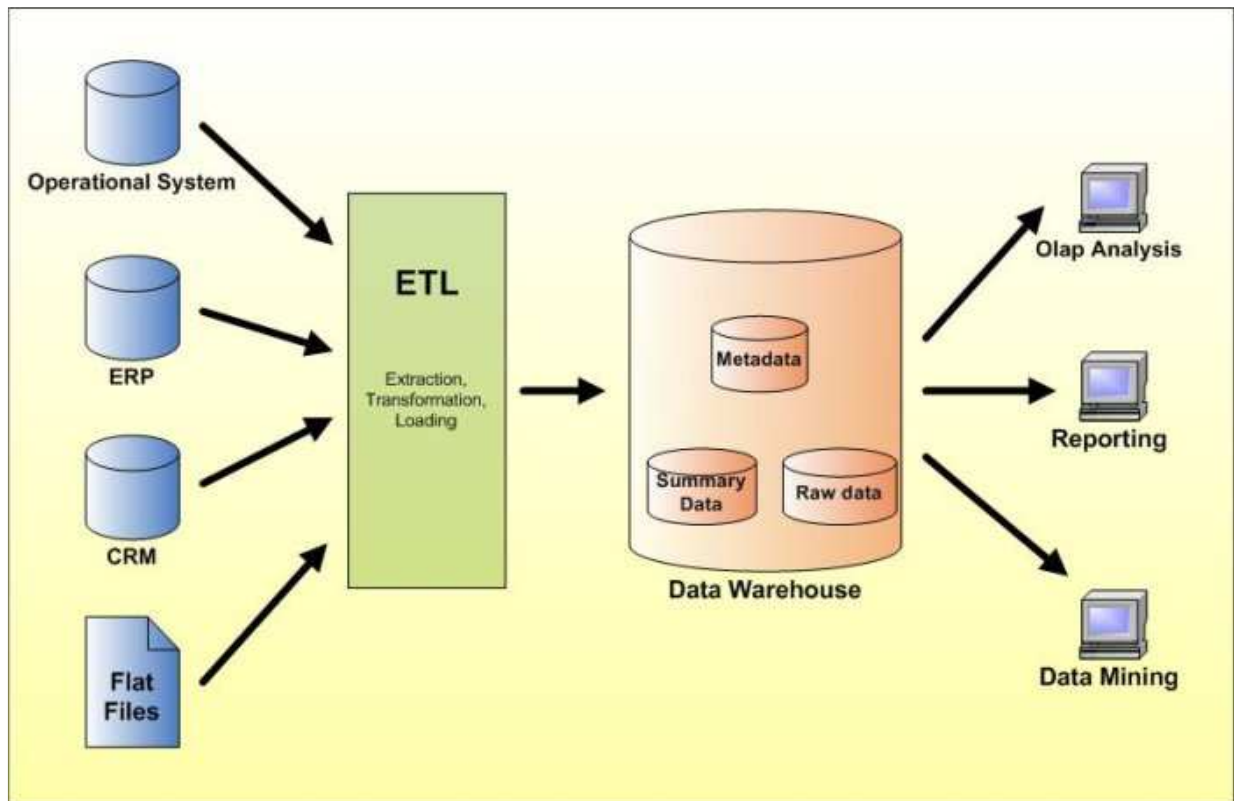
SOLUCIÓN OLAP

Componentes:

AGRUPADOR	COMPONENTE	FUNCIONES/CARACTERISTICAS	PRODUCTO
DW	ETL	Extraccion, Transformacion y Carga de datos.	Oracle Warehouse Builder, DataStage.
	DBMS	<ul style="list-style-type: none"> • Escalable • Procesamiento paralelo de consultas; • Carga e indexación de datos • Tablas particionadas • Adm. Seguridad, resguardo, performance, registro de uso, etc. • Creación y mantenimiento de Metada de BD 	Oracle, SQL Server, DB2, MySQL, etc.

AGRUPADOR	COMPONENTE	FUNCIONES/CARACTERISTICAS	PRODUCTO
OLAP	Aplicaciones Frontend	<ul style="list-style-type: none"> • Petición de datos • Inspeccionar datos • Formatear datos • Generación de reportes • Explotación de cubos (drill, slice) 	Oracle Discoverer, MS Analysis Services, Microstrateg y Olap Services, Qlikview, Penthao, etc.
	Metada de aplicación	<ul style="list-style-type: none"> • Almacenamiento de la estructura del Modelo Multidimensional: Dimensiones, Jerarquías, Medidas. 	Oracle Discoverer, MS Analysis Services, Microstrateg y Olap Services, Qlikview, Penthao, etc.

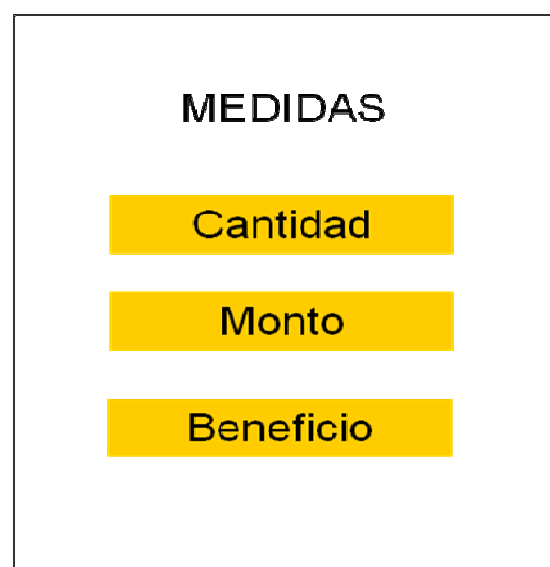
Esquema de una solución BI:



El modelo multidimensional:

- Permite ver los datos desde múltiples perspectivas
- Muestra medidas, dimensiones y sus interrelaciones
- Utiliza el vocabulario del usuario

Ejemplo: MM de ventas



Elementos:

- Medidas:
 - Información cuantitativa
 - Representan el cuánto de una consulta
 - Primitivas o calculadas
 - Se almacenan en la FACT TABLE

- Dimensiones:
 - Calificadores que dan sentido a las medidas
 - Son el que, quién, cuándo, dónde de una consulta
 - Se almacenan en las TABLA DE DIMENSIONES, junto con sus atributos
 - Se guardan como códigos numéricos o pocos caracteres
 - Pueden tener jerarquías, que son distintos niveles de sumarización
 - Pueden tener múltiples jerarquías
 - Día → Semana → Año
 - Día → Mes → Año

- Fact table (Tabla de hechos):
 - Es el centro del modelo dimensional
 - Tiene punteros a las claves de menor nivel de cada dimensión
 - Contiene las medidas.

Dimensiones: atributos

- Son campos que amplían información sobre cada elemento de la Dimensión
- Facilitan las consultas a los usuarios finales
- EJ: Producto (Id, Descripción, Línea, rama)

Dimensiones: tabla producto

ID	DESC_PROD	TIPO	DESC_TIPO	DESTINO	DESC_DESTINO
1	TV	101	Electrónica	3	Hogareña
2	Radio	101	Electrónica	3	Hogareña
3	Heladera	102	Blanca	3	Hogareña

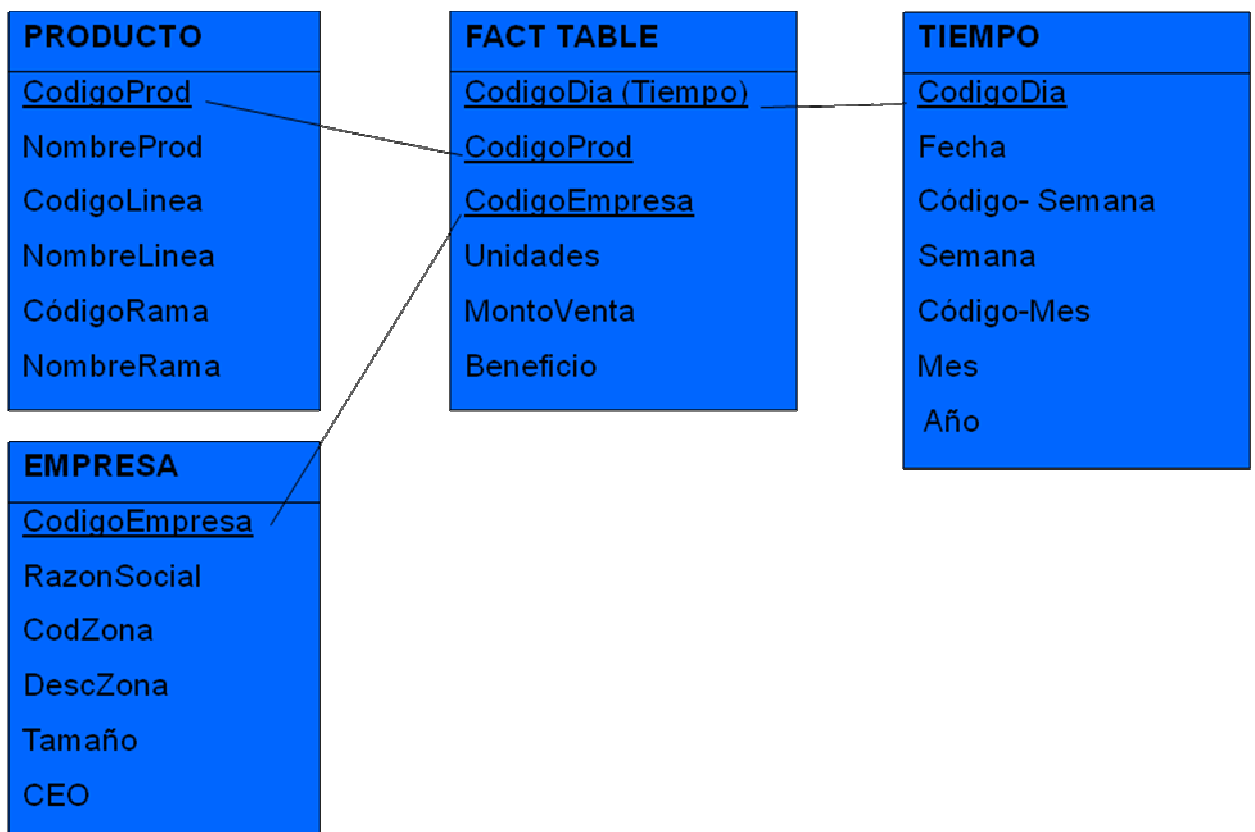
Fact table: ejemplo

Empresa	Producto	Tiempo	Unidades	Monto Ventas	Beneficio
100	1	1340	30	800.76	270.12
200	1	1341	20	740.7	260.5
100	2	1341	45	115	43.21

El esquema estrella:

- Es un tipo de diseño especial para los procesos analíticos
- Cada tabla de dimensiones se vincula con la Fact Table siempre por el mismo campo (ID correspondiente)
- Es simple e intuitivo
- De mantenimiento flexible

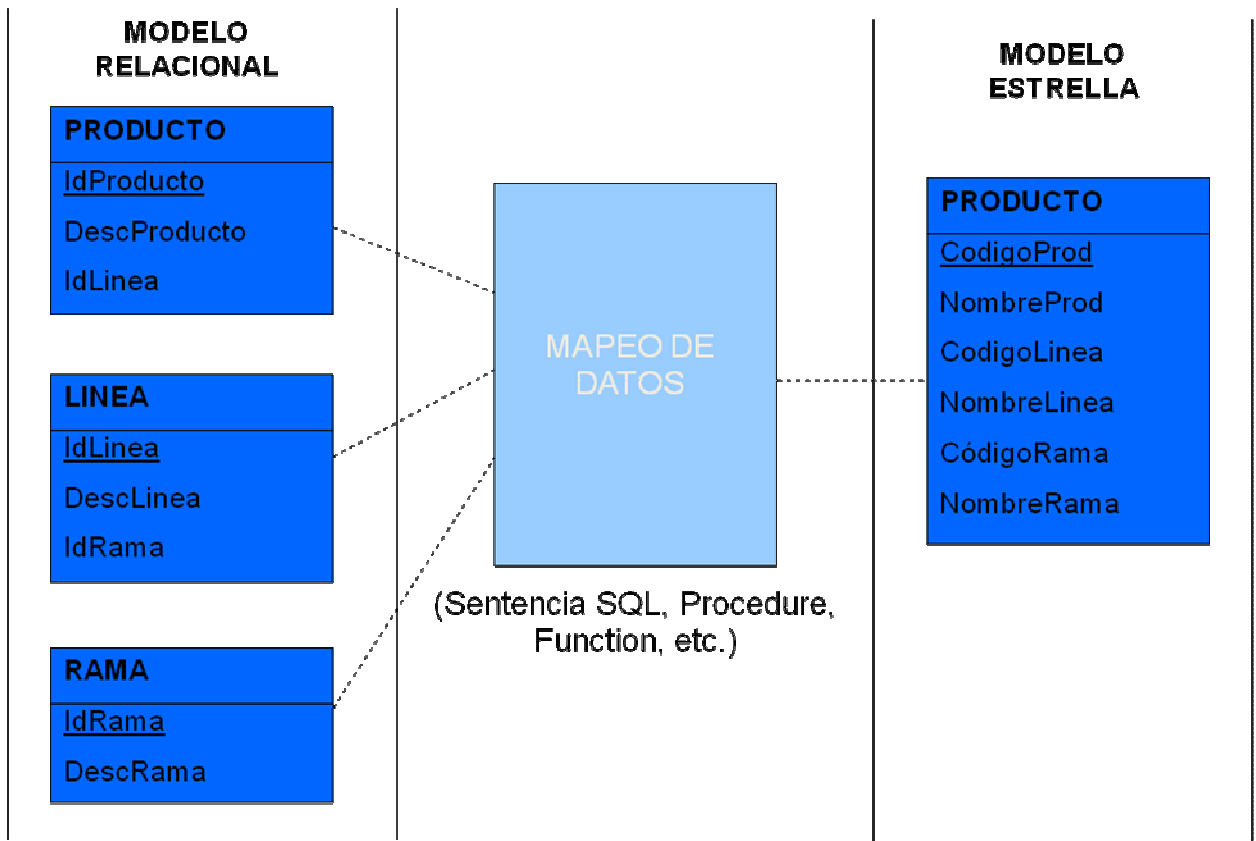
Esquema Estrella Simple: ejemplo



Creación del modelo multidimensional:

- Identificar Medidas
- Identificar Dimensiones y Jerarquías
- Determinar la granularidad
- Construir el Modelo Estrella
- Verificar el modelo con los usuarios y refinarlo
- Población del Modelo (Mediante el proceso de Mapeo)

Mapeo:



Modelo multidimensional conceptos avanzados

El hipercubo

- Mas de tres dimensiones, supero las aristas del cubo (Región, mes, productos, cliente)
- Esta compuesto por una serie de ejes
- No tiene límites de dimensiones
- Cualquier combinación es valida
- La medida la trabaja como una dimensión más
- Los miembros de las dimensiones son las distintas medidas
- No genera el problema de la cantidad de aristas del cubo.
- Las múltiples dimensiones y múltiples niveles por dimensión es el eje del hipercubo.
- Una celda es la intersección de un miembro de cada dimensión
- El OLAP de DB2 trabaja con esto.

Representación de Hipercubos

- Representación de una dimensión con una medida, sin problemas en la pantalla
- Representación de una dimensión con varias medidas, sin problemas en la pantalla
- Representación de dos dimensiones (tiempo y producto) con varias medidas, en la pantalla plana
- Representación de tres dimensiones (tiempo y producto y depósito) con varias medidas, en la pantalla plana

Otra forma:

- Cada dimensión lógica es representada por una línea vertical, segmentada en tantas partes como elementos tiene la dimensión
- Otra línea vertical representa las medidas
- Es posible desplazarse en forma independiente por cada línea y encontrar distintas intersecciones de información válida.

Densidad del Cubo

- El cubo es denso cuando las dimensiones tienen todos sus miembros. Solo se da en muy pocos casos. Sino se forman agujeritos de información. Esto sucede en las combinaciones de dimensiones donde no hay medidas asociadas.

Las Dimensiones

- Calificadores que dan sentido a las medidas de la FACT-TABLE
- Son jerárquicas y permiten distintos niveles de agregación.
- Son las medidas que el usuario quiere saber
- Son los criterios por lo que muestro la información
- No incluir las combinaciones de dimensiones donde no hay medidas asociadas. Produce cubos no densos (con agujeritos de información)
- Puede haber más de una jerarquía en la misma dimensión.
- Distintos elementos pueden tener distintas jerarquías
- Puede haber hijos con más de un padre
- Se cuenta con formulas para asociar las dimensiones.
- Para descubrirlas preguntar: Quién, cuándo y dónde?
- Las tablas de las dimensiones son más cortas y anchas que la FACT-TABLE
- Los atributos pertenecen a las dimensiones no a la FACT-TABLE
 - EJ: Tiempo es dimensión jerárquica
Día->Mes->Trimestre->año
- Sexo es dimensión no jerárquica
- Puede contener medidas en el caso de usar hipercubos. DB2.

Agregar una Dimensión

- Se puede agregar. EJ: vendedor. Implica modificación de la FACT-TABLE para agregar el vínculo. Si no califica las medidas de la FACT-TABLE no puedo agregarla a menos que modifique la FACT-TABLE previamente.

La agregación

- La agregación es el proceso mediante el cual la información de bajo nivel se resume anticipadamente y se coloca en tablas especiales que almacenan la información resumida, o Agregada. La técnica de agregación es parte integral de la solución al problema de performance.

Tablas de las Dimensiones

- Tienen a ser mas anchas que largas
- Tienen claves que no provienen de la fuente de datos
- Usan claves numéricas simples
- Están desnormalizadas
- Si falta un dato EJ Cliente debo poner desconocido
- Si es muy ancha la rompo en dos (mini-dimensiones)
- Posee tuplas para representar el elemento nulo
- Suele usar mas de una tabla para el mismo criterio (si tiene muchos atributos / atributos que se usan juntos / atributos poco frecuentes)

Dimensión Degenerada

- No la soportan todos los productos
- Es tener una dimensión sin tabla
- EJ: guardar el número de factura. Tener tantos registros como la FACT-TABLE.

Dimensiones Jerárquicas

- Reflejan la realidad
- La mayoría de las dimensiones lo son dado que la realidad es jerárquica
- Facilita reportes con distinto nivel de detalle
- Permite subir y bajar de nivel a los usuarios
- Permite distintos niveles de agregación
- Puede ser vista como un árbol donde los miembros de menor nivel son las hojas
- Los grupos jerárquicos deben ser flexibles

Mini-Dimensiones

- Abrir una dimensión en dos dimensiones más pequeñas
- La divido por los datos que mas uso en una y los que menos uso en otra

Miembros de las dimensiones

- Los distintos clientes
- Las distintas fechas
- Los miembros de las dimensiones no se suman

Los Atributos

- Facilitan las consultas a los usuarios finales.
- Describen los códigos. Nombre, tamaños, jefe, etc.
- Se pueden cambiar el valor agregando desconocido o valor por defecto o calculando si es posible para los faltantes.

Atributos Variables

- Son los que pueden cambiar con el tiempo. EJ: cambio de % IVA. Pueden ser convertidos a una nueva dimensión para resolver problemas entre datos nuevos y anteriores en el DataWarehouse.

Las Medidas

- Son lo que quiero mostrar
- Información cuantitativa
- Representan el cuanto
- Pueden ser primitivas o calculadas
- Se almacenan en la FACT-TABLE
- Cantidad de alumnos
- Cantidad de pesos

Agregar una medida

- Dependiendo del tipo si es derivada (fácil calculo y listo) o no.
- Existen múltiples soluciones, la idea es encontrar la óptima.

Tipos de Medidas

- Aditivas
 - Son las sumables. EJ: Unidades, peso.
- Semi-Aditivas
 - Stock. Se suma en algunas dimensiones y no en otras.
 - Al Stock no lo puedo sumar sobre tiempo. Solo puedo promediario sobre tiempo.
- No-Aditivas
 - Porcentaje. No tiene sentido sumarla.

FACT-TABLE

- Es el centro del modelo multidimensional
- Para tener distintas granularidades necesito distintas FACT-TABLE
- Medidas y punteros a los miembros de las dimensiones.
- No incluye las combinaciones de dimensiones para las cual no hay medidas
- No tiene referencias nulas a ninguna dimensión
- Por lo general tiene muchos registros
- Los atributos pertenecen a las dimensiones no a la FACT-TABLE
- Lleva un identificador (código) por cada dimensión
- Solo va día (o el menor nivel guardado en la tabla tiempo)
- Van todas las medidas
- Tiende a ser mas larga que ancha. (Inversa a tabla de dimensiones)
- Puede tener múltiples vinculaciones a la misma dimensión. EJ: tiempo, en la cual apunta a dos fechas distintas. EJ: código de día cuando se pidió y cuando se entrego.

Tiene sentido un FACT-TABLE sin medida?

- Si, solo si quiero saber que el hecho se produjo. (Es un caso atípico). EJ: institución medica y el registro de enfermedades. Marcan un evento y vinculación de las dimensiones.

Tiene sentido varias vinculaciones a una FACT-TABLE?

- EJ: fecha de orden y orden de entrega

Tiene sentido mas de una FACT-TABLE?

- Si, cuando las dimensiones no son comunes o cuando encuentro medidas que están calificadas por dimensiones distintas. Cuando están asociadas a distintas dimensiones. Se pueden vincular varias FACT-TABLE a las mismas dimensiones en una pregunta.

Jerarquía

- Varios niveles de detalle. EJ: día -> mes -> año
- Son distintos niveles de sumarizacion
- Están contenidas en las dimensiones

Granularidad

- Es el nivel elemental (Jerarquía mínima) EJ: tiempo es día
- Una vez definida no se puede minimizar
- Representa el mínimo nivel de detalla de la FACT-TABLE
- Para tener distintas granularidades necesito distintas FACT-TABLE

Que debo tener en cuenta después de la charla con el cliente para realizar un modelo multidimensional?

- Las Dimensiones y sus atributos
- Las jerarquías de las dimensiones
- La granularidad de las dimensiones

Representación del Modelo Multidimensional

- Cuando tengo mas de 2 dimensiones se complica la visualización por eso una solución es embeber dimensiones en columnas o filas pero los títulos se hacen más grandes.
- Representa datos en 3 dimensiones (columna, fila y pagina) EJ: fila: producto; pagina: día y columna: locales.
- No es la mejor para rebanar el cubo
- Usa tres aristas del cubo. Una dimensión en cada arista y la medida en la intersección.

Diseño incorrecto:

				Q1	Q2	Q3
Actual	Paris	Toys	Sales	320	225	700
			Costs	200	220	600
		Clothes	Sales	825	390	425
			Costs	750	250	630
	NYC	Toys	Sales	500	310	880
			Costs	450	500	850
		Clothes	Sales	210	625	875
			Costs	225	600	700
Plan	Paris	Toys	Sales	525	554	653
			Costs	603	600	725
		Clothes	Sales	750	365	320
			Costs	629	400	530
	NYC	Toys	Sales	460	520	810
			Costs	325	610	875
		Clothes	Sales	655	725	890
			Costs	780	650	889

Diseño correcto:

Store.Paris

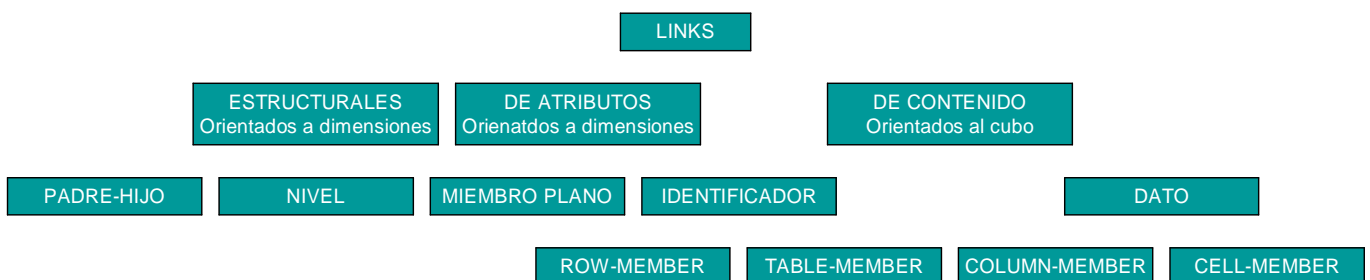
	Actual				Plan			
	Toys		Clothes		Toys		Clothes	
	Sales	Costs	Sales	Costs	Sales	Costs	Sales	Costs
Q1	320	200	825	750	525	603	750	629
Q2	225	220	390	250	554	600	365	400
Q3	700	600	425	630	653	725	720	530
Q4	880	850	875	700	893	875	890	889

Los datos:

- Si bien los datos son generalmente numéricos, podrían ser textuales, gráficos, sonidos, etc.
- Las herramientas OLAP agregan valor a los datos por organización y agregación
- Los datos numéricos se prestan para ambas situaciones.
- Los atributos son generalmente alfabéticos
 - Los hay que definen al miembro
 - Los hay que marcan estados del miembro
 - Ej: dirección de un cliente
 - Podrían considerarse como datos que varían sobre una sola dimensión
 - Se desaconseja su uso como atributos

LINKS

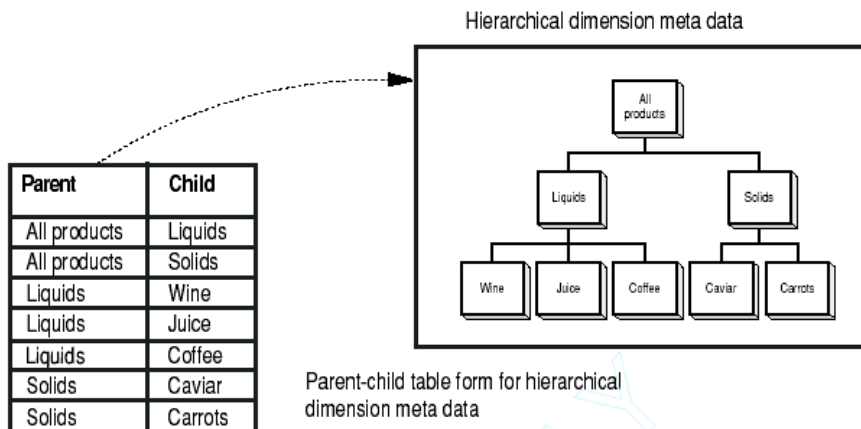
- Son modelos para persistencia de datos y estructura de un cubo OLAP
- Son el vínculo entre el Modelo Multidimensional de una herramienta OLAP y sus orígenes de datos externos.
- Tienen mucha relevancia cuando una herramienta OLAP no se provee de datos desde un DW bien diseñado.
- Permiten hacer el macheo de la FACT-TABLE con las dimensiones
- Permiten cargar el cubo
- Permiten definir los valores de las aristas del cubo (dimensiones)
- Una vez definidas las aristas del cubo cargo los atributos.



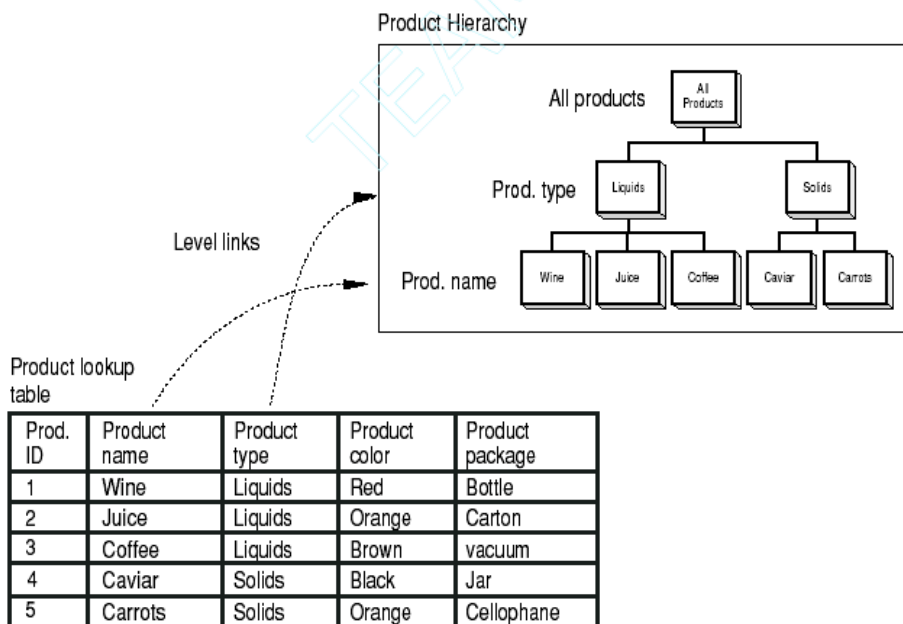
Tipos de LINKS

Estructurales

- No son valores
 - Definen la estructura
 - Permiten descubrir las dimensiones
 - Permiten descubrir las jerarquías de las dimensiones
 - Uso el concepto de link estructural para cargar las aristas del cubo
 - No es físico, solo es la técnica por la cual defino la estructura de la dimensión
 - Son los códigos de las tablas EJ: CodigoArticulo, CodigoLinea
- Se puede usar tablas de dos columnas (padre-hijo), que carecen de semántica de los niveles

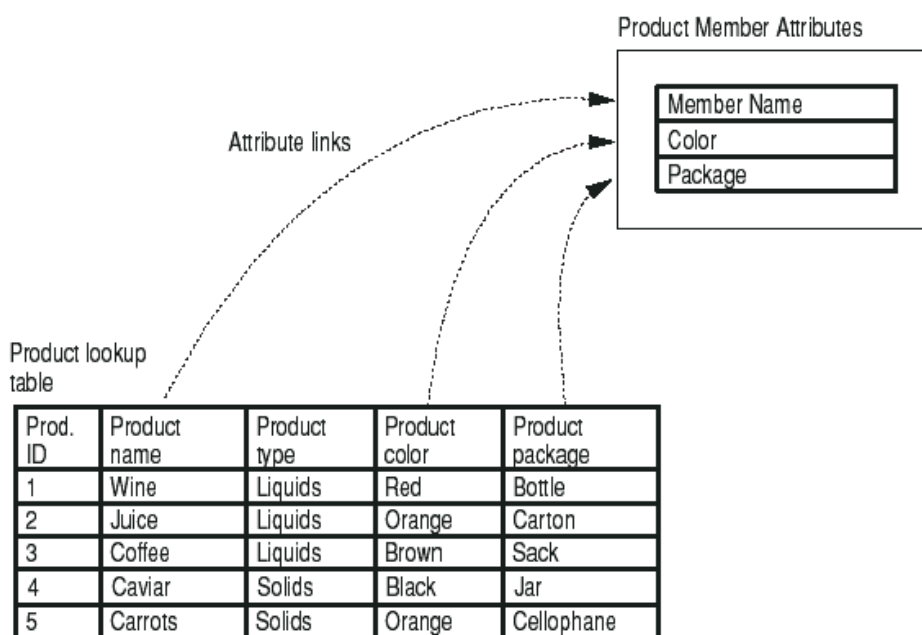


- Tablas completas, con tantas columnas como distintos niveles existan: T=(Producto, Nombre, Tipo, Línea, Color, envase)



De Atributos

- No son valores
- Lo que el usuario ve como elementos o miembros de las dimensiones
- Mapea los atributos a los elementos de las dimensiones
- El cubo esta vacío al momento de su definición
- Tengo las aristas del cubo
- Es una columna descripción que depende de un código.



De Contenido

- Cuando cargo los datos
- Los link de contenido junto con los link estructurales cargan el cubo
- Alimentan realmente al cubo
- Suponen sumarización previa cuando cargo desde sistemas transaccionales

Del Tipo RowMember

- No es un Valor
- Tiene titulo
- Identifican un Column Member
- Vinculo entre una columna de una tabla y una dimensión
- Los distintos valores son los miembros de la dimensión

Tiempo	Producto	Ventas
2-1-2002	Zapato	10
2-1-2002	Bota	4
3-1-2002	Zueco	5

- Vinculo entre una columna y medidas
- Los distintos valores son los nombres de las medidas

Tiempo	Producto	Medidas	Valores
2-1-2002	Zapato	venta	10
2-1-2002	Bota	venta	4
2-1-2002	Zapato	costo	5

Del Tipo ColumnMember

- Es un Valor
- Tiene titulo
- No necesita toda la fila para identificarse
- En el caso de ser dos columnas forman la dimensión Escenario (Proyectada)
- La dimensión Escenario no necesita toda la fila para identificarse
- Los distintos valores serán los distintos valores de esa medida

Tiempo	Producto	Venta
2-1-2002	Zapato	10
2-1-2002	Bota	4
2-1-2002	Zapato	5

- Vinculo en una columna y un miembro de una Dimensión

Tiempo	Producto	Medidas	Actual	Plan
2-1-2002	Zapato	Venta	10	12
2-1-2002	Bota	Venta	4	4
2-1-2002	Zapato	Costo	5	6

Del Tipo Cell

- Es un valor
- No tiene titulo, el nombre de columna no se asocia con nada.
- Necesita ser calificado por toda la fila
- Las medidas de los distintos valores son indicados por otra columna.
- La columna es un mix y depende de otras columnas.
- Forman parte del centro del cubo.
- Esta formado por distintos valores.

Tiempo	Producto	Medidas	Valores
2-1-2002	Zapato	Venta	10
2-1-2002	Bota	Venta	4
2-1-2002	Zapato	Costo	5

Del Tipo Table-Member

- Toda la tabla se asocia con un miembro de la dimensión
- Para los demás miembros de la dimensión no hay valores en la tabla

Tiempo	Producto	Medidas	Valores
2-1-2002	Zapato	Venta	10
2-1-2002	Bota	Venta	4
2-1-2002	Zapato	Costo	5

Formulas (rangos)

- Las elementales son las sumas
- Se definen sobre los ejes (No sobre celdas)
- Las formulas se derivan desde las hojas a la raíz
- Debo definir el orden de aplicación, dado que puede generar diferencias de resultados
- Debo definir los ejes sobre los cuales se aplica
- La formula de un total agregado se apoya en una estructura jerárquica.

Formulas Condicionadas

Dependen lo que pidan sumo o saco el promedio

Ventas Totales:

	Local 1	Local 2	Tot. Prod. (+)
Zapato	80	30	110
Bota	40	60	100
Tot. Local (+)			210

	Local 1	Local 2	Tot Prod. (+)
Zapato	80	30	
Bota	40	60	
Tot. Local (+)	120	90	210

Margen actual/planificado

	Actual	Plan	Act/Plan (/)
Venta	120	90	1.33
Costo	60	30	2
Margen (/)			0.67

	Actual	Plan	Act/Plan (/)
Venta	120	90	
Costo	60	30	
Margen(/)	2	3	0.67

Varianza en vtas actuales/planificadas

	Actual	Plan	Actual/Plan (/)
Zapato	120	100	
Bota	50	60	
Tot Esc. (+)	170	160	1.06

	Actual	Plan	Act/Plan (/)
Zapato	120	100	1.2
Bota	50	60	.83
Tot Esc. (+)			2.03

VISUALIZACION DE MULTIPLES DIMENSIONES:

Es parte importante de OLAP es lo que se vincula directamente con el usuario.

Tabular:

- Lo único que importa es el contenido
- Es la mejor para leer los valores exactos y actuales
- Es detallada al máximo
- EJ: permite ver valores y cantidades de ventas

Gráfica:

- Son gráficos contenidos sobre los números
- Sirve para descubrir tendencias
- Muestra relaciones entre los valores
- Aproximan más a la realidad
- Animaciones para mostrar la evolución
- Importa mucho el color, el tamaño de las rayas y de los dibujos
- EJ: permite ver un mapa de ventas zonas y regiones de un país

Crecimiento:

- Si aplicamos fórmulas a todos los niveles jerárquicos de todas las dimensiones las celdas crecen en forma exponencial

Veamos ejemplo sobre una dimensión:

18 miembros, 10 hojas y 8 derivados → un factor de crecimiento de $18/10$, 1.8

Ejemplo: Dimensión Productos

Todos	→ tipo1	→ talle1	→ producto1
			→ producto 2
		→ talle 2	→ producto 3
			→ producto 4
		→ talle 3	→ producto 5
			→ producto 6
	→ tipo2	→ talle 4	→ producto 7
			→ producto 8
		→ talle 5	→ producto 9
			→ producto 10

- 18 celdas, con 10 valores hoja y 8 derivados
- factor de crecimiento = $18/10 = 1.8$

- 18 celdas, con 10 valores hoja y 8 derivados
- factor de crecimiento = $18/10 = 1.8$

Producto y depósito

	prod.(10)	talles (5)	tipos (2)	todos(1)
1 país	10	5	2	1
3 regiones	30	15	6	3
4 ciudades	40	20	8	4
10 locales	100	50	20	10

**Las celdas representan las cantidades de totales necesarios*

- 324 celdas, con 100 valores hoja y 224 derivados
- $324/100=3.24 \rightarrow$ factor de crecimiento $=3.24=1.8^2$

Conclusión:

- Precomputar sólo aquellos datos
 - Usados más frecuentemente;
 - Cuyo cálculo depende de gran volumen de datos.

Arquitectura

Lo conceptual versus lo físico:

- El modelo multidimensional es la forma más natural de ver la información
- Es poco eficiente usar un formato que refleje fielmente la vista del usuario
 - Guardar una celda por cada combinación posible de valores de dimensiones, ofrecerá muchas celdas vacías (mucha dispersión).
 - No hacerlo dificultará el posterior acceso a cada celda.

Carga del data warehouse:

- Qué grado de actualidad deben tener los datos?
- Pueden quedar offline? Cuánto tiempo?
- Qué disponibilidad de almacenamiento hay?
- Cuál será el tiempo de carga?
- **Elección de arquitectura**
- Creación y mantenimiento de estructuras
- Soporte de actualización (refresh)
- Creación y mantenimiento de vías de acceso adecuadas

Arquitecturas:

- Por qué la importancia de la elección?

Porque la elección determina lo que podremos hacer y como lo vamos a hacer.

Características de cualquier arquitectura:

- Ser soporte de la toma de decisiones
- Permitir vista multidimensional de datos
- Drill (taladrar) sobre las dimensiones
- Slice (rebanar)
- Dice (cortar en cubos)

Memoria vs. Disco:

- Ventajas de datos activos en memoria:
 - Más rapidez (sólo tiempo de CPU)
 - Se disimulan errores de diseño y de tuning
 - Al existir menos datos precalculados, la actualización es más rápida y se requiere menos espacio en disco
- Desventajas de datos activos en memoria:
 - Soporta menos volumen de datos
 - La memoria es más cara
 - Deja de ser viable para muchas aplicaciones
- Realidad:
 - Productos que procesan en memoria utilizan memoria virtual en disco
 - Productos que trabajan en disco pueden, si es posible, operar en memoria de forma transparente (memoria cache)

Datos en disco: subcategorías de OLAP:

- ROLAP : Relational OLAP
La base de datos relacional es el soporte
- MOLAP : Multidimensional OLAP
Se usa un esquema multidimensional
- DOLAP : Desktop OLAP
Herramientas y cubo residen en la PC
- IOLAP/ HOLAP : Herramientas integradas/ híbrido

MOLAP: Características:

- Los datos se almacenan en forma similar a como se usarán
- Para esto se utilizan matrices multidimensionales que representan los cubos
- Usa formatos propietarios (no recibe los avances de RDBM);
- Los cubos incluyen datos precalculados
- Se debe tener en cuenta implicancias en tiempo de carga y almacenamiento
- Con limitaciones en el tamaño de los cubos
- Los datos deben cargarse periódicamente desde una fuente
- La carga y la "precalculación" requieren tiempo significativo
- Performance: rápido tiempo de respuesta
- Capacidad analítica: soporta complejas funciones

MOLAP: Almacenamiento:

- Utilización de Matrices Multidimensionales
- Uso de algoritmos especiales de indexación:
 - Se arman pequeños vectores, con compresión, con las dimensiones que más se usan juntas (el clustering favorece la búsqueda sobre algunas dimensiones)
- Compresión de datos dispersos
- La administración de la base (propietaria) requiere especialistas

VENTAJAS	DESVENTAJAS
<ul style="list-style-type: none"> • Buen tiempo de respuesta. <ul style="list-style-type: none"> ◦ Uso de funciones complejas. ◦ Ocultamiento de la complejidad de la BD. ◦ Es recomendable sobre un universo de datos no muy grande. 	<ul style="list-style-type: none"> • Mayor tiempo de carga de los datos y la estructura del cubo. • Requiere nueva base de datos. • Aislada de la fuente. • Requiere de especialistas para su administración. • Limitaciones en el tamaño del cubo (hasta 10 dimensiones). • Formato propietario.

ROLAP: Características:

- Almacenamiento de los datos en una BD Relacional
- Tecnología mas reciente
- No es la misma base OLTP
- Se utiliza el esquema estrella
- Crece con el poder de los motores relacionales
- Es muy escalable, tanto en dimensiones como en miembros de una dimensión
- Metadata propietaria de aplicación genera el modelo multidimensional
- No utiliza datos precalculados
- Puede eventualmente analizar datos atómicos (acceso a la BD OLTP)
- Acceso a los datos mediante SQL
- Capacidad analítica: soporta funciones complejas
- Performance:
 - Cada "drill", "slice" y "dice" son nuevos query's
 - Es importante el uso de indexación de la BD
 - Depende del poder del motor relacional

ROLAP: Consideraciones especiales:

- La administración de la base corre por cuenta de los técnicos existentes
 - Requiere expertos en el diseño
 - Sufre las limitaciones del SQL para análisis complejos
- Estrategias para enfrentar estas limitaciones:
- Extender SQL con funciones analíticas (SQL Cube y Rollup)
 - Manejar SQL multipasos
 - Trasladar los datos seleccionados a un server con procesamiento multidimensional (Opción híbrida de OLAP)
- Dificultades por existencia de más de una Fact Table
 - Necesidad de definir las precalculaciones como un buen compromiso entre almacenamiento y cálculo

VENTAJAS	DESVENTAJAS
<ul style="list-style-type: none"> • Solución mas escalable. • Utilización de gran numero de dimensiones. • Posibilidad de acceso a base OLTP. • Se aprovecha la capacidad instalada de los SGBD relacionales. • Soporta un universo de datos muy grande. • Posibilidad de utilizar SGBD gratuitos. 	<ul style="list-style-type: none"> • Baja performance en análisis complejos (Cada acción un query). • Requiere nuevo diseño de BD. • Requiere de especialistas para su diseño. • Limitaciones en el uso de datos precalculados. • Necesidad de utilización de Metadata de aplicación.

ROLAP – MOLAP: Herramientas:

ROLAP	MOLAP
<ul style="list-style-type: none"> • Oracle Discoverer; • Microsoft Analysis Services; • MicroStrategy; • Bussiness Objects; • Mondrian (Open Source); • Penthao (Open Source). 	<ul style="list-style-type: none"> • Microsoft Analysis Services; • Oracle OLAP; • Hyperion; • Palo (Open Source)

DOLAP (Desktop OLAP):

- Trabajar localmente con el cubo y sus datos de forma offline
- Útil para usuarios móviles
- Herramientas fáciles de usar
- Performance: rápido tiempo de respuesta
- Capacidad analítica: limitada

VENTAJAS	DESVENTAJAS
<ul style="list-style-type: none"> • Rápido tiempo de respuesta. <ul style="list-style-type: none"> ○ Óptimos para usuarios móviles. ○ Fácil de usar. 	<ul style="list-style-type: none"> • Desconectada de la fuente. • Rápidamente los datos devienen obsoletos. • Debe customizarla para cada usuario. • Capacidad analítica limitada.

IOLAP/HOLAP:

- Solución Integrada o Híbrida
- Puede combinar bases relacionales y multidimensionales (MOLAP y ROLAP)
- Interface simple para query's, reportes y análisis
- Permite modo online y offline
- Puede analizar desde datos atómicos a toda la DW
- Puede distribuir datos en las PC's para agilizar tiempo de respuesta;
- Performance: optimizada dependiendo el tipo de análisis;
- Capacidad analítica: análisis con agregaciones utiliza MOLAP y análisis detallado ROLAP.

VENTAJAS	DESVENTAJAS
<ul style="list-style-type: none"> • Tiempo de respuesta optimizado. • Explotación de los beneficios del resto de las arquitecturas. 	<ul style="list-style-type: none"> • Tecnología reciente. • Mucha dependencia de expertos. • Productos costosos.

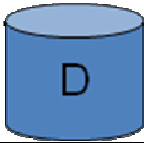




Elección de la arquitectura:

- Cualquier decisión dependerá de:
 - El tamaño de la base
 - La cantidad de dimensiones
 - La escalabilidad demandada
 - El tiempo de respuesta
 - El grado de dispersión (muchos requieren algoritmos más complejos)
 - La frecuencia de la actualización
 - Necesidad de compartir datos con otras aplicaciones
 - La infraestructura instalada

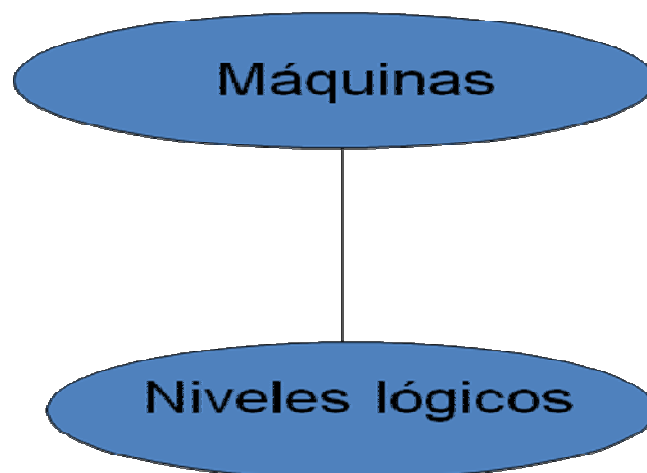
Distribución de datos y procesos

- Datos compartidos por múltiples usuarios lleva a pensar en una arquitectura cliente-servidor, pero:
 - Dónde se realizan los procesos de transformación, "precalculación", armado de consulta y muestra del resultado?
- Soluciones existentes hoy:
 - Arquitecturas cliente-servidor, multicapas, con procesamiento distribuido entre ellas
 - La vuelta del "mainframe" bajo la forma de Webserver y Navegadores Web
 - PCs independientes

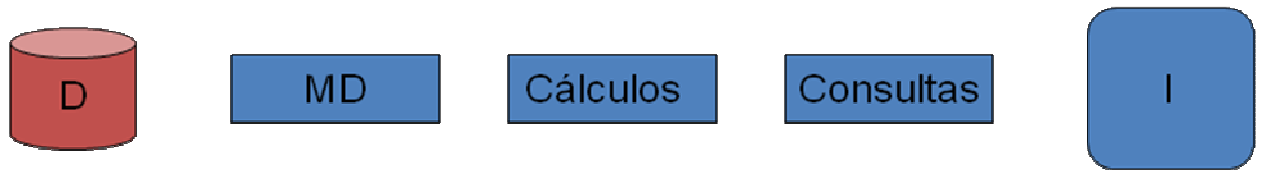
Arquitectura cliente-servidor

Niveles lógicos	
<u>Archivos de datos</u> : A ser compartidos entre los usuarios según niveles de seguridad	
<u>Manejo de datos</u> : Motor propietario MDB o RDBMS; interpone la metadata entre los datos y las aplicaciones	
<u>Cálculos del vuelco</u> : Cálculos "adelantados", en general en el server	
<u>Consultas</u> : Repartidas entre cliente y servidor	
<u>Interface de presentación</u> : En el cliente	

Relación M:N

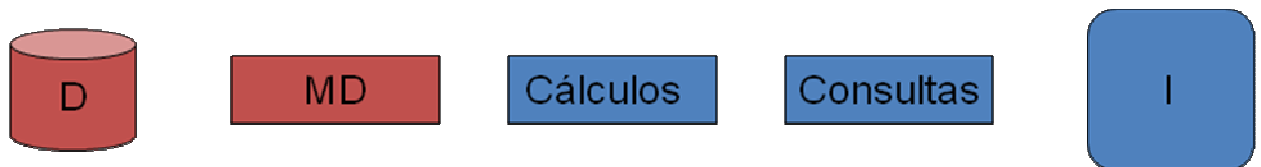


Arquitectura cliente-servidor



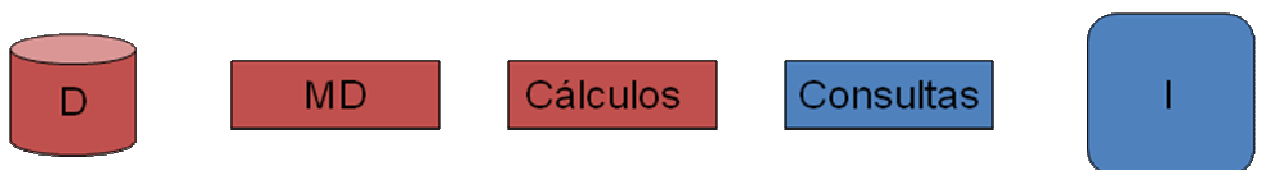
Archivo compartido:

- File-Server económico
- Clientes con gran capacidad (lógica del negocio, administración de datos, y presentación)
- Excesivo tráfico de red para procesar en el cliente
- Seguridad manejada desde el cliente
- No es un verdadero Cliente-Servidor



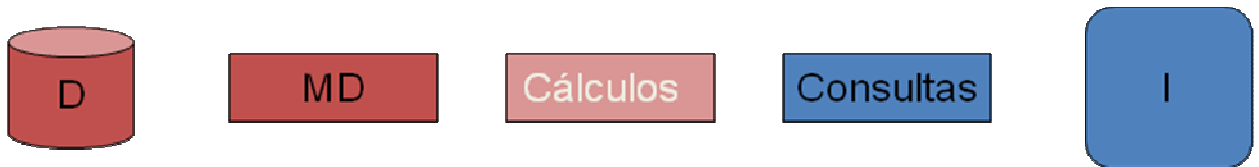
Base Compartida:

- Server para base de datos relacional
- Hay menos tráfico de red
- Se puede manejar seguridad en el server
- Soporta grandes bases de datos
- Una mayor cantidad de usuarios requiere de refinamiento en la administración de la base de datos



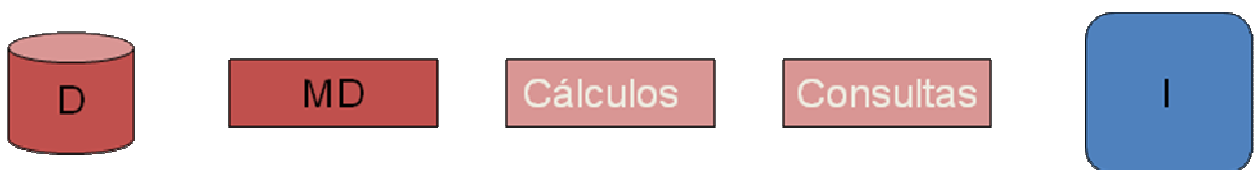
Server OLAP 2 Capas:

- Server de base de datos y de aplicación
- Cliente mas pequeño
- Optima perfomance con poco tráfico
- Mas cerrada que una solución de 3 capas
- En general son productos de uso más específico



Server OLAP 3 Capas:

- Server de base de datos
- Server de aplicación (Flexibilidad comercial)
- Cliente mas pequeño
- Reduce tráfico de red
- Hay dos tipos de server sobre los cuales hacer tuning



Server OLAP 3 Capas con WebServer:

- Server de base de datos
- Server de aplicación (Flexibilidad comercial)
- Web Server integrado a la solución
- Browser conectado por intranet o internet
- Sin costo de mantenimiento de red
- Cliente muy delgado, con distintas plataformas, sin necesidad de mantenerle versiones
- Requiere cliente conectado
- Pierde funcionalidad (applets disponibles)

Server Céntrica	Cliente Céntrica
Datos y cubos en el server	Datos y cubo en el server con opción de bajada al cliente
Procesos en el server	Proceso en el server y el cliente
Dependientes de la red	Asincrónica / Desconectada
Buena performance de query	Más flexible

Arquitectura Datamart:

Dependiente	Independiente
Conectado a un DW empresarial	Sistema aislado
Usa un subconjunto de datos del DW	Los datos provienen de OLTP o son ingresados manualmente

ODS (Operational Data Store)

Características:

- Datos
 - integrados
 - orientados a un tema
 - actualizados en tiempo casi real
 - volátiles
 - corrientes
 - detallados

Beneficios:

Reportes operacionales más rápidos;
Accesibilidad a datos críticos;
Visión completa de un tema;
Con posibilidad de replicar datos en los sistemas operacionales;
Ayuda a alimentar el datawarehouse.

Integración:

- Integración con sistemas operacionales: se actualiza en tiempo casi real, con las siguientes modalidades:
 - Por procesos batch con datos integrados de los sistemas operacionales
 - Desde el ODS se pueden actualizar las bases operacionales mediante triggers
 - Acceso totalmente integrado

Tratamiento de datos

Datos incompletos:

- Causa:
 - Atributos de interés que no están disponibles
 - Tuplas con valores ausentes
 - Atributos que fueron considerados sin importancia
 - Fallas de los sistemas OLTP al grabar
 - No se consideró guardar la historia de los cambios
- Solución: completar

Completar datos incompletos:

- Sin afectar el resultado
 - Ignorar la tupla
 - Llenar el valor manualmente
- Afectando al resultado
 - Usar una constante global
 - Usar el valor promedio
 - Usar el valor promedio para los de la misma clase
 - Usar el valor más probable (regresión, árbol)

Datos con ruido:

- Causa:
 - Instrumentos de recolección defectuosos
 - Errores humanos en el ingreso de datos
 - Limitaciones tecnológicas (tamaño del buffer)
- Solución: aplanar

Aplanar datos con ruido:

- Encajado: suaviza por consulta a su entorno; se arman cajas equiprofundas y se reemplaza cada valor por su media en la caja
- Clusters: para detectar los outliers, que se analizan manualmente
- Regresión: ajustando los valores según una función

Datos inconsistentes:

- Causa:
 - Inconsistencias en nombres o códigos
 - Tuplas duplicadas
- Solución: detectarlas y eliminarlas

Datos excesivos, que afectan la celeridad del proceso de explotación:

- Solución: reducirlos sin afectar su calidad
 - De datos : por agregación (menos filas)
 - De dimensiones: por correlación de atributos redundantes (menos columnas)
 - De espacio: por mecanismos de compresión
 - De valores: ver siguiente

Datos: otras transformaciones:

- Generalización de valores de un atributo (mayor jerarquía);
- Desratización de valores
- Construcción de nuevos atributos que explicitan relaciones entre datos
- Normalización de datos
 - $v = (v - \min) / (\max - \min)$
 - $v = (v - \text{media}) / \text{desvío}$

Situación actual:

Necesidad del DW:

- Porque los datos no son consistentes
- Porque los sistemas operacionales no pueden convivir con largos query's sin perder performance
- El foco de atención de los ejecutivos varía
- Varía sin patrón previsible
- Requieren respuestas inmediatas
- El DW está cuando se lo necesita
- Provee información detallada y sumariada
- Es ideal para detectar tendencias (inf. histórica)
- Evita buscar la fuente para responder a una consulta
- Evita trabajosas extracciones de datos no integrados

El mercado hoy:

- Crecimiento en la proporción de empresas que usan/planean usar DW
 - hoy: 90%
- Crecimiento de las aplicaciones
 - El 60 % de las empresas cree que va a aumentar número de usuarios y tamaño de la DW
- Principal usuario
 - mercado minorista, con esquema estrella otras industrias con abundancia de información (telecomunicaciones, finanzas, transporte, salud), investigando otros modelos lógicos

Un poco de historia:

- En el inicio:
 - compañías con grandes recursos
 - con buenas soluciones tecnológicas
- Luego
 - proveedores abren el juego ofreciendo pequeñas "soluciones quick-start": hard, soft y servicios con bajo costo y tiempo (15-90 días)
- Conviene un pequeño proyecto?
 - Muestra el valor potencial del recurso
 - No tiene posibilidades de crecimiento
 - Enfrenta al problema de la escalabilidad

Consideraciones sobre el crecimiento:

En millones de bytes (megabytes)

1.000(Giga)

Mayoría de Cías.

1.000.000(Tera)

Grandes Cías (30%)

1.000.000.000(Peta)

Unas pocas Cías

Escalabilidad:

- Debe preverse aún con un gran proyecto (100 GB)
- Debe preverse en varias direcciones
 - Espacio en disco
 - Usuarios
 - Ciclos de CPU

Escalabilidad Hard:

- TRES categorías de plataformas escalables
 - SMPs: multiprocesador simétrico (hasta docenas, que comparten memoria física y bus)
 - Clusters: conectar múltiples nodos SMP's con una conexión potente
 - MPP's: interconexión con ancho de banda creciente en función de los nodos

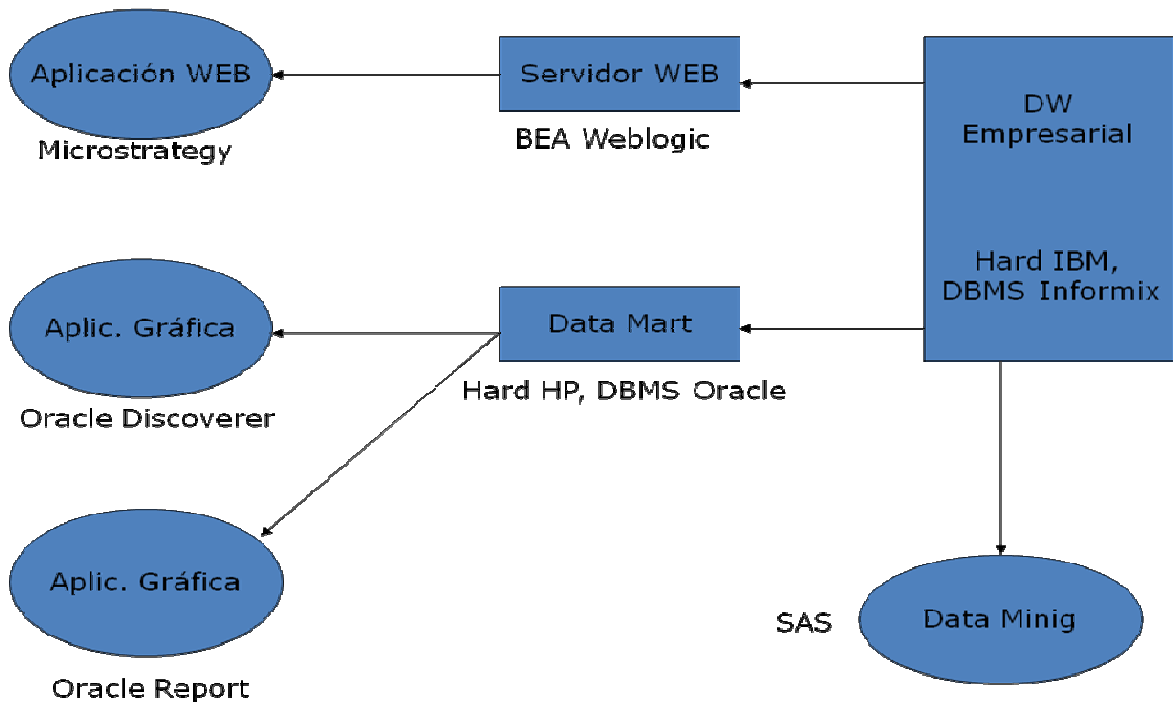
Escalabilidad Soft:

- Debe tomar ventaja de los múltiples procesadores y nodos

Herramientas:

- De modelado y diseño de datos (CASE) o modelos de datos prefabricados por industria
- De ETL
- De administración
 - Planificación de capacidad
 - Back up
 - De seguridad
 - De registro de uso
 - De control de rendimiento
- De usuario final

Ejemplo:



Herramientas de usuario final:

- Herramientas para query's y reportes (no especiales para este entorno)
 - Representación tabular, con sumarización
- Herramientas MULTIDIMENSIONALES que hablan con distintas arquitecturas (ROLAP; MOLAP...)
 - Representación multidimensional (vista)
 - No priorizan el evento sino la acumulación de eventos a lo largo del tiempo
 - Posibilidad de pivotear los resultados
- Herramientas de Data Mining
 - No sólo facilidad de acceder a información consistente
 - Descubrimiento de patrones y tendencias ocultas,
 - Sin tener que formular las consultas necesarias
 - Sin intuiciones previas a confirmar
- Herramientas de Data Mining: Ejemplos
 - Identificar 'canasta' de mercado
 - Identificar patrones de conducta de clientes para encontrar los destinatarios de la publicidad de un nuevo producto
 - Encontrar la mejor combinación de tratamientos para una enfermedad

La visualización:

- Crece la importancia de las técnicas de visualización de los datos obtenidos de cualquier herramienta (es parte del OLAP)
 - A los existentes barras, tortas....
 - Combinación de colores y trazos
 - Animación para mostrar la evolución

Evoluciones divergentes:

- Data Mart (+)
- DW operacional o ODS (-)

Evoluciones Data Mart:

- Subset de una gran DW
- Mejor performance
- Mas simple para entender y mantener
- Mayor autonomía de usuarios, que sólo manipulan sus datos
- Replica datos y descentraliza accesos

Evoluciones DW operacionales:

- Combina DW con sistemas OLTP
- Permite analizar y actuar sobre la base operacional al mismo tiempo
- Los usuarios no son los mismos

Evoluciones:

- SOFT: facilidades de DW
- HARD: tecnología escalable
- HERRAMIENTAS DE ACCESO: más simples y poderosas

Webhousing

Qué es?

Integración de las Tecnologías Web (Internet, Intranets) y DW.

Beneficios:

- Acceso unificado a información consistente
- Explotación de ventajas del browser
 - Mas económico que otras herramientas
 - Más sencillo
 - Requiere menos recursos en cada PC (las herramientas de interfaz gráfica son muy pesadas)
- Explotación de las intranet's
- Evita distribución de herramientas
- Tiene las mismas ventajas que cualquier otra aplicación en la web
- Reduce la complejidad de la plataforma de los clientes
- Se aprovecha toda la infraestructura de red interna
- El costo es inferior al de otra herramienta

Tipos de Servicios:

- Publicación de datos del DW en la Intranet/Internet
- Distribución de Reportes
- Aplicaciones Dinámicas

Implicancias:

- Crea nuevas demandas a la tecnología de DW
 - La necesidad de ajustar la aplicación como server-céntrica(el procesamiento en el server);
 - Las consideraciones de seguridad para datos críticos;
 - Limitar a los usuarios inexpertos que pueden colapsar el sistema.

Desventajas:

- Ocuparse de la seguridad de los datos que viajan en la red
- Necesidad de dialogar con DW Server mediante un Web Gateway (Ej: Librerías OCI para Oracle, Provider for OLAP Services en .NET para SQL Server)
- Velocidad de respuesta
- El Web Server no es Server en el sentido tradicional: Solo se utiliza para transmitir mensajes
- Limitaciones del browser para visualizar reportes complejos
- timeout, porque no hay forma de saber si el cliente sigue activo
- cada drill otro query? (no hay memoria del contexto)
- dificultad para transmitir imágenes pesadas

Los pasos para un efectivo DW

Análisis comercial:

- Exploración (rediseño de procesos, benchmarking..)
- Descubrimiento de una oportunidad
- Identificar objetivos y alcance del proyecto
- Planificación del proyecto y de la infraestructura necesaria
- Reunión y documentación de requerimientos

Estructuración y administración de datos:

- Arquitectura de la base de datos y su compatibilidad con la infraestructura tecnológica
- Diseño de la base de datos
- Extracción: elegir herramientas, identificar sistemas de origen y datos de origen, desarrollar rutinas o códigos
- Transformación ...
- Carga.....
- Implementación y prueba

Selección de herramienta OLAP:

- Determinar tipo de Arquitectura de la aplicación;
- Determinar recursos de interfaz (técnicas de presentación de datos, look-and-feel, etc.)
- Comparar alternativas Propietarias vs Open Source
- Selección de alternativa
- Implementación
- Pruebas
- Distribución y soporte

Funciones en el desarrollo del DW:



Peligros:

- El síndrome del DW como panacea
- Hablar con los usuarios finales equivocados
- Dedicar demasiado tiempo a investigar y perder de vista los destinatarios
- Estancar un proyecto en la creación de metadata
- Desviarse analizando "cosas interesantes"
- Utilizar SSD sin tener opción a decisiones
- Conflictos entre sectores comerciales e informáticos
- Falta de marketing interno
- Olvidar que las aplicaciones tienen un tiempo de vida útil
- Centrarse en los procesos ETL en desmedro de la administración del DW

BI Tendencias

BI Governace:

- Definición y ejecución de la Infraestructura para implementación de BI
- Dirección y Administración de todas las tareas implicadas en soluciones BI:
 - Reingeniería decisional
 - Arquitectura Data Warehouse
 - Selección de herramientas OLAP y DM
- Definición de Ciclo de vida para proyectos BI
- Planes de formación
- Optimización y Tuning
- Soporte de procesos ETL
- Auditorias

BI Open Source:

- Implementación de una Infraestructura de BI utilizando e integrando herramientas y plataformas Open Source
- Opción atractiva para pequeñas y medianas empresas con cultura Open Source

Ventajas:

- Bajo costo
- Solución mas personalizada
- Independencia de proveedores
- Posibilidad de agregar nuevas funcionalidades

Desventajas:

- Necesidad de personal adiestrado en OS
- Dispersión del Soporte
- Implantación de Metodologías formales de control de cambios en la Plataforma OS
- Implementación más lenta

	OPEN SOURCE	PROPRIETARIO
DATABASES	PostgreSQL EnterpriseDB MySQL	Oracle DB2 SQL Server
BI PLATFORMS	Pentaho Jasper Intelligence	BusinessObjects Oracle Microstrategy
ETL	Kettle (aka Pentaho Data Integration) Octopus	Informatica DataStage Oracle Warehouse Builder
QUERY & REPORTING	JFreeReport BIRT JasperReports	BusinessObjects Cognos Oracle Discoverer/Reports
ANALYTICS	R Weka	SAS S-Plus
OLAP	Mondrian + JPivot PALO	Oracle Analytical Workspaces MSFT Analysis Services Hyperion Essbase
APPLICATION SERVERS	JBoss	WebLogic WebSphere Oracle IAS
LANGUAGES	Python Ruby Perl PHP	Java / J2EE
PORTALS	JBoss Portal JetSpeed Liferay	WebLogic Portal WebSphere Portal Oracle Portal
CONTENT MANAGEMENT	Alfresco	Documentum