Unidad 9: DATAWAREHOUSING y OLAP

Bases de Datos

Dentro de una organización o empresa coexisten dos grupos diferentes de aplicaciones



Aplicaciones Tradicionales



Son el soporte para las transacciones del día a día (altas, bajas, modificaciones)



Aplicaciones de Análisis



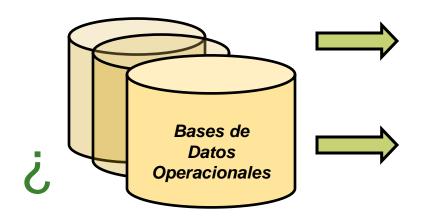
Permiten "analizar" el negocio y dar soporte a la toma de decisiones

- ¿En esta misma época , cuanto vendí el año pasado?
- ¿Cual es el mes de menos ventas?
- ¿Cual es el producto mas vendido?
- ¿Cual es el producto mas vendido el año pasado?

Product Dimension Total Products

Se necesitan datos históricos





También llamadas
Transaccionales y Tradicionales

Contienen los datos de las transacciones diarias

Pueden proveer los datos necesarios para las aplicaciones de análisis que dan soporte a la toma de decisiones



Si, en parte. Sin embargo, los datos provistos serían <u>parciales</u> y su acceso <u>no sería eficiente</u>

Detallando algunos problemas... aún pudiendo encontrar la <u>información</u> necesaria:

- Todos los datos necesarios <u>no están on-line</u>, deben rastrearse los backups correspondientes
- Los datos están esparcidos en distintas bases de datos, inclusive en otras fuentes (internas y/o externas)
- Altera el trabajo transaccional diario, deben postergarse a horarios sin carga de trabajo
- Los <u>tiempos de respuestas</u> no son los esperados
- La <u>estructura de los datos</u> está pensada para dar soporte a tareas transaccionales

Solución:

Almacenar los datos en un <u>sistema separado y específico</u> (datos históricos)



Datawarehouse Almacén de Datos Bodegas de Datos



Información precisa y en un tiempo razonable

Datos Tradicionales



Datos Datawarehouse



objetivo de explotación diferente

BD orientadas al proceso



OLTP

(On Line Transaction Processing)

BD orientadas al análisis



(On Line Analytic Processing)

Comparando OLTP y OLAP...

OLTP vs OLAP

- Datos dinámicos y de detalle
- Muchos usuarios (administrativos)
- Gran cantidad de transacciones concurrentes
- Orientado a los procesos de la organización

- Datos estáticos (históricos), detalle y de resumen
- Pocos usuarios (estratégicos)
- Baja o media cantidad de transacciones
- Orientado al análisis de datos

En esta área <u>existen discrepancias</u> en muchos aspectos, <u>inclusive en el concepto</u> mismo.

- Inmon -

"Un datawarehouse es una colección de datos orientados a temas integrados (provenientes de diversas fuentes), no-volátiles, variante en el tiempo, organizados para soportar necesidades empresariales"

- Ralph Kimball –

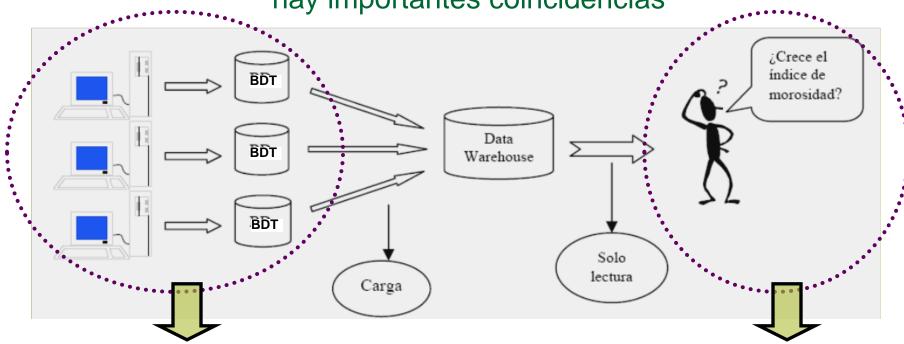
"A <u>data</u> warehouse is a copy of transaction data <u>specifically structured</u> for querying and analysis"

- Larry Greenfield – Disiente con Kimball:

✓ La <u>forma</u> en que los datos estén almacenados no determina que "<u>algo" califique o no</u> como datawarehouse

✓ No todos los datos son transaccionales

Si bien existen opiniones opuestas, hay importantes coincidencias



Fundamentalmente alimentados por bases de datos transaccionales

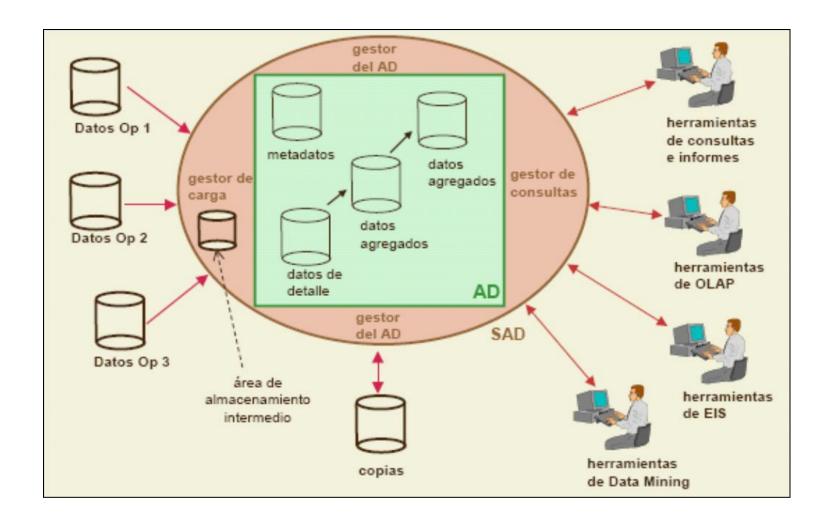
Sistemas de soporte a la decisión

- Procesamiento Analítico-

Concepto: Resumen Características Principales

- Agrupa y mantiene <u>datos transaccionales y no transaccionales</u> <u>de distintas fuentes</u>, incluso externas.
- Mantiene <u>datos históricos</u>.
- La forma en la que se almacenan los datos no es fija.
- El tipo de usuario correspondiente es de nivel gerencial.
- El tipo de proceso a realizar es de análisis.
- El objetivo de todo datawarehouse es proveer información suficiente y oportuna asistiendo a la toma de decisiones para alcanzar los objetivos del negocio.

Arquitectura



Arquitectura: Fuentes de Datos

- Bases de Datos (Internas o Externas)
- Datos contenidos en estructuras no base de datos (Internas o Externas)

Arquitectura: Gestor de Carga

Permite realizar la <u>extracción</u> de los datos desde las fuentes y los <u>carga</u> al datawarehouse



Encargado de soportar el proceso ETL (Extraction – Transformation- Load)

- Extracción
- Transformación de los datos: Limpieza, Consolidación, Agregación de datos, etc.
- Carga Inicial
- Recarga Periódica (Refresco): Operación que propaga los cambios desde las fuentes

Arquitectura: Datos (Almacén)

En general, refiere a la base de datos física que contiene los datos

Arquitectura: Metadatos

Datos sobre los datos:

- Estructura de datos de las fuentes de datos
- Reglas de proceso para <u>transformar los datos</u> de <u>origen</u> a los datos que contendrá el data warehouse (limpieza, cálculo y equivalencias, definiciones de agregación, etc.)
- Calendarios de ejecución de los procesos, etc.
- <u>Estructura</u> de los <u>datos del data warehouse</u>

Arquitectura: Herramientas Clientes/Usuarios

Permiten extraer información/conocimiento

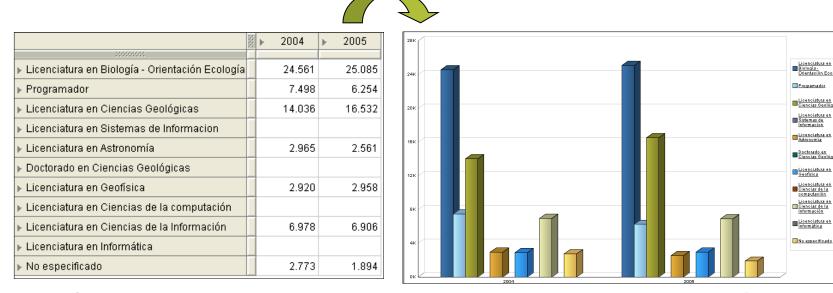
- Herramientas para diseñar consultas e informes (Presentación y Visualización)
- Herramientas estadísticas
- Herramientas de análisis de datos (OLAP)
- Herramientas de minería de datos
- Herramientas de EIS

Las herramientas OLAP:

- Están <u>basadas</u>, generalmente, en <u>sistemas o interfaces</u> <u>multidimensionales</u>.
- Utilizan <u>operadores específicos</u> (además de los clásicos): <u>drill dwon,</u> <u>roll up, pivot, slice, etc.</u>
- El <u>resultado</u> se <u>presenta</u> generalmente de <u>manera matricial</u>, y permite <u>generar diferentes tipos de gráficos</u>.

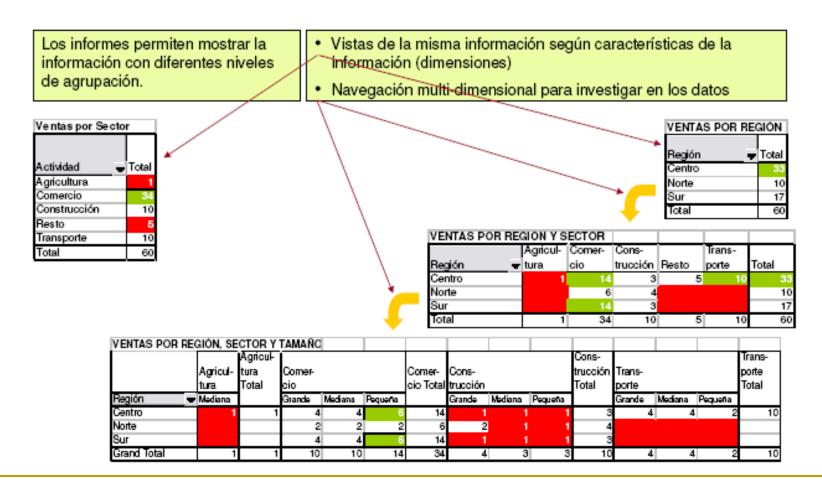
En resumen, permiten realizar diferentes combinaciones de datos para visualizar los resultados hasta un grado determinado de detalle, permitiendo navegar por sus dimensiones y analizar sus datos desde diferentes puntos de vista

Un ejemplo de la funcionalidad permitida en una herramienta OLAP



Cantidad de prestamos bibliotecarios con dos ejes de análisis (carrera y año) en formato de tabla y gráfico.

Otro ejemplo de la funcionalidad en una herramienta OLAP



La tecnología OLAP generalmente se asocia a los almacenes de datos, aunque se puede tener DW sin OLAP y viceversa

Herramientas de Minería

Si bien existen herramientas que permiten en análisis de datos, como las OLAP



Existen casos donde no son adecuadas

¿Por qué?

- La cantidad de información es demasiada
- Se necesita un análisis "inteligente"



Minería / Data Mining

Datawarehouse vs. Datamart

Es frecuente encontrar los términos datawarehouse y datamart usados en forma equivalente.

Sin embargo, existen **semejanzas** y **diferencias** entre ambos conceptos.

Semejanzas:

- Contienen datos históricos provenientes de fuentes operacionales.
 - El tipo de proceso que se efectúa sobre ambos, es de análisis.

Diferencias:

En este punto, las opiniones se dividen en dos grupos





Según Inmon

Según Kimball

Concepto: Datawarehouse vs. Datamart

Inmon

- Un <u>datawarehouse es fuente</u> de datos de todo <u>datamart</u>
- Los <u>"clientes" naturales</u> de un <u>datamart</u> son las herramientas <u>OLAP</u>
- El tipo de estructura de un datamart son diferentes a las de un datawarehouse

Kimball

- La diferencia está dada fundamentalmente por el tamaño de la base de datos, en términos de objetos de análisis que mantienen (no en cuanto a extensión sino a intensión)
- El tipo de estructura en un datamart y en un datawarehouse es la misma

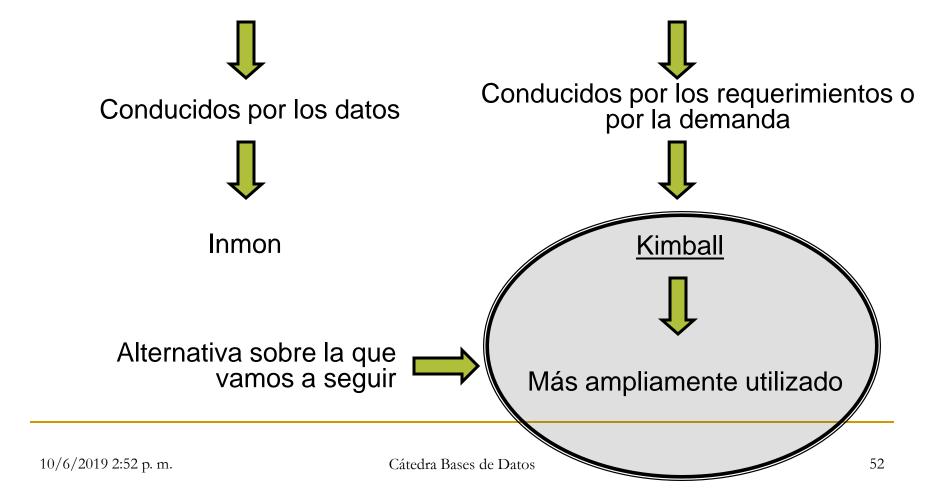
Repasando...

- Necesidades
- Concepto
- Arquitectura
 - Fuentes de datos
 - Gestor de Carga: Asiste en el proceso ETL
 - Gestor del almacén
 - Usuarios o Clientes
- Datamarts vs. Datawarehouse

Diseño Multidimensional CONSTRUCCIÓN DE UN DATAWAREHOUSE

Construcción/Diseño de un datawarehouse

De los métodos existentes se distinguen 2 enfoques



Proceso de Construcción (Conducido por los Requerimientos)

Recordemos...

Objetivo de todo datawarehouse



Proveer información suficiente y oportuna asistiendo a la toma de decisiones <u>para alcanzar los **objetivos del negocio**</u>

Proceso de Construcción: Requerimientos

La fase de relevamiento



Alineada a los objetivos del negocio

Proceso de Construcción: Requerimientos

La <u>fase de relevamiento</u> que debe <u>identificar</u>?



Datos a contener en el DW, además de:

- <u>Exactitud de los datos</u>: Conformidad del dato en relación al valor en el mundo real. <u>Dos factores influyen en este aspecto</u>, exactitud de los datos en las fuentes y errores en el proceso de carga.
- **Completitud**: Capacidad del sistema de representar todos los estados del mundo real
- Consistencia
- Oportunidad: Nivel de actualización necesaria para la tarea que lo usa



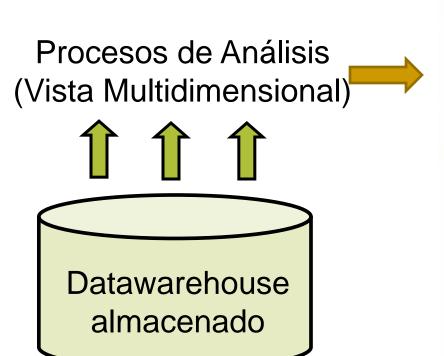
Fuentes de donde extraer los datos, además de:

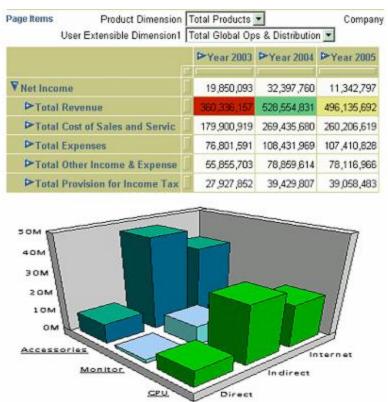
- Disponibilidad
- Tipo:
 - Concretas: Los datos se encuentran tal como los necesitamos
 - □ Adicionales: Es necesario combinar datos de ≠ fuentes



Políticas de actualización

Proceso de Construcción: Diseño



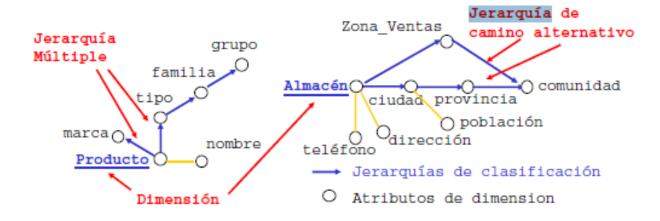


Proceso de Construcción: <u>Diseño</u>

La <u>mayor parte de la bibliografía</u> adopta (tácitamente) una perspectiva <u>relacional</u> debido a la importancia y supremacía de las bases de datos relacionales

Proceso de Construcción: <u>Diseño</u>

Comúnmente el modelado conceptual se realiza a través de gráficos a cíclicos dirigidos:

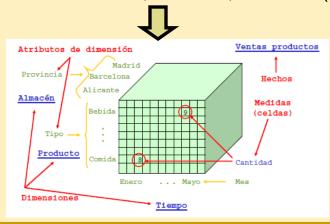


Proceso de Construcción: <u>Diseño</u> - Enfoque Relacional

La única estructura de datos que soporta el enfoque relacional es la relación (tabla)

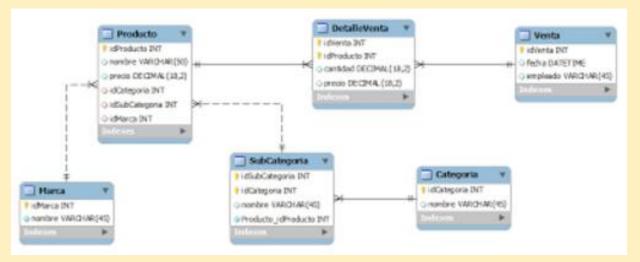


- Un datawarehouse implementado en un motor relacional estará formado por tablas
- Sin embargo, deberá permitir análisis dimensionales, disponiendo la información preparada para su vista de manera multidimensional, es decir, un cubo (arreglo n-dimensional)



Enfoque Relacional

Como ya sabemos, la <u>estructura propia de las bases</u> de datos relacionales son las tablas, con claves primarias y foráneas



Las <u>tablas</u> deberán estar <u>vinculadas</u> de tal manera de poder brindar la vista multidimensional



Esquemas

Estrella, Copo de Nieve, Constelación de Estrellas, Estrella Agrupado, etc.

Cualquiera sea el esquema o modelo utilizado:

- Tabla de hechos
- Tabla de dimensiones

Tabla de Hechos (variables dependientes)

 $\hat{\mathbb{T}}$

Describen

<u>datos sobre actividades básicas</u>

de la organización que son objeto de análisis

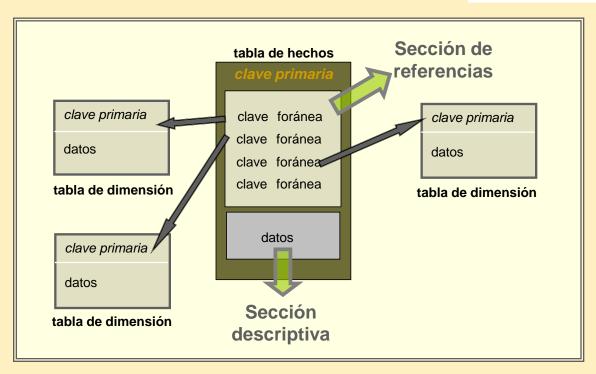
Tabla de Dimensiones (variables independientes)



Describen los <u>objetos</u> relevantes de la organización por los cuales se analiza la actividad

Provincia Madrid Barcelona Alicante Medidas (celdas) Producto Comida Repro ... Mayo Mes Mes Tiempo

Esquema Estrella:



Esquema Estrella:

- Una tabla de hechos y n tablas de dimensiones
- Las tablas de dimensiones, se relacionan directamente a la tabla de hechos, a través de claves foráneas

Atributos Tabla de hechos:

- Sección de referencias (dimensiones)
- Sección descriptiva (medidas)

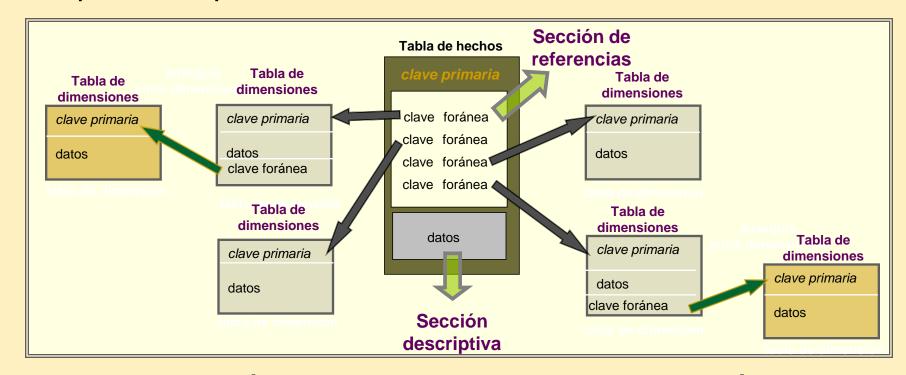
Atributos Tabla de dimensiones:

- Identificador del objeto
- Atributos descriptivos, es decir, propiedades del objeto

Claves utilizadas:

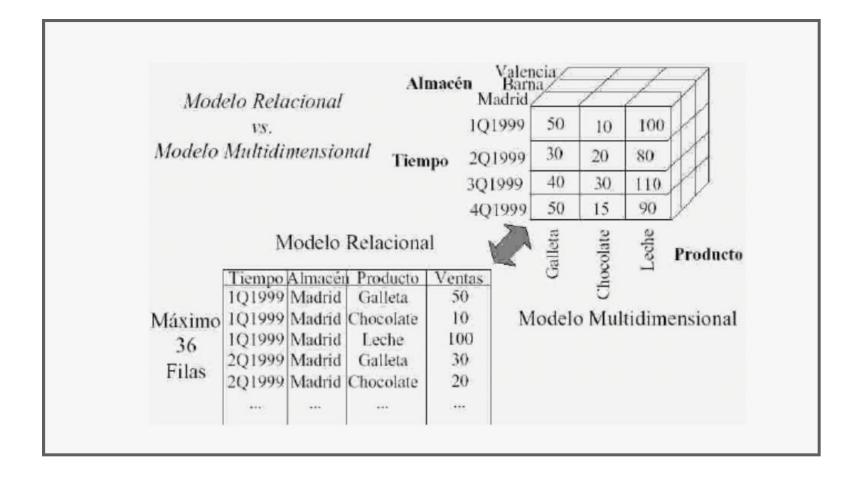
- Autogeneradas (subrrogadas):
 - Aumentan el rendimiento
 - Mas fáciles de manejar en los procesos de ETL
- Con significado semántico

Esquema Copo de Nieve:



Extensión del esquema estrella cuando las jerarquías dimensionales quedan explicitas, separadas en diferentes tablas

Proceso de Construcción: <u>Comparación de Modelos</u>



Normalización:

Las <u>desventajas</u> de un esquema de base de datos relacional no normalizado están relacionadas fundamentalmente con los <u>problemas de update</u>

- En un datawarehouse no se efectúan "update" como en una base de datos transaccional
- Se incrementa en forma periódica

Datawarehouses



Normalización pierde importancia

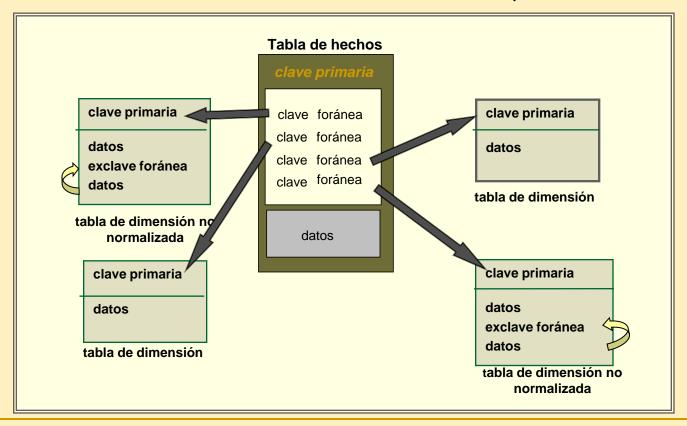


Se pueden generar esquemas no normalizados para alcanzar una mejor performance

Proceso de Construcción:

Diseño - Enfoque Relacional

Esquema Estrella (no normalizado 3FN):



Dimensión Tiempo:

- Presente en todo dw
- Aunque SQL ofrece funciones sobre el tipo DATE, la representación de la dimensión permite representar otros atributos no calculables en SQL
- Atributos frecuentes: nro. de día, nro. de semana, valores absolutos del calendario juliano que permiten ciertos cálculos aritméticos, día de la semana (lunes, etc.), día festivo, fin de semana

Proceso de Construcción:

<u>Diseño</u> - Enfoque Relacional

Analizando el enfoque Relacional



Ventajas:

- Tecnología madura
- Extremadamente utilizada y conocida
 - Sistemas altamente escalables
- Poseen lenguaje de consulta standard
 - En general, las organizaciones ya poseen las licencias



Desventajas:

- Las estructuras del modelo no son naturales para el procesamiento OLAP
- Por ello necesitan contar con una capa intermedia que mapee ambas estructuras provocando una baja performance en las consultas

Otros aspectos a considerar en la fase de diseño:

- Aditividad de las medidas
- Tipos de Jerarquías
- Relación muchos a muchos entre las Tabla de Hechos y Tabla de Dimensión
- Granularidad
- Ventana de Tiempo a mantener en el DW
- Refresco

- Granularidad:
- La granularidad tiene que ver con el nivel de detalle de la información almacenada en las estructuras, es decir, el nivel de detalle de los datos
- A mayor nivel de granularidad:
 - Mayor tiempo de procesamiento en la carga
 - Menor volumen de información
 - Consultas con mejor performance
 - Escasa flexibilidad ante consultas no planificadas
- Analizar cuidadosamente el nivel de detalle adecuado en cada caso de manera de poder <u>balancear ventajas</u> y <u>desventajas</u>

- Granularidad:
- Dependiendo de la implementación elegida, corresponderá:
 - Modelo Relacional: Cuan resumida o no es la información de la tabla de hechos
 - Modelo Multidimensional: El grado de detalle de la variable dependiente

Ventana de Tiempo a mantener en el DW

Para ello se deberá responder a la siguiente pregunta: ¿Cuánta historia debe mantener el datawarehouse?

- Refresco:
- Definir el intervalo de tiempo oportuno entre cada actualización del dw
- Se deberá balancear:
 - Sobrecarga del ambiente operacional
 - Actualidad del datawarehouse

Lenguaje SQL

El standard SQL da soporte a las Bases de Datos multidimensionales, con extensiones que proveen las funcionalidades necesarias



cláusula group by

funciones específicas

Lenguaje SQL

Cláusula group by:

GROUP BY < grouping specification>

```
<grouping specification>::=
  <grouping column reference list> |
  ROLLUP (<grouping column reference list>) |
  CUBE (<grouping column reference list>) |
  GROUPING SETS (<grouping set list>) | ()
```

Operador Rollup

El operador **ROLLUP** mantiene los grupos formados por GROUP BY y además añade los superagregados de la primera columna del group by.

 Obtener el nº de unidades pedidas por categoría, país y año, con los subtotales por categoría.

SELECT CategoriaNombre, Pais, Año, SUM(UnidadesLinea) AS suma

FROM Ventas_Fact INNER JOIN Producto_Dim ON
Ventas_Fact.ProductoKey = Producto_Dim.ProductoKey INNER JOIN
Cliente_Dim ON Ventas_Fact.ClienteKey = Cliente_Dim.ClienteKey
INNER JOIN Tiempo_Dim ON Ventas_Fact.TiempoKey =
Tiempo_Dim.TiempoKey

GROUP BY ROLLUP (CategoriaNombre, Pais, Año)

Operador Cube

El operador **CUBE** mantiene los grupos formados por GROUP BY y además añade los superagregados para cada columna.

· Obtener el nº de unidades pedidas por categoría, país y año con todos sus subtotales

SELECT CategoriaNombre, Pais, Año, SUM(UnidadesLinea) AS suma

FROM Ventas_Fact INNER JOIN Producto_Dim ON
Ventas_Fact.ProductoKey = Producto_Dim.ProductoKey INNER JOIN
Cliente_Dim ON Ventas_Fact.ClienteKey = Cliente_Dim.ClienteKey
INNER JOIN Tiempo_Dim ON Ventas_Fact.TiempoKey =
Tiempo_Dim.TiempoKey

GROUP BY CUBE (CategoriaNombre, Pais, Año)

Operador Grouping Set

El operador GROUPING SET permite construir en una sola consulta múltiples grupos. Los grupos se combinan con un UNION ALL para ofrecer el resultado final

Obtener el nº de unidades pedidas por categoría y país y por país y año

SELECT CategoriaNombre, Pais, Año, SUM(UnidadesLinea) AS suma

FROM Ventas_Fact INNER JOIN Producto_Dim ON
Ventas_Fact.ProductoKey = Producto_Dim.ProductoKey INNER JOIN
Cliente_Dim ON Ventas_Fact.ClienteKey = Cliente_Dim.ClienteKey
INNER JOIN Tiempo_Dim ON Ventas_Fact.TiempoKey =
Tiempo Dim.TiempoKey

GROUP BY

GROUPING SETS ((CategoriaNombre, Pais),

(Pais, Año))

CategoriaNombre	Pais	Año	suma
Bebidas	UK	NULL	397
Bebidas	USA	NULL	1352
Cámicos	UK	NULL	132
Cárnicos	USA	NULL	885
NULL	UK	2003	358
NULL	UK	2004	173
NULL	UBA	2003	1274
NULL	USA	2004	993

Roll-up Drill-down

Sobre las tablas del esquema SH de Oracle

Ejemplos:

Agrupamiento común

```
select prod_id, cust_id, sum(amount_sold)
from sh.sales
group by prod_id, cust_id;
```

2. Disminuyo el nivel de detalle respecto de (1):

```
select prod_id, sum(amount_sold)
from sh.sales
group by prod_id;
(ROLL-UP)
```

3. Aumento el nivel de detalle :

```
select prod_id, cust_id, channel_id, sum(amount_sold)
from sh.sales
group by prod_id, cust_id, channel_id;
(DRILL-DOWN)
```



Procesamiento OLAP TABLAS DINÁMICAS EN EXCEL

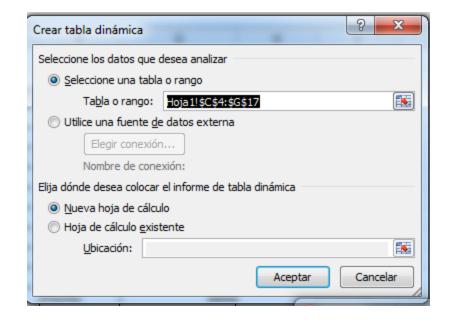
Tablas Dinámicas

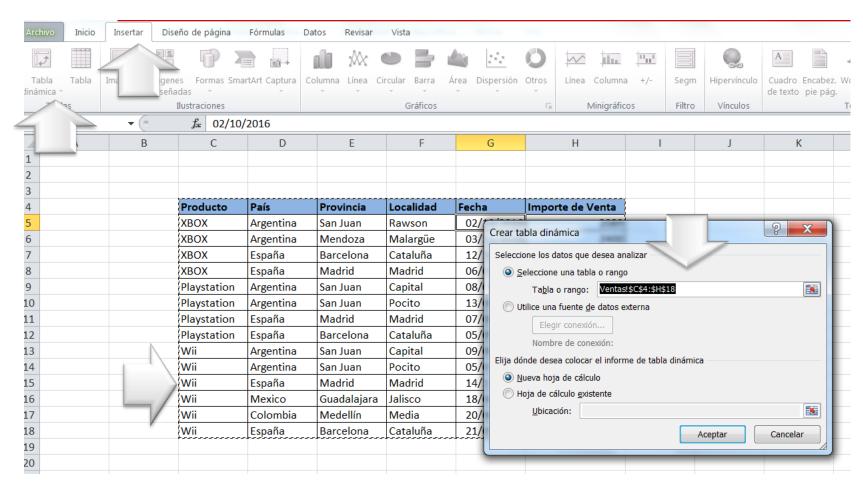
Las tablas dinámicas (también conocidas como Pivot Tables), son una herramienta para analizar grandes cantidades de datos en forma resumida y ordenada.

Pasos para la creación de Tablas Dinámicas:

Selecciona el rango de celdas que contienen la información.

Pacer clic sobre la cinta Insertar y luego en Tabla dinámica.

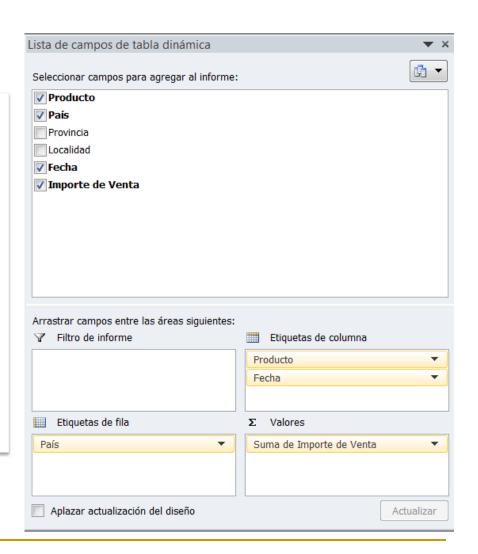




3°

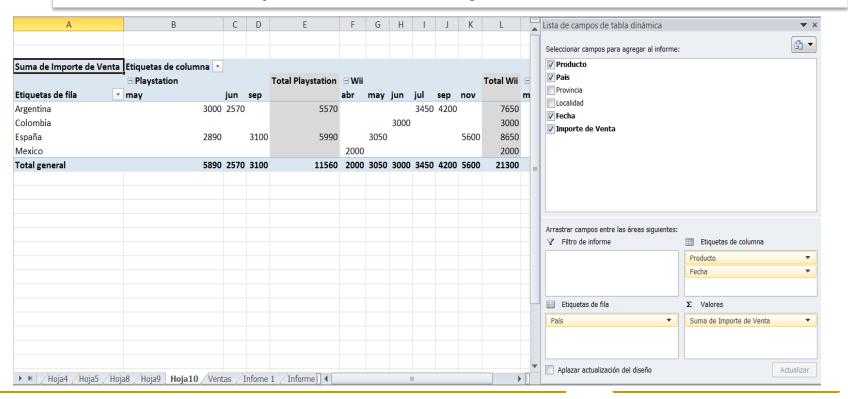
4°

 A continuación se muestra en la parte derecha una ventana con todas las columnas (El nombre se determina por el texto en la primera celda de cada columna de la selección),



5°

 Se deben "Seleccionar las columnas para agregar al informe" a una o más de las cuatro secciones que están debajo.



Secciones para agregar datos a una tabla dinámica

6°

- Filtro de informe: Columnas por las que puede filtrarse la tabla dinámica.
- Etiquetas de columna: Información que se mostrara como nuevas columnas.
- Etiquetas de fila: Información que se mostrará como filas.
- Valores: Valores que van a totalizarse (por ejemplo sumas, promedios, etc.)





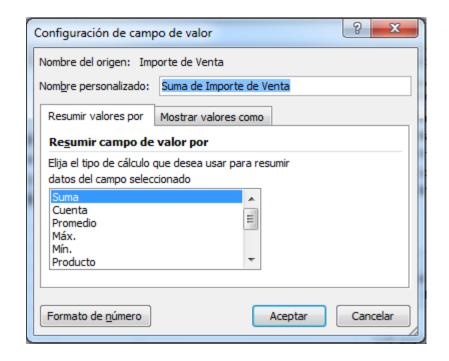




А	В	С	D	E	F	G	Н	-1	J	K	L	M	N	0	Р	Q
Suma de Importe de Venta Etiquetas de columna 🔻																
	■ Playstation			Total Playstation	∃Wii						Total Wii	\exists XBOX			Total XBOX	Total general
Etiquetas de fila	may	jun	sep		abr	may	jun	jul	sep	nov		mar	oct	nov		
Argentina	3000	2570		5570				3450	4200		7650		2089	2400	4489	17709
Colombia							3000				3000					3000
España	2890)	3100	5990		3050				5600	8650	2600	2500		5100	19740
Mexico					2000						2000					2000
Total general	5890	2570	3100	11560	2000	3050	3000	3450	4200	5600	21300	2600	4589	2400	9589	42449

Cambiar la forma en que se totalizan los valores (suma, mayor, promedio, etc)

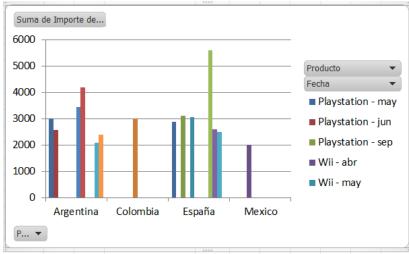




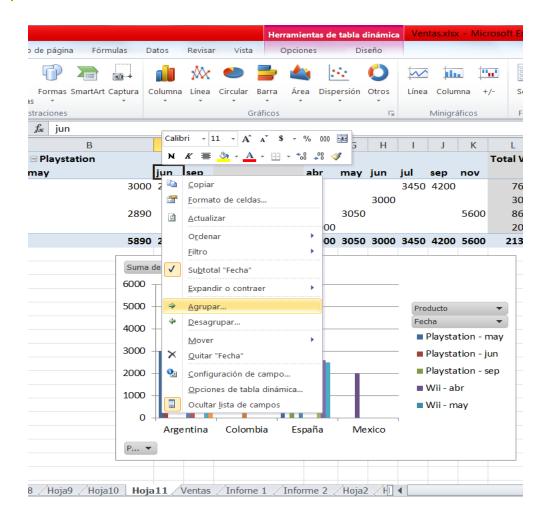
Creación de Gráficos Dinámicos

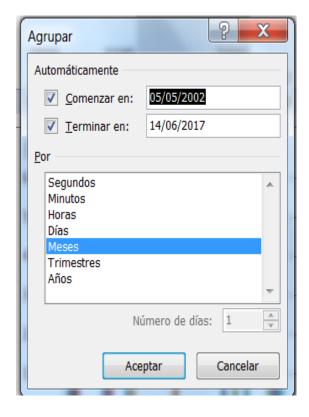
- Hacer clic en cualquier punto de la tabla dinámica para que aparezcan las herramientas de la misma en la cinta.
- Insertar seleccionar el gráfico que se desee.



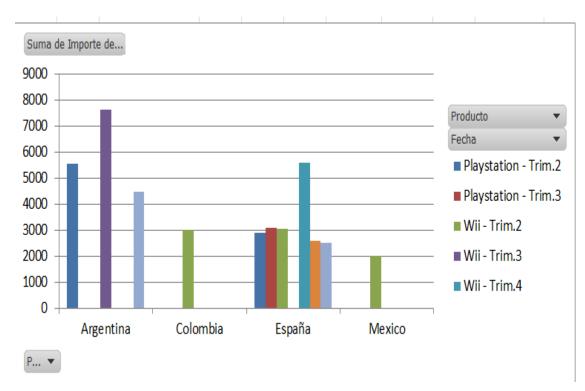


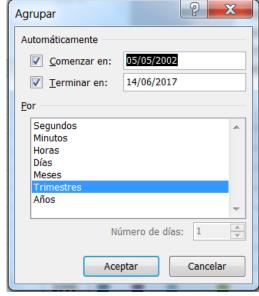
Creación de Gráficos Dinámicos





Creación de Gráficos Dinámicos





Importante cualquier cambio en la tabla de datos requiere una actualización de la tabla dinámica, para ello pulsar Alt+F5

Revisando lo visto...

- Introducción Conceptos
- Aspectos de Diseño (Relacional)
- Procesamiento OLAP:
 - Cláusula HAVING Cube, Rollup, Grouping Set
 - Tablas Dinámicas en Excel

FIN