Diploma de Especialización en Computación.

Sistemas de Data Warehousing

Instituto de Computación - Facultad de Ingeniería
Julio 2003

Plan General

- **Introducción al Data Warehousing.**
  - Motivaciones y conceptos generales.
  - Características Técnicas.
- **Diseño Conceptual.**
  - Conceptos generales y proceso de diseño.
  - Modelos Multidimensionales.
  - Estrategia basada en requerimientos.
  - Estrategia basada en datos.
- **Diseño Lógico.**
  - Conceptos generales y proceso de diseño.
  - Diseño Lógico Multidimensional.
  - Diseño Lógico Relacional.
  - Proceso de Carga y Actualización.
- **Aspectos Tecnológicos y Metodológicos**
  - Arquitecturas de Sistemas de DW
  - Tecnologías en DBMS.
  - Incorporación de la tecnología.
- **Conclusiones y Perspectivas.**

Presentación

- **Docentes:**
  - Grupo Concepción de Sistemas de Información.
    - http://www.fing.edu.uy/inco/grupos/csi
  - Dr. Ing. Raul Ruggia (encargado del curso).
  - Ing. Adriana Marotta MSc,
  - Ing. Veronika Peralta MSc,
  - A/C Lorena Etcheverry,

Temario Introducción

- **Introducción al Data Warehousing:**
  - Introducción.
    - Motivaciones.
    - Conceptos generales.
  - Características Técnicas.
    - Herramientas Front-End.
    - El Data Warehouse.
    - Proceso de ETL.
  - Conclusiones.

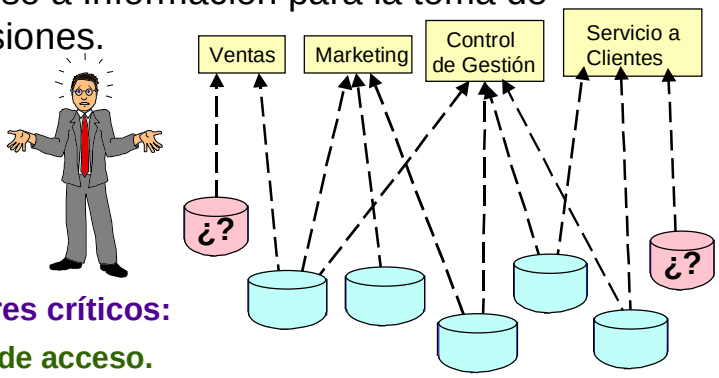
Introducción

Introducción

- **Temas:**
  - Motivaciones: La información y las organizaciones
  - Características de los Sistemas de DW
  - Desarrollo de Sistemas DW.

Motivaciones

- **Problemática planteada:**
  - Acceso a Información para la toma de decisiones.



☛ **Factores críticos:**

- **Tiempo de acceso.**
- **Integración y Calidad de información.**

La información y las organizaciones

- **Las organizaciones tienen necesidad de:**
  - Conocimiento:
    - Materia prima para toma de decisiones.
    - Es lo que se desea construir.
  - Información:
    - Materia prima para conocer los fenómenos reales.
    - Un item de datos es información según el contexto de toma de decisiones.
  - Datos:
    - Materia prima de la información.
    - Generados por procesos que no necesariamente los explotan.

La información y las organizaciones

- **Los datos existen, pero ...**
  - No siempre se acceden facilmente.
  - No siempre se explotan.
    - Un reporte de los Laboratorios Bell indica que la cantidad de datos se duplica cada 5 años, y que solo se usa un 5% de ella.
- **La información suele ser dificil de obtener:**
  - Deben obtenerse los datos:
    - A partir de los cuales se construye la información.
    - Que definen el contexto del mismo.
  - En un cierto contexto, un item puede ser información:
    - Dependiendo del tipo de decisiónes a tomar.
    - Dependiendo de la persona encargada.
    - Dependiendo de la calidad de su valor.

La información y las organizaciones

- **Y los sistemas de información tradicionales ...**
  - Orientados a sistemas operacionales.
  - Asociados a procesos productivos.
  - Procesan grandes cantidades de transacciones.
- **Pueden resolver estas necesidades ?**

Sistemas de Producción y de Decisión

- **Sistemas orientado a la Producción:**
  - Prioridad:
    - tiempo de respuesta a transacciones read-write.
  - Se manejan datos actuales muy detallados.
  - Estables y de larga vida util.
- **Sistema orientado a la Decisión:**
  - Prioridad:
    - expresividad y eficiencia en consultas complejas.
  - Datos actuales+históricos resumidos.
  - En constante evolución.

Sistemas de Producción y de Decisión

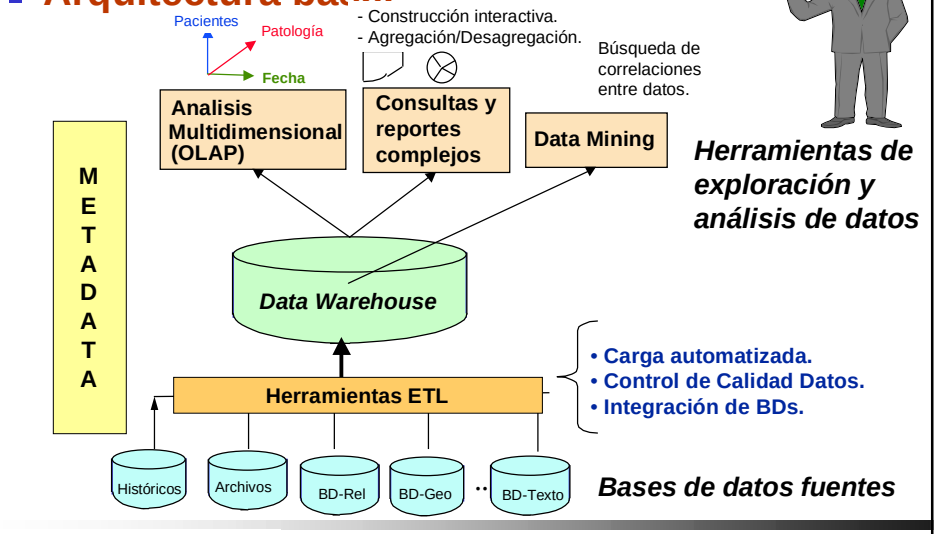
- **Conclusión.**
  - Se trata de sistemas con objetivos diferentes.
  - Se construyen para ser eficientes en sus objetivos.
- **No es posible usar uno para las tareas del otro.**

Sistemas de Data Warehousing

- **Atacan la problemática planteada:**
  - Generar Información para toma de decisiones.
- **siguiendo los principios:**
  - Construir Información a partir de datos objetivos.
  - Integrar diferentes fuentes de datos.
  - Ofrecer al usuario final mecanismos flexibles para el acceso a la información:
    - Pre-programada.
    - Libre, exploratoria.
    - A través de los objetos de su negocio.
    - Observando los datos en formatos especializados.

Sistemas de Data Warehousing

- **Arquitectura base.**



Sistemas de Data Warehousing

- **Algunos conceptos:**
  - Diccionario de Datos o Metadata:
    - Asocia objetos del negocio a datos en BDs.
  - Análisis multidimensional y herramientas OLAP:
    - Modelamiento del problema en dimensiones.
  - Data Mining:
    - Búsqueda de correlaciones entre datos.
  - Calidad de Datos
    - Se agregan criterios de Relevancia y Pertinencia de Datos.

SDW: Visión General

- **Definiciones:**
  - Data Warehouse [Inmon 94]:
    - Es un conjunto de datos orientados a temas, integrados, no volátiles e históricos, organizados para soportar un proceso de toma de decisiones.
  - Sistema de Data Warehousing:
    - Es un sistema informático capaz de ofrecer información para toma de decisiones, y cuya pieza principal es un Data Warehouse.

Sistemas de Data Warehousing

- **Definiciones (cont.):**
  - Datos Orientados a Temas:
    - En los DW, los datos se organizan en torno a los Temas principales de la organización
  - Datos integrados:
    - Heterogeneidad de datos:
      - Diferentes áreas de la organización.
      - Diferentes tipos (tradicionales, geográfico, documentos).
    - Aspectos a resolver en la integración:
      - Unificación de conceptos.
      - Construcción del dato integrado a partir de los fuentes.

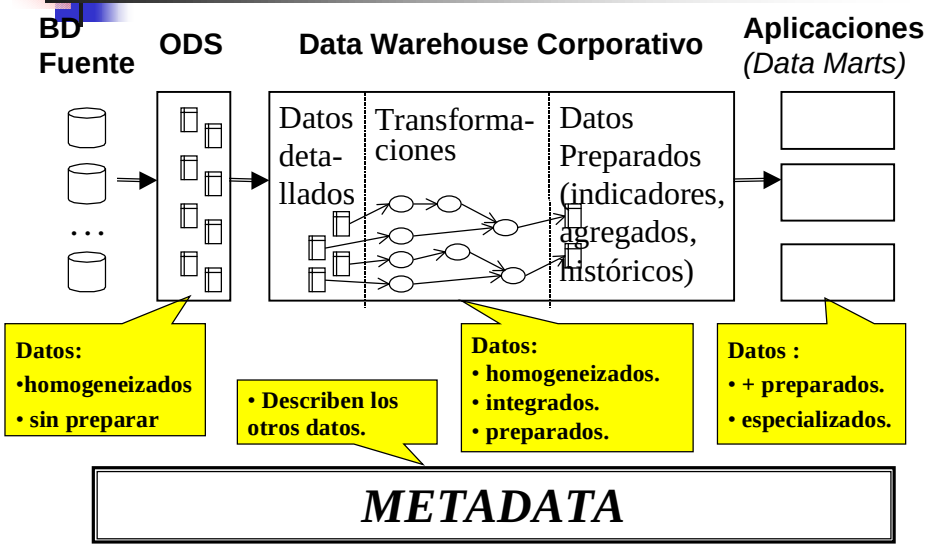
Sistemas de Data Warehousing

- **Definiciones (cont.):**
  - Datos históricos:
    - Se deben manejar los datos con su referencia temporal.
  - Datos no volátiles:
    - Los datos deben ser lo suficientemente estables como para permitir análisis “largos” sin que cambien durante el mismo.
    - Esto se obtiene como consecuencia de:
      - La historización.
      - La planificación de la carga.

Sistemas de Data Warehousing

- **Los Data Marts.**
  - Son aplicaciones de análisis de datos en áreas precisas de negocios.
  - Por ejemplo:
    - Ventas, Marketing, Recursos Humanos.
  - Toman sus datos del Data Warehouse.
  - Priorizan la funciones de análisis de datos:
    - Interfaces a usuario.
    - Indicadores específicos al área de negocio.
  - Normalmente basados en OLAP.

Estructura del Data Warehouse



Estructura del Data Warehouse

- **Clases de datos.**
  - Datos detallados.
    - Corresponden al punto de inserción de datos desde los sistemas de producción.
    - Diferencia con datos de BD-Fuentes:
      - Están en formato homogeneo (via ODS).
      - Pasaron un primer control de calidad.
      - Pueden ya haber pasado un primer proceso de agregación.
    - Generalmente se trata de volúmenes importantes de datos.

Estructura del Data Warehouse

- **Clases de datos (cont.):**
  - Datos agregados.
    - Resultantes de aplicar funciones de totalización sobre datos correspondientes a un objeto del problema. Por ejemplo: total mensual de ventas por producto.
    - Interés:
      - Información significativa para analizar.
      - Permiten reducir volúmenes de datos.
    - Cálculo interactivo plantea problemas de performance.
  - Datos historizados:
    - Datos (base o agregados) a los cuales se les agrega una marca de tiempo.
    - Generan volúmenes importantes de datos al acumular cada conjunto de datos correspondiente a un valor de tiempo.

Estructura del Data Warehouse

- **Clases de datos (cont.):**
  - Metadatos.
    - Consiste en información sobre los datos del DW.
    - Incluye información sobre:
      - Semántica de los datos y su localización en el DW.
      - Localización de los datos en los sistemas de producción y reglas de transformación.
      - Especificación de formulas de calculo de agregados.
      - Información sobre frecuencias de carga, mecanismo de historización, etc.
    - Constituye una pieza clave para:
      - El control de calidad de los datos.
      - La explotación eficaz del DW.

Estructura del Data Warehouse

- **Tipos de Operaciones/Transformaciones (1):**
  - Extracción de datos.
    - Consiste en extraer los datos de la BD fuente y cargarlo en el ODS o DW.
  - Filtrado.
    - Consiste en filtrar datos no admisibles en el DW.
  - Modificación de formato o valores.
    - Consiste en adaptar formatos o valores para que cumpla pautas definidas en el DW.
  - Integración.
    - Consiste en integrar datos provenientes de dos fuentes.

Estructura del Data Warehouse

- **Tipos de Operaciones/Transformaciones (2):**
  - Cálculos y Consolidaciones (Agregaciones).
    - Consiste en calcular indicadores a partir de datos base. Pueden implicar consolidaciones.
  - Generación de datos históricos (historizacion).
    - Consiste en agregar marcas de tiempo a datos.
  - Generación de versiones.
    - Consiste en agregar atributos diferenciadores de diferentes versiones de un objeto base.
    - La historización permite hacer esto marcando la versión con un valor temporal.

Propiedades de los Sist. DW

- **Un Sistema de DW debería :**
  - Mantener una relación adecuada con BD Fuentes:
    - Acceso a BDs heterogeneas y multiplataforma.
    - Independiente de los Sistemas de Producción.
  - Permitir acceso efectivo a usuarios finales:
    - Soportar múltiples tipos de usuarios.
    - Ofrecer Interfaces a usuario avanzadas.
  - Funcionar en arquitecturas de varios niveles.
  - Interactuar con ambientes de Metadata.

Acceso a BD Fuentes heterogeneas

- **BD Fuentes heterogeneas:**
  - Diferentes modelos de datos:
    - Relacional.
    - Archivos legados (legacy).
    - Geográficos.
    - Documentos electrónicos.
    - Fuentes externas de datos (P.ej: cotizaciones bolsa).
  - Diferentes formatos:
    - Diferentes modelizaciones de información similar.
      - Claves diferentes para los mismos objetos.

Independencia de Sist. de Producción

- **Relaciones SDW y Sist. Producción:**
  - Coordinación:
    - El DW se alimenta a partir de datos de Sists. Prod.
  - Independencia:
    - Razones de performance:
      - Un SDW "pesado" no debe acceder on-line a BD-Prod.
        - Recargaria el Sist. Prod.
        - La performance de los dos se degradaría.
    - Razones de independencia lógica.
      - Los SDW suelen ver los datos de producción con una perspectiva histórica.
        - No siempre es deseable una coordinación fuerte.

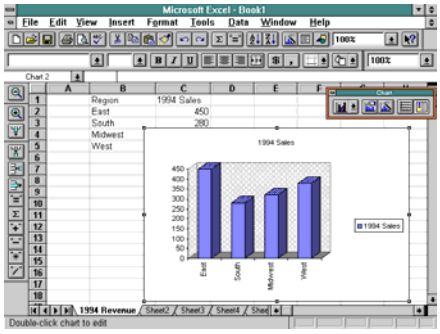
Soportar múltiples tipos de usuarios

- **Diferentes niveles jerárquicos:**
  - Directivos.
  - Gerentes de área.
  - Mandos técnicos.
- **Diferentes funciones:**
  - Planificación.
  - Control.
  - Análisis.

Interfaces avanzadas a usuario

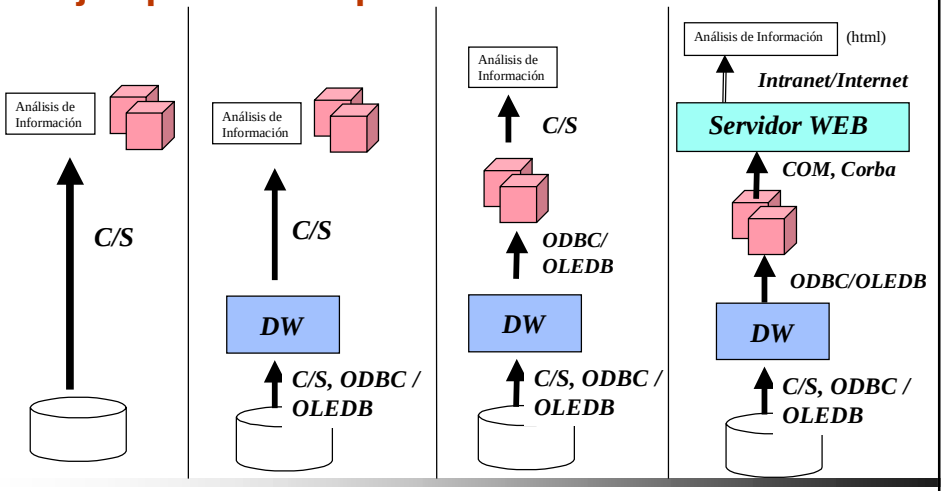
- **Interfaces a usuario especializadas.**
- **Por qué ?**
  - Optimizar el tiempo del usuario.
  - Principio:

    A cada tipo de usuario o aplicación se le ofrece la interfaz más adecuada.

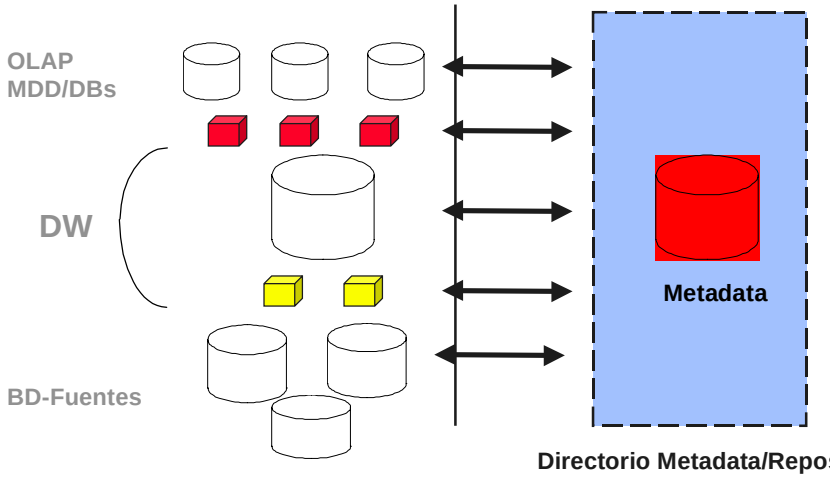


Arquitecturas en varios niveles

- **Ejemplos de Arquitecturas de SDW:**



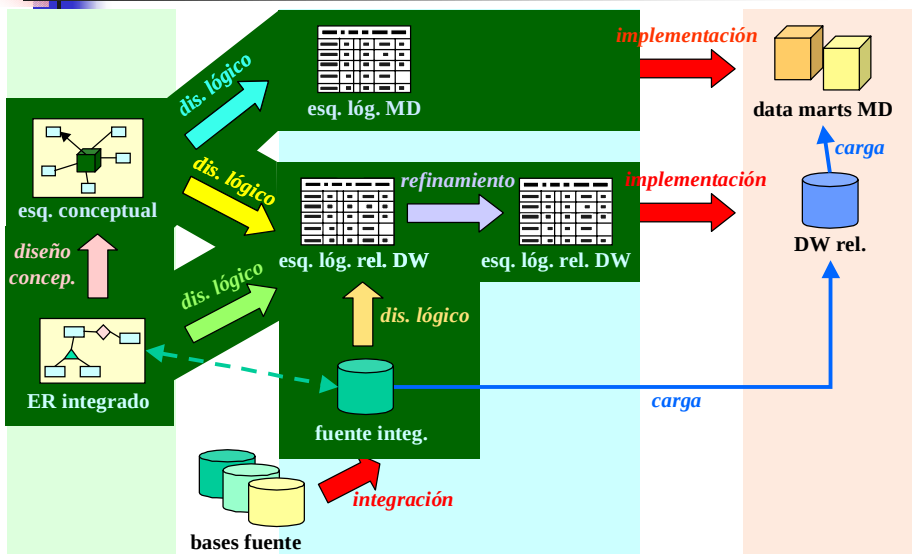
Interacción con Metadata



Desarrollo de Sistemas DW

- Fases:
  - Etapa 1: Descubrir y definir iniciativas.
  - Etapa 2: Determinación de la infraestructura.
  - Etapa 3: Desarrollo de aplicaciones.
- Componentes a desarrollar:
  - Adquisición de datos.
  - Almacenamiento del DW.
  - Mecanismos de acceso por parte de usuarios.

Proceso de Desarrollo



Factores de éxito

- **Un proyecto DW se considera exitoso si:**
  - Integra información heterogenea.
    - De diferentes tipos.
    - De diferentes orígenes.
  - Hace visible y manejable la información útil.
  - Incluye datos de calidad validada.
  - Ofrece acceso directo a usuarios.
  - El sistema se populariza.

Errores a evitar

- **Se debe evitar:**
  - Establecer expectativas demasiados altas.
  - Cargar el DW con todo lo disponible.
  - Elegir un DW manager sin orientación al negocio.
  - Diseñar el DW igual que un sistema de producción.
  - Ignorar fuentes de datos externas.
  - Ignorar la evolutividad del sistema.

Beneficios esperables

- **Se obtiene:**
  - Acceso interactivo e inmediato a información estratégica de un área de negocios.
  - Permite toma de decisiones basadas en datos objetivos.
  - Los beneficios aumentan :
    - cuanto más importantes son las decisiones.
    - cuanto más crítico es el factor tiempo.
  - Capitalización de datos en bases heterogeneas:
    - Archivos, dbf, etc.

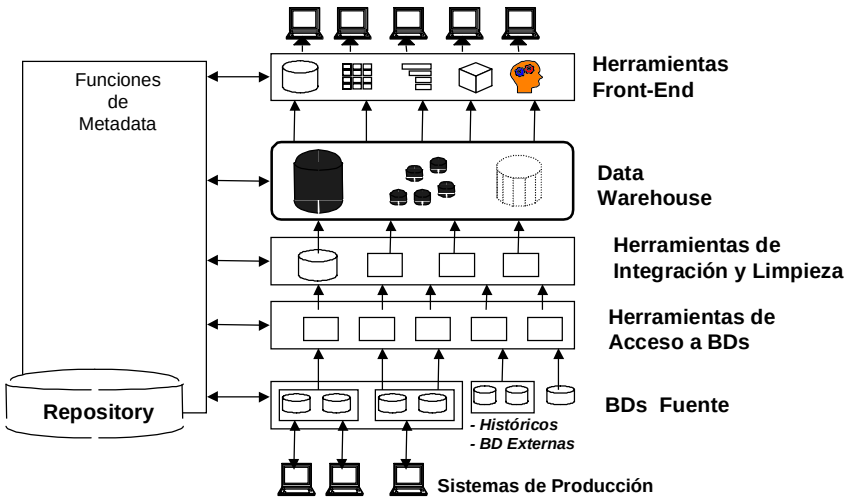
Características Técnicas

Características Técnicas

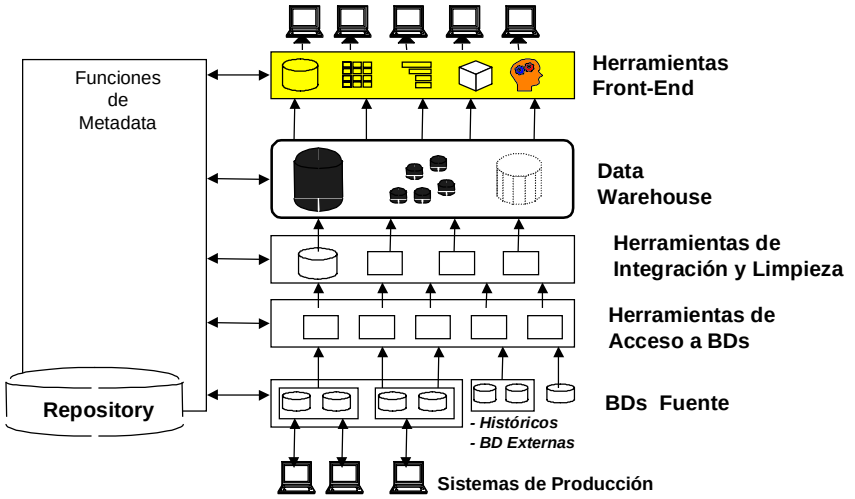
Temas:

- Herramientas Front-End.
- El Data Warehouse.
- Proceso de ETL.

Arquitectura Base



Herramientas Front-End



Herramientas Front-End

- **Introducción:**
  - Son herramientas usadas por el usuario para acceder a la información.
- **Objetivos:**
  - Ofrecer al usuario final mecanismos de acceso eficaces.
    - Mecanismos simples.
    - Mecanismos potentes.
  - Tener conección eficaz al DW.

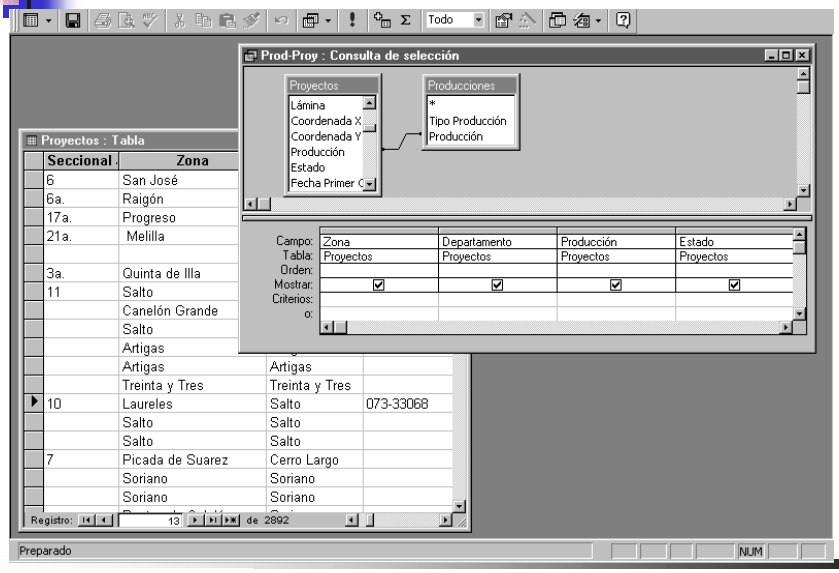
Herramientas Front-End

- **Diferentes tipos:**
  - BDs Escritorio.
  - Planillas Electrónicas.
  - Herramientas de consulta y reportes.
  - Herramientas OLAP.
  - Herramientas de Data Mining.

BD de Escritorio

- **Funcionalidades base:**
  - Estructuración Relacional de los datos:
    - Similar a la de los servidores relacionales.
    - Permite almacenar cantidades reducidas de datos manteniendo la estructura de la BD en el servidor.
  - Permite desarrollo de aplicaciones de apoyo a la decisión:
    - Sobre datos locales y a pequeña escala.
    - Personalizadas.
- **Ejemplos:**
  - Access, Paradox, dBase, FoxPro, Clipper.

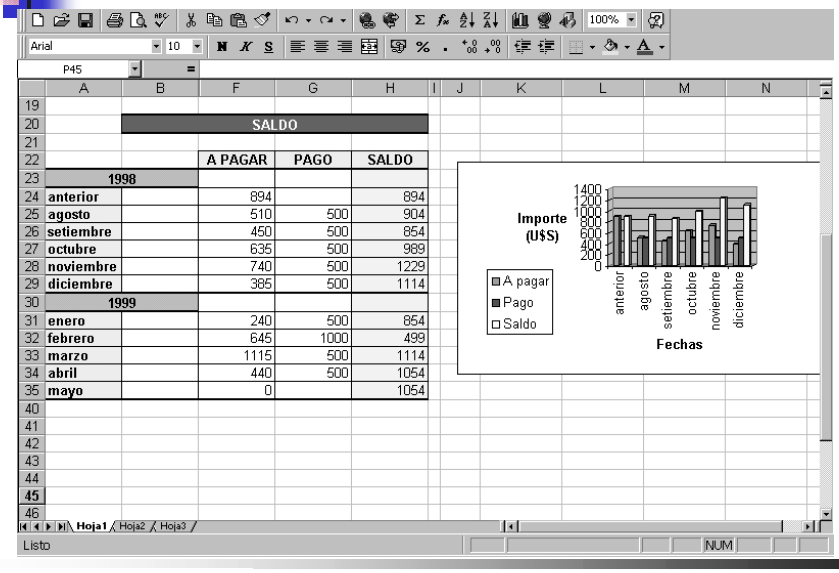
BD de Escritorio (II)



Planillas Electrónicas

- **Funcionalidades base:**
  - Estructuración de datos y operaciones cercanas a la visión del usuario.
  - Bien integradas con procesadores de texto.
  - En las últimas versiones se les integra más a bases de datos.
- **Ejemplos:**
  - Excel, 1-2-3, Quatro Pro

Planillas Electrónicas (II)

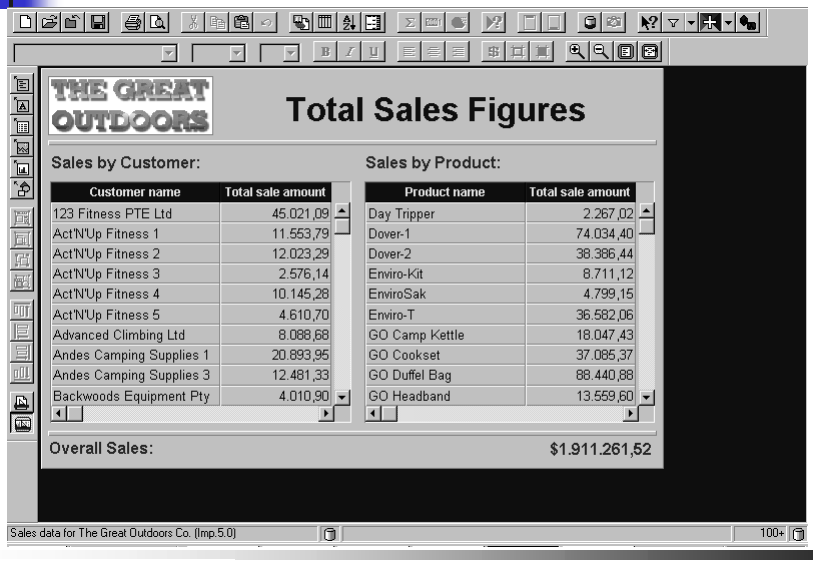


	A	B	F	G	H
20		SALDO			
22			A PAGAR	PAGO	SALDO
23	1998				
24	anterior		894		894
25	agosto		510	500	904
26	setiembre		450	500	854
27	octubre		635	500	989
28	noviembre		740	500	1229
29	diciembre		385	500	1114
30	1999				
31	enero		240	500	854
32	febrero		645	1000	499
33	marzo		1115	500	1114
34	abril		440	500	1054
35	mayo		0		1054

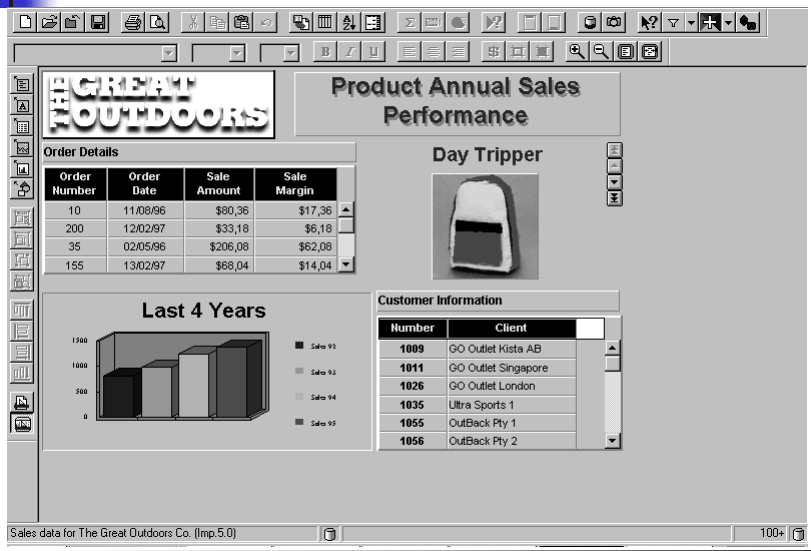
Herramientas Consultas y Reportes

- **Funcionalidades base:**
  - Construir facilmente consultas/reportes complejos.
  - Muy buenos para construir reportes no previstos.
  - Incorporan lenguajes para manejo de datos.
    - Incluyen funciones de todo tipo.
  - Ofrecen diferentes niveles de complejidad orientada a diferentes tipos de usuario:
    - Construcción de reporte complejo desde cero.
    - Construcción de reporte en base a templates.
    - Ejecución parametrizada de reportes.
    - Ejecución fija de reporte.

Herramientas Consultas y Reportes



Herramientas Consultas y Reportes



Herramientas Consultas y Reportes

- **Productos:**
  - Business Objects.
  - Andyne - GQL.
  - Seagate - Crystal Reports.
  - Soft AG - Esperant
  - Platinum Technology - Forrest & Trees, InfoQuery
  - Cognos - Impromptu
  - Oracle - Discoverer
  - IBM - Application System, QMF
  - Brio - BrioQuery
  - Informix - Viewpoint

Herramientas OLAP

- **Funcionalidades base:**
  - Permiten consultar datos :
    - Interactivamente y en forma eficiente.
    - Usando mecanismos comprensibles para usuarios.
      - Una consulta corresponde a cruzar dimensiones y elegir la medida en el cruzamiento.
  - Funcionalidades adicionales:
    - Rankings.
    - Visualización gráfica.
  - Funcionalidades de herramientas:
    - Integración con BDs Relacionales.
    - Integración con herramientas de escritorio.
    - Interfaces tipo API.

Herramientas OLAP

- **Introducción:**
  - Implementan Modelos Multidimensionales.
    - Los Modelos MD representan los datos como dimensiones en un hipercubo.
  - Tecnología en pleno desarrollo y expansión.
  - Diferentes alternativas tecnológicas:
    - ROLAP vs. MOLAP vs. HOLAP:
      - *ROLAPs*: actuan directamente sobre BD Rel.
      - *MOLAPs*: trabajan sobre almacenamiento especializado.
      - *HOLAP:* intentan aplicar ambas estrategias.

OLAP - Modelos Multidimensionales

- **Motivaciones:**
  - Representar los datos en forma más cercana a la intuición del usuario.
  - Resolver problemas planteados en sistemas relacionales.
- **Principios generales:**
  - La información se representa como:
    - cuadros de doble o triple entrada.
    - cubos de "n" dimensiones.
  - Una BD-MD incluye varias dimensiones.

OLAP - Modelos Multidimensionales

- ***Ejemplo: Análisis de ventas de autos***

Tabla:

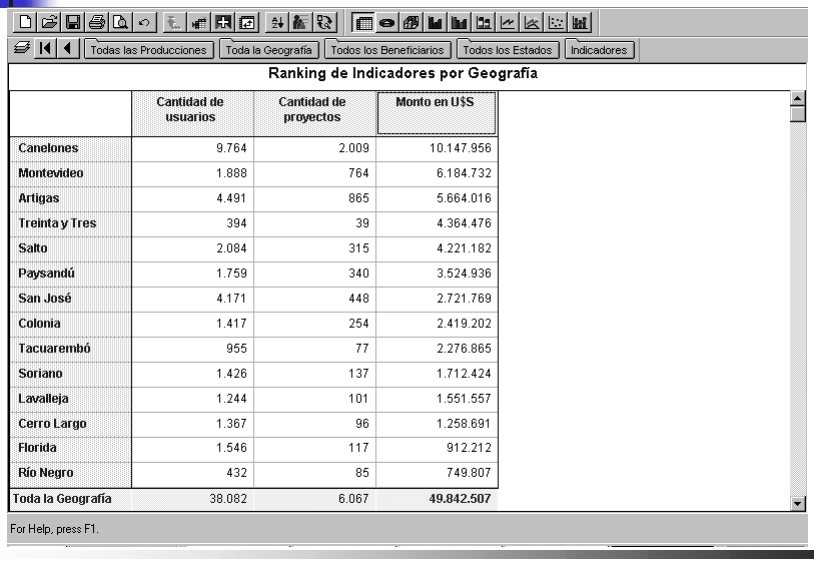
MODELO	COLOR	VOLUMEN-Ventas
MINI VAN	BLUE	6
MINI VAN	RED	5
MINI VAN	WHITE	4
SPORTS COUPE	BLUE	3
SPORTS COUPE	RED	5
SPORTS COUPE	WHITE	5
SEDAN	BLUE	4
SEDAN	RED	3
SEDAN	WHITE	2

Cuadro:

MODELO	Blue	Red	White
Mini Van	6	5	4
Coupe	3	5	5
Sedan	4	3	2

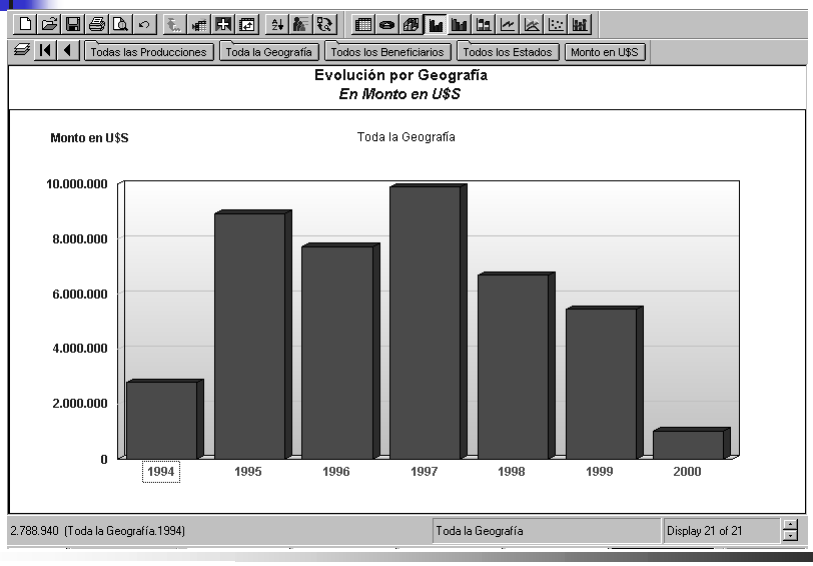
COLOR

Herramientas OLAP - Ejemplo



	Cantidad de usuarios	Cantidad de proyectos	Monto en U$S
Canelones	9.764	2.009	10.147.956
Montevideo	1.888	764	6.184.732
Artigas	4.491	865	5.664.016
Treinta y Tres	394	39	4.364.476
Salto	2.084	315	4.221.182
Paysandú	1.759	340	3.524.936
San José	4.171	448	2.721.769
Colonia	1.417	254	2.419.202
Tacuarembó	955	77	2.276.865
Soriano	1.426	137	1.712.424
Lavalleja	1.244	101	1.551.557
Cerro Largo	1.367	96	1.258.691
Florida	1.546	117	912.212
Río Negro	432	85	749.807
Toda la Geografía	38.082	6.067	49.842.507

Herramientas OLAP - Ejemplo



Herramientas OLAP

- **Cualidades de herramientas OLAP:**
  - Acceso a BDs.
    - Buena conectividad.
  - Valorización de datos.
    - Capacidad de cálculos.
    - Capacidad de operaciones de análisis.
    - Variedad de presentación de resultados.
  - Adaptación a diferentes tipos de usuario.
  - Control a operaciones del usuario.
    - Control de usuarios.
    - Evitar operaciones que "cuelguen" el sistema.

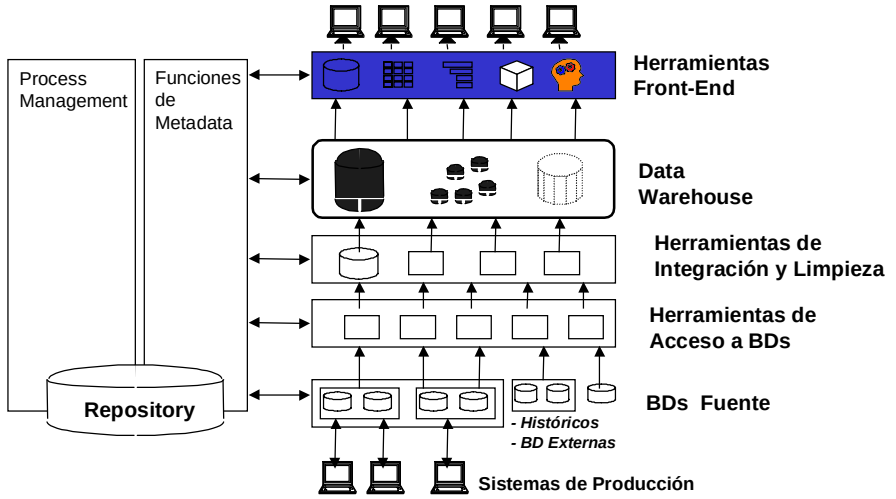
OLAP - Productos

- **Ejemplos:**
  - Hyperion Essbase.
  - MicroStrategy DSS
  - Oracle Express.
  - Pilot Lightship
  - Informix Metacube
  - Cognos PowerPlay
  - Microsoft SqlServer OLAP Services

Data Mining

- **Objetivos:**
  - Explorar BDs buscando relaciones desconocidas entre los datos.
- **Por ejemplo:**
  - Relaciones entre enfermedades y decesos.
    - Algunas candidatas a nuevas causas de decesos.
    - Otras podrían ser datos erróneos.
- **Qué incluye ?**
  - Un conjunto muy amplio y heterogéneo de técnicas y herramientas.

Data Mining en contexto DW



Data Mining en contexto DW

- **Diferencias con OLAP.**
  - Data Mining usa mecanismos de:
    - Descubrimiento de información, Pattern-matching,
    - Deducción de reglas, ... y otros

  para determinar relaciones claves entre los datos.
  - Los algoritmos de Data Mining pueden estudiar varias dimensiones de datos simultáneamente y descubrir los que tienen comportamiento especial.
  - La iniciativa es del algoritmo y no del usuario.

Aplicación : criterios generales

- **Etapas en uso de DM:**
  - Identificación del problema.
  - Definición de la *Estrategia* de resolución.
  - Aplicación de DM para generar un *Modelo*.
  - Manipulación del Modelo obtenido.
  - Medición de resultados obtenidos.
- **DM provee feedback a otros procesos:**
  - Construcción del DW.
    - Estructuración de los datos.
    - Definición de indicadores.
  - Estructuración/Análisis de datos OLAP post-DM.
    - En base a resultados obtenidos.

Estrategias para Data Mining

- **Introducción.**
  - Las estrategias para Data Mining corresponden al tipo de estudio que se desea realizar.
  - Las estrategias no son algoritmos en si mismas, sino formas de encarar el problema planteado.
  - Cada estrategia generará un Modelo, a través de la ejecución de un algoritmo.
- **Algunas estrategias.**
  - Clasificación, Clustering, Asociación, Optimización, Predicción.

Estrategias : Clasificación.

- **Objetivo:**
  - Clasificar registros según una variable objetivo, teniendo en cuenta valores de otros atributos.
- **Ejemplo:**
  - Se tiene una BD histórica con datos variados de clientes y un atributo de calificación de calidad (variable objetivo).
  - Dado un nuevo registro, del cual se desconoce su valor de variable objetivo, se quiere clasificar según los valores de los atributos.
- **Observaciones:**
  - Es de tipo *aprendizaje dirigido*, ya que se define la variable objetivo

Estrategias: Clustering

- **Objetivo.**
  - Generar grupos con registros según su “similaridad” en valores de atributos variados.
- **Ejemplo:**
  - Dada la BD del caso de Clasificación, generar grupos de clientes que tienen comportamiento similar sobre el conjunto de atributos.
- **Observaciones.**
  - Se trata de *aprendizaje no-dirigido.*
  - Se modela como un espacio n-dimensional de puntos, con una dimensión or atributo y un punto por registro.

Estrategias: Visualización

- **Objetivo.**
  - Representar situaciones de problema en forma visual, de forma de facilitar su análisis.
- **Ejemplo:**
  - Mostrar las distribuciones de ventas de productos en ciudades, teniendo en cuenta las características demográficas.
- **Observaciones.**
  - Se basa en técnicas de Interfase Hombre-Máquina y de comunicación de información en forma gráfica.

Estrategias: Asociación

- **Objetivo.**
  - Generar reglas de tipo **`IF A1,…An THEN B`**, donde A1 …,An son fenónemos en el problema.
- **Ejemplo:**
  - Se tiene una BD con tickets de supermercado. Y se quiere generar reglas que relacionen los productos comprados, hora de compra, dia, mes, y perfil de cliente.
  - `IF TipoCliente=1 AND CompraProd=p1 THEN CompraProd=P2;`
- **Observaciones.**
  - También se lo llama *Market Basket Análisis.*

Estrategias: Optimización.

- **Objetivo.**
  - Seleccionar una combinación de productos (o resultados) que mejor alcanza los objetivos de negocios.
- **Ejemplo:**
  - Lograr una combinación de cantidades producidas en diferentes productos que tienen sus costos y precios de venta.
- **Observaciones.**
  - Son casos de optimización lineal y no-lineal.

Estrategias: Estimación.

- **Objetivo.**
  - Realizar clasificaciones pero con una variable objetivo continua y no discreta.
- **Ejemplo:**
  - Para el caso de los clientes, tomar como variable la ganancia esperada que generan.

El Proceso de Data Mining.

- **Introducción.**
  - Aplicar Data Mining corresponde más a un proceso que a una operación individual.
- **Pasos:**
  - Preparación de datos.
  - Definición de estudio.
  - Construcción de Modelo.
  - Entender y aplicar el Modelo.

Proceso: Preparación de datos.

- **Definición.**
  - Consiste en la generación de una base de datos sobre la cual se pueda aplicar el estudio deseado.
- **Aspectos a resolver:**
  - Limpieza de datos.
  - Valores nulos.
  - Derivación de datos.
  - Integración (merge) de datos.

Proceso: Definición de estudio

- **Definición.**
  - Consiste en definir los resultados a obtener, el tipo de estrategia y el alcance del estudio.

- **Aspectos a resolver:**
  - Definir los límites.
    - De qué se parte y qué se quiere obtener.
  - Elegir el tipo de estudio, incluyendo la estrategia.
  - Especificar los elementos a analizar.
    - Datos relevantes, valores resultados.
  - Definición de la muestra.
    - ¿ Como tomar una muestra representativa ?

Proceso: Construcción de Modelo

- **Definición.**
  - Consiste en construir un modelo abstracto que representa el problema y que manipulándolo se tratan de resolver los requerimientos.
- **Aspectos a resolver:**
  - Precisión (accuracy).
  - Comprensibilidad (understandability).
    - Qué entradas afectan la salida.
    - Por qué tiene éxito o falla.
  - Performance.
    - Qué tan rápido genera el modelo.
    - Qué tan rápido se obtienen las conclusiones deseadas.

Proceso: Entender y aplicar el Modelo

- **Definición.**
  - Consiste en asociar el modelo resultante al problema real de forma de comprenderlo.

- **Implica:**
  - Validar los resultados del modelo.
  - Extraer elementos relevantes y descartar las distorsiones.
  - Concluir qué fenómeno ocurre u ocurrirá.

Modelos y sus características

- **Modelos de Data Mining:**
  - Un *Modelo* es una representación de un problema que, instanciado con valores, genera resultados.
  - Por ejemplo: se tienen modelos predictivos, de clasificación, series de tiempo, clustering, etc.
  - Los modelos poseen ciertos atributos:
    - Underfitting y Overfitting.
    - Dirigido o no dirigido.
    - Explicabilidad de resultado.
    - Facilidad de aplicación.

Modelos y sus características

- **Underfitting y Overfitting:**
  - ***Overfitting:*** más info que la deseable.
    - Todos los elementos se comportan como el set de entrenamiento (memorización del training set).
    - Se tiene información redundante dentro de los campos considerados, obteniendo un modelo trivial.
  - ***Underfitting:*** menos info que la deseable.
    - No se llegan a obtener patrones de interés sobre los datos (e.g. con bajo impacto predictivo).
    - Puede ser consecuencia de la desactualización de modelos en el tiempo.

Modelos y sus características

- **Dirigidos vs. No dirigidos.**
  - ***Dirigidos:*** la forma de la salida del modelo se especifica previo a su construcción.
    - El modelo se entrena sobre casos donde la salida está determinada (e.g. red neuronal con salida a estimar conocida).
  - ***No dirigidos:*** el propio modelo determina cuál será su salida.
    - Por ejemplo: estrategia de clustering donde el modelo son los clusters identificados.

Modelos y sus características

- **Explicabilidad.**
  - Resulta clarificante de interés conocer las razones que determinan los resultados.
  - Diferentes técnicas aportan distintos niveles de explicabilidad sobre sus resultados.
- **Facilidad de aplicación.**
  - Está asociado a la facilidad de uso, de comprensión de los resultados, de claridad de los resultados, de practicidad y conexión a bases de datos.

Algoritmos de Data Mining.

- **Introducción.**
  - El *Modelo* resultante del proceso de Data Mining es generado por algoritmos a través de productos de software.
- **Tipos de algoritmos.**
  - Árboles de Decisión.
  - Algoritmos Genéticos.
  - Redes Neuronales.
  - Estadísticos.
  - Algoritmos avanzados de asociación.
  - Algoritmos para Optimización.

Técnicas para Data Mining

- **La elección de una combinación particular de técnicas dependerá**
  - Problema a resolver / análisis DM.
  - naturaleza de los datos disponibles.
  - Características conocidas sobre los tipos de *Modelos* generados por las técnicas:
    - Underfitting & Overfitting
    - Dirigidos vs No dirigidos
    - Explicabilidad
    - Facilidad de aplicación

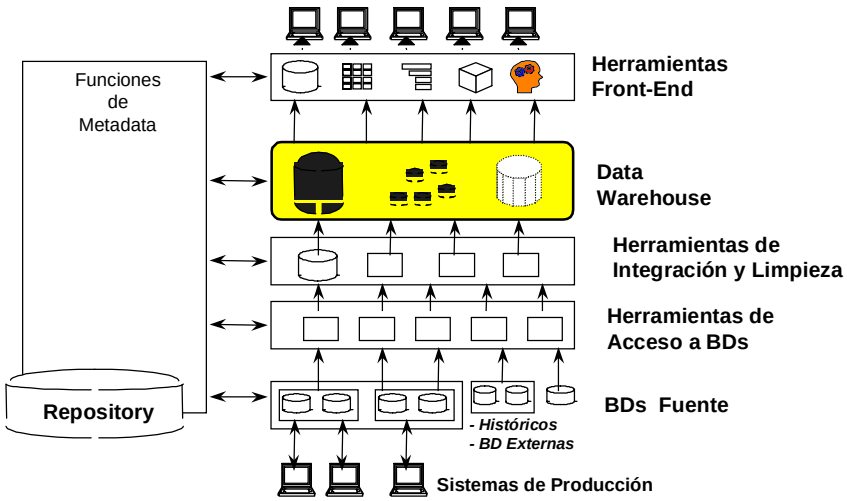
Data Mining

- **Sintesis.**
  - Area con fuertes componentes matemáticas.
  - Nuevos productos:
    - Accesibles en precio.
    - Explotables por usuarios no expertos.
  - Se prevee un gran impacto:
    - en el diseño de Sistemas DW.
    - en la explotación de Sistemas DW.
  - Todavía trabajo por hacer en la integración a los Sistemas DW.

Características Técnicas

Temas:

- Herramientas Front-End.
- ***El Data Warehouse.***
- Proceso de ETL.



El Data Warehouse

- **Aspectos técnicos relacionados.**
  - Tipo/s de DBMS.
    - Relacional, Multidimensional, o ambos.
  - Concepción del DW.
  - Diseño e implementación de la carga.
  - Administración del DW:
    - Cómo describir y documentar los datos ?
    - Qué información hay que monitorear ?
    - Cómo organizar y realizar la administración del DW ?
    - Mediante qué tipo de herramientas ?

DBMS para el Data Warehouse

- **DBMSs Relacionales:**
  - Solución "universal".
  - Soportan el grueso de las aplicaciones DW.
  - Dificultades para resolver eficientemente consultas dimensionales.
- **DBMSs Multi-Dimensionales:**
  - Representan los datos del problema en términos de dimensiones.
  - Estructuras de almacenamiento están diseñadas para optimizar consultas dimensionales.

DBMSs Relacionales

- **Mecanismos que mejoran performance.**
  - Arquitecturas paralelas.
    - Solución de hardware.
    - Multiplicar recursos de procesamiento.
  - Mecanismos del DBMS.
    - Independientes del hardware.
    - Indices binarios.
      - Nuevas estructuras de datos para acceso más eficiente.
    - Algoritmos de StarJoin.
      - Nuevos tipos de algoritmos orientados a consultas dimensionales sobre BDs Relacionales

Diseño del Data Warehouse

- **Elementos base:**
  - Las operaciones principales son consultas.
  - La carga/actualización no es transaccional.
  - Importancia de la calidad y facilidad de acceso.
- **Por lo tanto …**
  - El DW se construye en capas asignando propiedades a las tablas de cada una.
  - Se suele desnormalizar y materializar cálculos.
- **En cuanto a complejidad …**
  - El diseño del DW y programación de la carga constituyen las tareas más costosos y complejas.

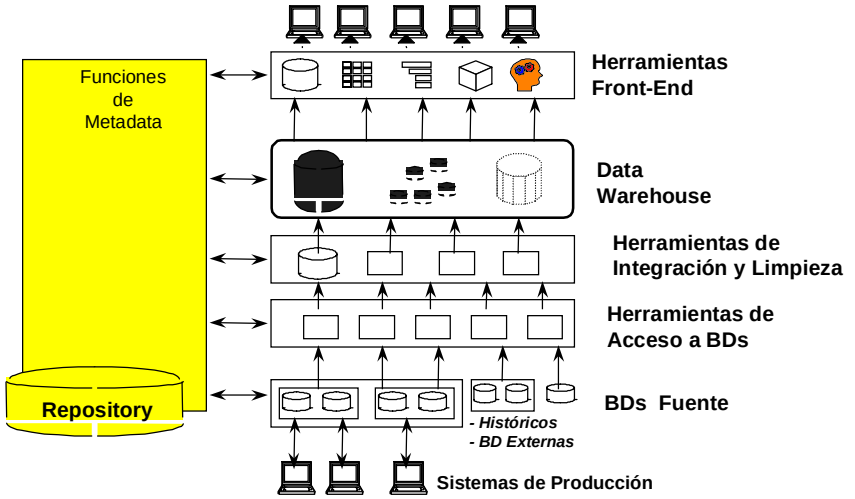
Administración del DW

- **Aspectos principales a prever.**
  - Gestión de la Metainformación.
  - Monitoreo del DW.
- **Resulta importante:**
  - Prever los recursos necesarios.
    - Personas, Herramientas, Datos.
- **Herramientas especializadas:**
    - *Data Warehouse Repository*.

Administración del DW: Metainformación

- **La metainformación.**
  - Información sobre los datos en el DW.
    - Constituye un elemento clave para los Sist. DW.
      - Para el mantenimiento del sistema.
      - Control y verificación de la calidad de los datos.
    - Enriquece los datos almacenados en el DW.
      - Asociandolos a objetos del negocio real.
  - Tipos de metainformación:
    - Propiedades (o atributos).
    - Relaciones con otros objetos.

Gestión de la Metainformación



La Metainformación

- **Items de metainformación:**
  - Semántica (de datos en el DW).
    - Qué significa ese dato ?
    - Con qué temática se relaciona el ítem ?
  - Origen.
    - Cúal es su origen ? (BD, cálculo, ...)
  - Reglas de cálculo.
    - Cómo se calcula el ítem de datos ?
  - Reglas de agregación.
    - Cuál es el conjunto de datos fuente ?

La Metainformación

- **Items de metainformación (cont.):**
  - Almacenamiento, formato.
    - Cómo se almacena y con qué formato ?
  - Uso.
    - Qué programas lo usan ?
  - Datos fuentes.
    - De qué tablas se extrae el ítem ?
  - Carga.
    - Con qué frecuencia se cargan los datos del DW ?
    - Cómo se realiza las historización.

La Metainformación

- **Gestión de la metainformación.**
  - Es un problema en si mismo,
    - además de la administración del DW.
  - Concierne funciones de:
    - Modelado de datos.
    - Almacenamiento.
    - Acceso.
  - Por lo que:
    - Resulta interesante contar con herramientas especializadas en Gestión de Metainformación.
      - El ***Data Warehouse Respository***.

Data Warehouse Repository

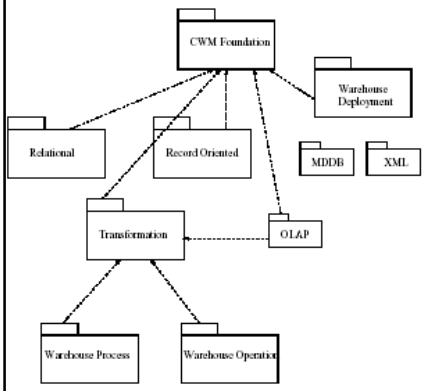
- **Qué es:**
  - Es un sistema que almacena y soporta operaciones sobre la Metadata.
  - Puede ser usado en diferentes contextos:
    - Sistema de Data Warehousing.
    - Para organizar la Metabase Corporativa de una org.
    - Como base para herramientas CASE.
- **Vocación (función principal):**
  - Federar la metainformación disponible sobre los diferentes tipos de datos.

Modelos de Metadata

- **OIM (Metadata Coalition) (1999-2001).**
  - Objetivo: soportar a interoperabilidad entre herramientas heterogeneas a nivel empresarial.
  - Diseñado para acompañar todas las fases del desarrollo de Sistemas de Información.
- **CWM (Object Management Group) (2000- ...)**
  - Objetivo: soportar interoperabilidad en herramientas y sistemas de Data Warehousing.
  - Especializado en soporte a Data Warehouse y Business Intelligence.

Modelos de Metadata: CWM

- **Estructura:**
  - Packages que cubren todas las áreas.



The CWM Metamodel

Management	Warehouse Process			Warehouse Operation		
Analysis	Transformation		OLAP	Data Mining	Information Visualization	Business Nomenclature
Resource	Object Model	Relational	Record	Multidimensional		XML
Foundation	Business Information	Data Types	Expression	Keys and Indexes	Type Mapping	Software Deployment
	Object Model					

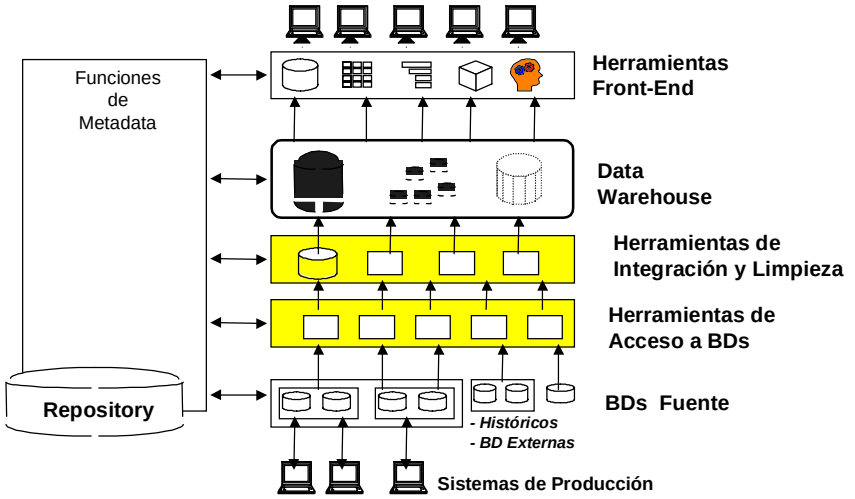
Características Técnicas

Características Técnicas

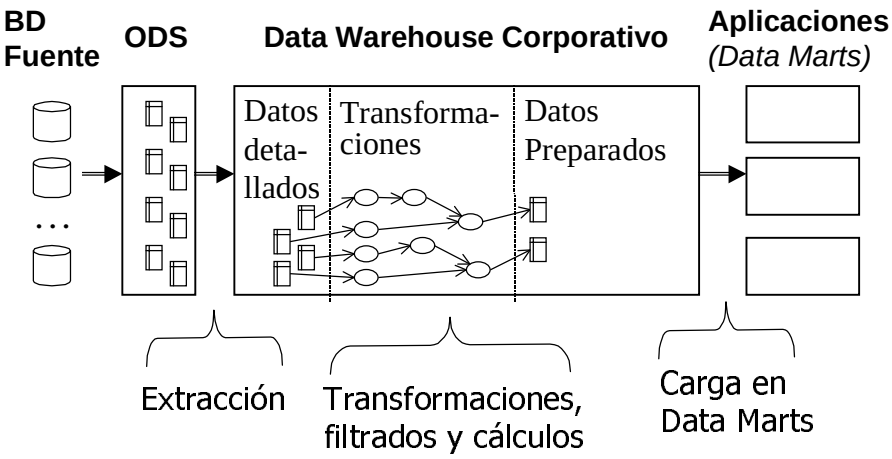
Temas:

- Herramientas Front-End.
- El Data Warehouse.
- ***Proceso de ETL.***

Integración y Consolidación de datos



Proceso ETL en el DW



Tecn. de Integración y Consolidación

- **Qué operaciones conciernen ?**
  - La *identificación* de los datos fuente que interesan.
  - La *Extracción* de datos de las BDs fuente.
  - La *Integración* de lo extraído.
  - La *Historización* y *Consolidación* de los datos en valores *agregados*.
  - El *Control de Calidad* de los datos que se insertan en el DW.

Realización de Integración y Consolidación

- **Problemas a resolver:**
  - Diseño de la configuración:
    - Qué operaciones se realizan en qué momento.
    - Algunas pueden estar particionados en diferentes etapas.
  - Aspectos técnicos :
    - Diseñar procesos de extracción, integración, etc.
  - Definición de frecuencia de operaciones.
    - Grado de sincronización del DW con BDs fuente.
    - Coordinar esta frecuencia con la lógica de las consolidaciones.

Resumen (1)

- Los Sistemas de DW son una pieza clave en el proceso de toma de decisiones:
  - Acercan la información al usuario.
- Los Sistemas DW permiten revalorizar los datos en la empresa:
  - Integran datos en diferentes formatos.
- Los Sistemas DW no son productos monolíticos sino composición de diferentes soluciones técnicas.
  - Construcción del Diccionario de Datos, Diseño de Base de Datos, Conectividad, Control de calidad de datos, etc .

Resumen (2)

- **Herramientas Front-End:**
  - Tipos muy diferentes:
    - Desde planillas … OLAP … Data Mining.
    - OLAP es la actualmente dominante.
    - Data Mining es la emergente.
  - Enfasis :
    - Usabilidad por parte de usuarios finales.
    - Conección a la Arquitectura del SDW.
- **Proceso de Carga y Actualización.**
  - Corresponde a la actividad más costosa.