Almacenes de datos

Juan Carlos Trujillo Mondéjar

IWAD: Ingeniería del Web y Almacenes de Datos

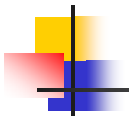
Dpto. Lenguajes y Sistemas Informáticos
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Indice

- Introducción
- Diseño de almacenes de datos: visión práctica
- Integración de fuentes: Procesos ETL
- Diseño de almacenes de datos: UML
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

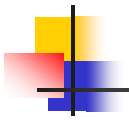


Indice

- Introducción
  - Sistemas de apoyo a la decisión
  - El almacén de datos (AD)
  - Diferencia sistemas transaccionales y de AD
  - Arquitectura de almacenes de datos
- Diseño de almacenes de datos: visión práctica
- Integración de fuentes: Procesos ETL
- Diseño de almacenes de datos: UML
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Almacenes de Datos



Indice

- Introducción
  - Sistemas de apoyo a la decisión
  - El almacén de datos (AD)
  - Diferencia sistemas transaccionales y de AD
  - Arquitectura de almacenes de datos
- Diseño de almacenes de datos: visión práctica
- Integración de fuentes: Procesos ETL
- Diseño de almacenes de datos: UML
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Introducción
Sistemas de apoyo a la decisión

- Empresas en la actualidad
  - Sistemas orientados a los procesos diarios de la empresa
    - Sistemas de Procesamiento Transaccional en Línea (*On-Line Transactional Processing, OLTP*)
    - Compras de productos, ventas, pedidos, gestión de clientes, ..
    - Optimizados para la edición e inserción de datos
  - Aproximadamente el 90% de SGBD son relacionales
    - SGBD eficientes, robustos, etc.
  - Datos históricos → almacenamientos externos

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Introducción
Sistemas de apoyo a la decisión

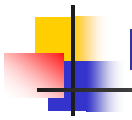
- Entornos económicos altamente competitivos
  - Empresas necesitan adoptar decisiones estratégicas
    - ¿ Qué tipo de cliente me ha estado comprando el BMW 320i durante los últimos 10 años ?
    - ¿ Ha variado un cliente sus gustos de compra de vehículos? ¿Ha estado comprando el mismo vehiculo de soltero que de casado?
    - ¿ Qué descuento deberiamos ofrecer para incrementar significativamente las ventas ?
- Sistemas de apoyo a la decisión

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Introducción
Sistemas de apoyo a la decisión

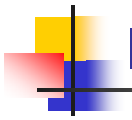
- ¿ Son válidos los sistemas OLTP para tales entornos ?
  - Algunos problemas
    - Gran volumen de datos históricos no disponibles en sistemas diarios OLTP
    - Normalmente en distintas fuentes de datos
      - Proveedores, Clientes, componentes, productos defectuosos, etc.
    - Los directivos/analistas no saben manejar tales sistemas

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es



Indice

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es



Indice

- Introducción
  - Sistemas de apoyo a la decisión
  - El almacén de datos (AD)
  - Diferencia sistemas transaccionales y de AD
  - Arquitectura de almacenes de datos
- Diseño de almacenes de datos: visión práctica
- Integración de fuentes: Procesos ETL
- Diseño de almacenes de datos: UML
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Introducción
El almacén de datos

- El Almacén de datos (Data Warehouse, DW)
  - Repositorio de datos históricos para ser utilizados por los Sistemas de Apoyo a la Decisión
  - Son sistemas eminéntemente de consulta enfocados a extraer conocimiento de los datos históricos almacenados
  - El análisis de los datos → On-Line Analytical Processing (OLAP)
    - Utilizan el modelado multidimensional (cubos, hipercubos, etc)

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Introducción
El almacén de datos

- Definición según W. Inmon (1992)
  - "Una colección de datos orientados por tema, variables en el tiempo y no volátiles que se emplea como apoyo a la toma de decisiones estratégicas"

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Introducción
El almacén de datos

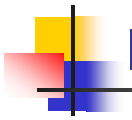
- Orientados por tema
  - El diseño enfocado a responder eficientementa a consultas estratégicas
    - Actividades de interés: compra, ventas, alquileres,...
    - Contexto de análisis: clientes, vendedores, productos, etc...
  - El modelado Multidimensional (primera aproximación)
    - Hechos → actividades de interés
    - Dimensiones → contexto de análisis

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Introducción
El almacén de datos

- Integrados
  - Datos integrados de distintas fuentes de datos operacionales
- Variables en el tiempo
  - Datos relativos a un periodo de tiempo y se incrementan periódicamente
- No volátiles
  - Los datos almacenados *normalmente* no se modifican ni actualizan nunca (casi nunca), sólo se insertan nuevos datos

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es



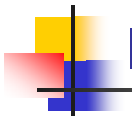
Indice

- Introducción
  - Sistemas de apoyo a la decisión
  - El almacén de datos (AD)
  - Diferencia sistemas transaccionales y de AD
  - Arquitectura de almacenes de datos
- Diseño de almacenes de datos: visión práctica
- Integración de fuentes: Procesos ETL

Diseño de almacenes de datos: UML

- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es



Indice

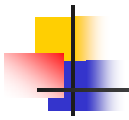
- Introducción
  - Sistemas de apoyo a la decisión
  - El almacén de datos (AD)
  - Diferencia sistemas transaccionales y de AD
  - Arquitectura de almacenes de datos
- Diseño de almacenes de datos: visión práctica
- Integración de fuentes: Procesos ETL
- Diseño de almacenes de datos: UML
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Introducción
Diferencia sistemas transaccionales y de AD

	OLTP	AD / OLAP
Usuario	Profesional de TI	Analista de Información
Función	Operaciones diarias	Apoyo a la decisión
Diseño de BD	Orientada a la aplicación (Basado en EE-R)	Orientado al tema/negocio (Multidimensional, ej. estrella)
Datos	Actuales, Aislados	Históricos, Consolidados
Vistas	Detallados, Planos, Relac.	Agregados, Multidimensional
Destino / utilización	Estructuradas, repetitivas	Ad-Hoc
Unidades de trabajo	Transacciones simples	Consultas complejas
Acceso	Lectura/escritura	Lectura mayoritariamente
# Registros accedidos	Decenas	Millones
# Usuarios	"Miles"	"Centenares"
Tamaño de la BD	100 MB-GB	100 GB-TB
Medidas de rendimiento	Cantidad de transacciones	Cantidad de consultas, Respuesta

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

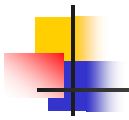


Indice

- Introducción
  - Sistemas de apoyo a la decisión
  - El almacén de datos (AD)
  - Diferencia sistemas transaccionales y de AD
  - Arquitectura de almacenes de datos
- Diseño de almacenes de datos: visión práctica
- Integración de fuentes: Procesos ETL
- Diseño de almacenes de datos: UML
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Almacenes de Datos

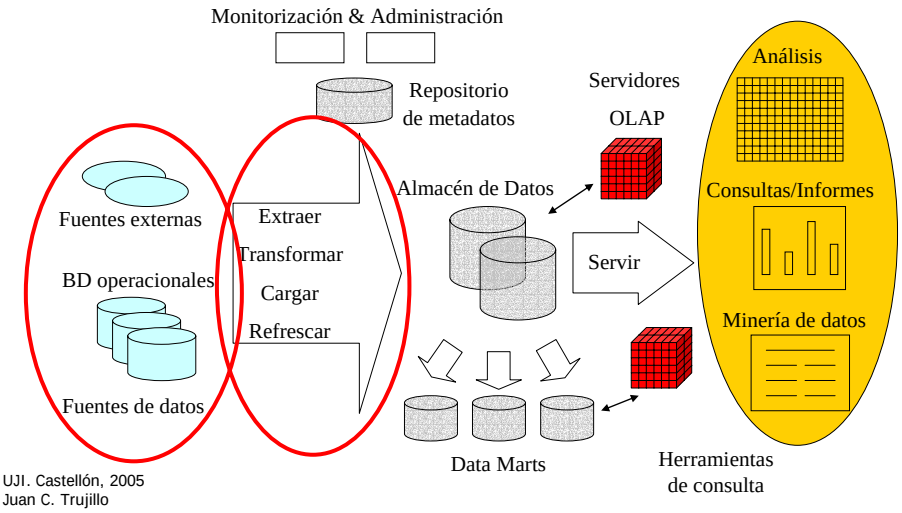


Indice

- Introducción
  - Sistemas de apoyo a la decisión
  - El almacén de datos (AD)
  - Diferencia sistemas transaccionales y de AD
  - Arquitectura de almacenes de datos
- Diseño de almacenes de datos: visión práctica
- Integración de fuentes: Procesos ETL
- Diseño de almacenes de datos: UML
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Introducción
Arquitectura de almacén de datos



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

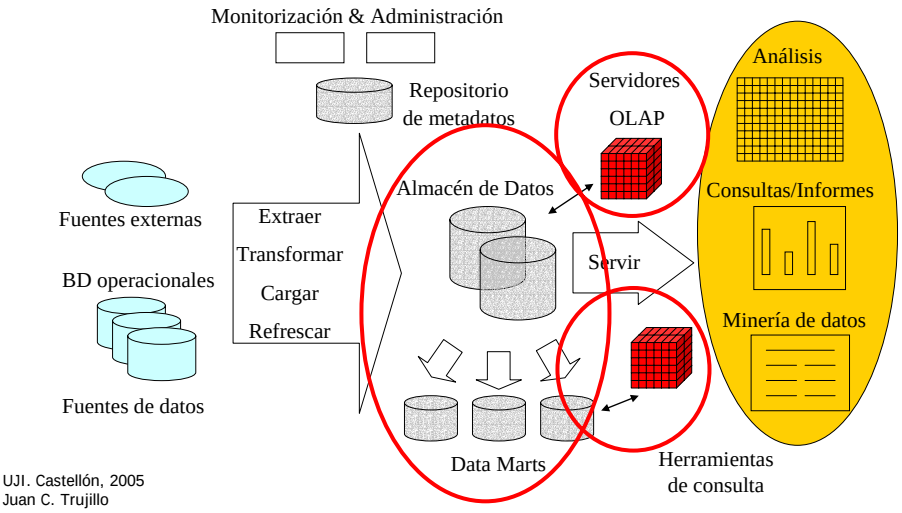
Introducción
Arquitectura de almacén de datos. Procesos ETL

- Procesos para poblar de datos el almacén (ETL)
  - Extracción (Extraction)
  - Limpieza (Cleaning) y Transformación (Transformation)
  - Carga (Loading) y Refresco

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Introducción

Arquitectura de almacén de datos



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Introducción

Arquitectura de almacén de datos. Servidores OLAP

- Servidores de consulta ROLAP
  - Utilizan tecnología Relacional (Relational OLAP)
  - Utilizan extendiones del SQL estándar para soportar el acceso multidimensional a los datos
  - Métodos de implementación adecuados para representar los datos multidimensionales en tecnología relacional
  - Ventaja: Basado en un estándar

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Introducción

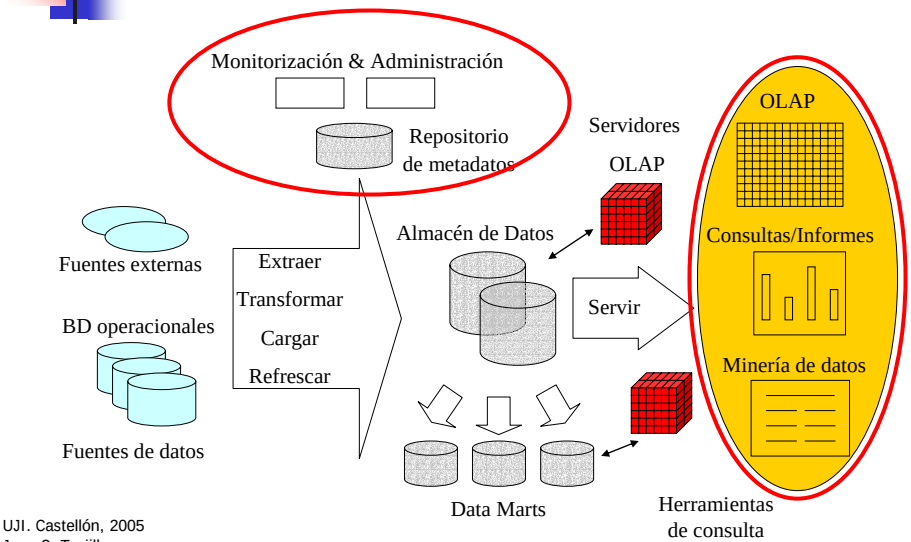
Arquitectura de almacén de datos. Servidores OLAP

- Servidores de consulta MOLAP
  - Utilizan tecnología Multidimensional (Multidimensional OLAP)
  - Los datos están almacenados directamente en matrices
  - Operaciones de consulta están implementadas directamente sobre estas matrices
    - No están basados en SQL estándar
  - Ventaja: Suelen ser más rapidos que los ROLAP
  - Inconveniente: no basados en un estándar

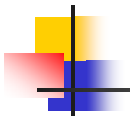
UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Introducción

Arquitectura de almacén de datos



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es



Indice

- Introducción
  - Sistemas de apoyo a la decisión
  - El almacén de datos (AD)
  - Diferencia sistemas transaccionales y de AD
  - Arquitectura de almacenes de datos
- Diseño de almacenes de datos: visión práctica
- Integración de fuentes: Procesos ETL
- Diseño de almacenes de datos: UML
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Almacenes de Datos

Indice

- Introducción
- Diseño de almacenes de datos: visión práctica
- Integración de fuentes: Procesos ETL
- Diseño de almacenes de datos: UML
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Indice

- Introducción
- Diseño de almacenes de datos: visión práctica
  - Bases de datos transaccionales vs. almacenes de datos
  - Modelado Multidimensional (MD)
  - Esquema estrella y variantes (en SGBDR)
  - Dimensión tiempo
  - Algunos datos del mercado
- Integración de fuentes: Procesos ETL
- Diseño de almacenes de datos: UML
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

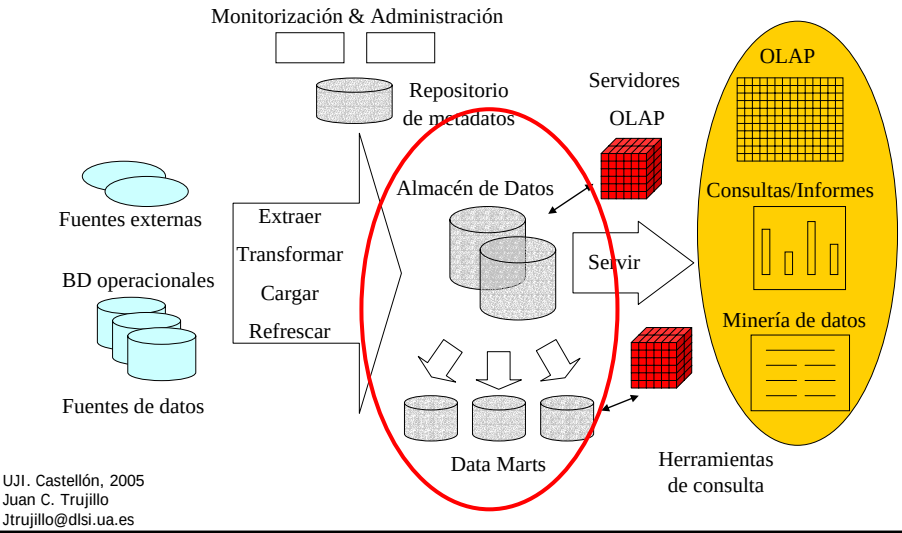
Almacenes de Datos

Indice

- Introducción
- Diseño de almacenes de datos: visión práctica
  - Bases de datos transaccionales vs. almacenes de datos
  - Modelado Multidimensional (MD)
  - Esquema estrella y variantes (en SGBDR)
  - Dimensión tiempo
  - Algunos datos del mercado
- Integración de fuentes: Procesos ETL
- Diseño de almacenes de datos: UML
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica
Bases de datos transaccionales vs. Almacenes de datos



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Almacenes de Datos

Diseño de AD: visión práctica
Bases de datos transaccionales vs. Almacenes de datos

- Bases de datos transaccionales (relacionales)
  - Normalización
    - Optimizadas para edición e inserción de datos
- Diseño conceptual → Modelo EER
- Diseño lógico → Modelo Lógico Relacional
- Diseño físico → Modelo físico (Indices, particionamiento,...)

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica

Bases de datos transaccionales vs. Almacenes de datos

- Almacenes de datos
  - Des - normalización
    - Optimizadas para consultas complejas
      - Reduce número de objetos y de relaciones entre éstos
    - Fácil interpretación por el analista de la información
- Diseño conceptual → Modelado MD (intuitivo)
- Diseño lógico → Esquema estrella (si SGBDR)
- Diseño físico → Modelo físico (Indices, particionamiento,...)

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Indice

- Introducción
- Diseño de almacenes de datos: visión práctica
  - Bases de datos transaccionales vs. almacenes de datos
  - Modelado Multidimensional (MD)
  - Esquema estrella y variantes (en SGBDR)
  - Dimensión tiempo
  - Algunos datos del mercado
- Integración de fuentes: Procesos ETL
- Diseño de almacenes de datos: UML
- Conclusiones

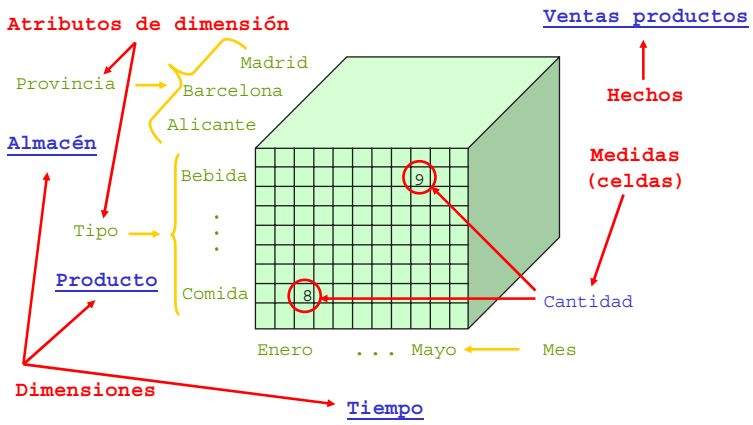
UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Indice

- Introducción
- Diseño de almacenes de datos: visión práctica
  - Bases de datos transaccionales vs. almacenes de datos
  - Modelado Multidimensional (MD)
  - Esquema estrella y variantes (en SGBDR)
  - Dimensión tiempo
  - Algunos datos del mercado
- Integración de fuentes: Procesos ETL
- Diseño de almacenes de datos: UML
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica
Modelado Multidimensional (MD)



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica
Modelado Multidimensional (MD)

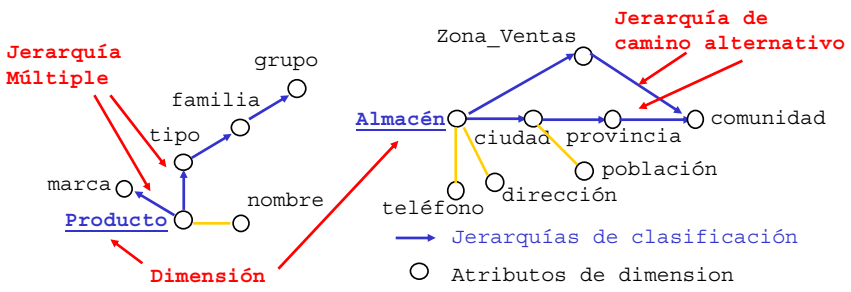
- Tablas multidimensionales

Ventas			Producto.Grupo = "Supermercado"			
			Comida		Bebida	
			Cong	Fresco	Refresco	Alcohol
Almacén. comunidad = "Comunidad Valenciana"	Alicante	Albatera	100	200	300	400
		Elche	500	600	700	800
	Valencia	Burjasot	900	1000	1100	1200
		Cullera	1300	1400	1500	1600

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica
Modelado Multidimensional (MD). Dimensiones

- Normalmente se representan mediante Grafos Acíclicos dirigidos (G.A.D.)



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica
Modelado Multidimensional (MD). Hechos

- Atributos de hecho o medidas
  - Atómicos
    - Ej. Cantidad vendida, precio, etc.
  - Derivados
    - Utilizan una fórmula para calcularlos
    - Ej. Precio_total = precio * cantidad_vendida

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica
Modelado Multidimensional (MD). Hechos

- Aditividad
  - Conjunto de operadores de agregación (SUM, AVG, etc.) que se pueden aplicar para agregar los valores de medidas a lo largo de las jerarquías de clasificación *(Kimball, 1996)*
  - Es aditiva → SUM sobre todas las dimensiones
  - Semi-aditiva → SUM sólo sobre algunas dimensiones
  - No aditiva → SUM sobre ninguna dimensión

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica
Modelado Multidimensional (MD). Consultas

- Definición de requerimientos iniciales de usuario
  - Están basados en jerarquías definidas en Dimensiones

Cantidad vendida de productos comestibles agrupados por su familia y tipo, de almacenes de la comunidad valenciana y, agrupados por la provincia y ciudad donde se vendieron

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica
Modelado Multidimensional (MD). Operaciones OLAP

- Operaciones de consulta (OLAP)
  - Roll-up
    - Agregar valores de medidas a lo largo de jerarquías de clasificación

Ventas'		Producto.Grupo = "alimentación"	
		Comida	Bebida
Almacén. comunidad = "Comunidad Valenciana"	Alicante	1400	2200
	Valencia	4600	5400

Grupo
Familia
Tipo

Provincia Ciudad

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica
Modelado Multidimensional (MD). Operaciones OLAP

- Operaciones de consulta (OLAP)
  - Drill-down
    - Desagregar valores de medidas a lo largo de jerarquías de clasificación

Ventas			Producto.Grupo = "alimentación"			
			Comida		Bebida	
			Cong.	Fresca	Refresco	Alcohol
Almacén. Comunidad= "Comunidad Valenciana"	Alicante	Albatera	100	200	300	400
		Elche	500	600	700	800
	Valencia	Burjasot	900	1000	1100	1200
		Cullera	1300	1400	1500	1600

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica
Modelado Multidimensional (MD). Operaciones OLAP

- Operaciones de consulta (OLAP)
  - Drill-accross
    - Consultar medidas de varios hechos en el mismo cubo
      - Ej. Que en la tabla MD analizaramos el ratio de ventas respecto de compras.
        - 1000 / 400

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica
Modelado Multidimensional (MD). Operaciones OLAP

- Operaciones de consulta (OLAP)
  - Slice-dice
    - Definir restricciones sobre niveles de jerarquías
      - Ej. Analizar datos donde el año sea 1999

Ventas'			Producto.Grupo = "Alimentación"	
			Comida	
			Congelada	Fresca
Almacén. Comunidad = "Comunidad Valenciana"	Alicante	Albatera	100	200
		Elche	500	600

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica
Modelado Multidimensional (MD). Operaciones OLAP

- Operaciones de consulta (OLAP)
  - Pivoting
    - Reorientar la vista multidimensional de los datos, es decir, cambiar la distribución de filas/columnas
      - Algunos autores consideran también el intercambio de medidas y hechos como pivoting (kimball, 1996) (Inmon, 1996)

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Indice

- Introducción
- Diseño de almacenes de datos: visión práctica
  - Bases de datos transaccionales vs. almacenes de datos
  - Modelado Multidimensional (MD)
  - Esquema estrella y variantes (en SGBDR)
  - Dimensión tiempo
  - Algunos datos del mercado
- Integración de fuentes: Procesos ETL
- Diseño de almacenes de datos: UML
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

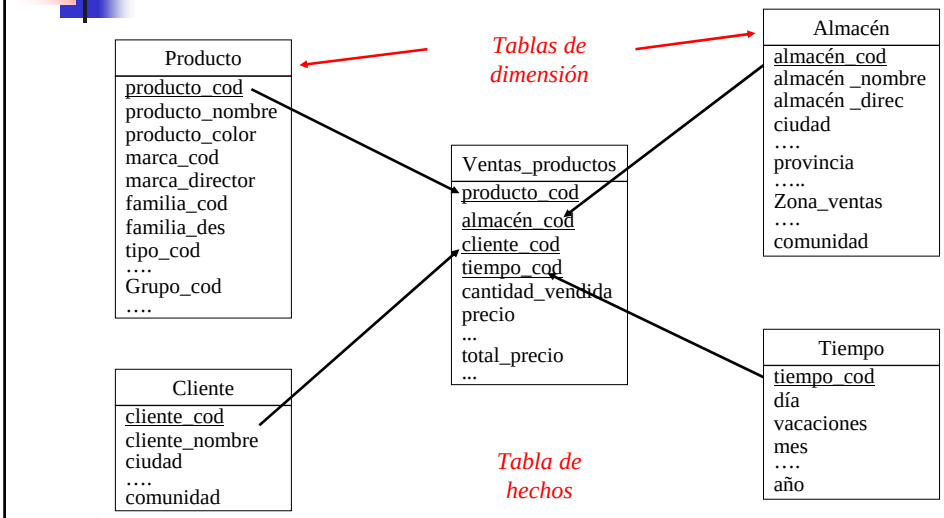
Almacenes de Datos

Indice

- Introducción
- Diseño de almacenes de datos: visión práctica
  - Bases de datos transaccionales vs. almacenes de datos
  - Modelado Multidimensional (MD)
  - Esquema estrella y variantes (en SGBDR)
  - Dimensión tiempo
  - Algunos datos del mercado
- Integración de fuentes: Procesos ETL
- Diseño de almacenes de datos: UML
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica
Esquema estrella y variantes (SGBDR)



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica
Esquema estrella y variantes (SGBDR)

- Tablas de dimensiones
  - Describen el contexto para analizar los hechos
    - Datos “textuales” (Alfanuméricos)
    - Datos desnormalizados → redundancia
  - Cada fila contiene su clave primaria y los atributos descriptores de todos los niveles de jerarquía
  - Tablas más pequeñas que las tablas de hechos

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica

Esquema estrella y variantes (SGBDR)

- *Tabla de dimensión producto*

Producto
producto_cod
producto_nombre
producto_color
marca_cod
marca_director
familia_cod
familia_des
tipo_cod
….
Grupo_cod
….

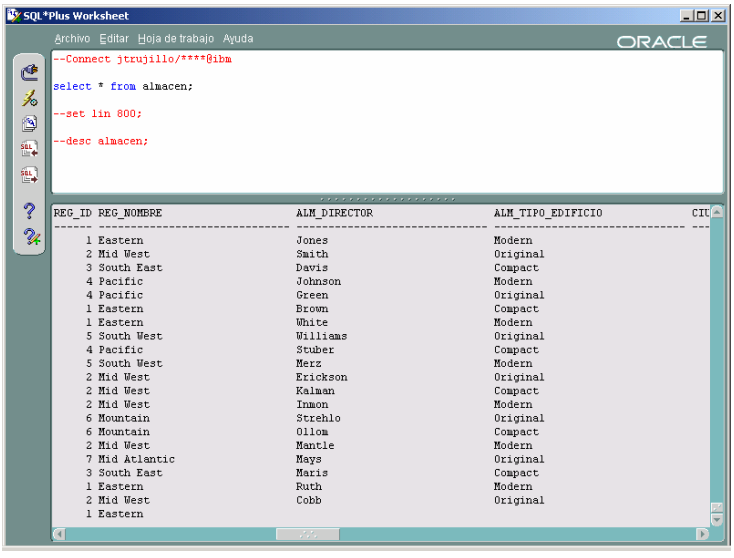
Producto_cod	Producto nombre ….	Family_Desc
1	Puleva Milk ½ L.	Productos lácteos
2	Puleva Milk 1L .	Productos lácteos
3	Yogourt Pascual	Productos lácteos
4	Mr. Proper	Productos limpieza
5	Ajax	Productos limpieza.

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Almacenes de Datos

Diseño de AD: visión práctica

Esquema estrella y variantes (SGBDR)



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

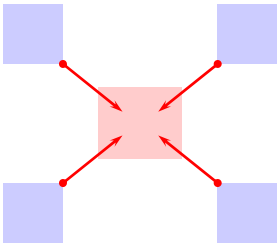
Diseño de AD: visión práctica
Esquema estrella y variantes (SGBDR)

- Tablas de hechos
  - Actividades básicas de empresa
  - Cada fila se compone de:
    - Clave primaria (compuesta por claves ajenas de las dimensiones)
    - Medidas → Datos numéricos
  - Generalmente relación m-n con dimensiones y, m-1 en particular con cada dimensión

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica
Esquema estrella y variantes (SGBDR)

- *Tabla de hecho "ventas"*

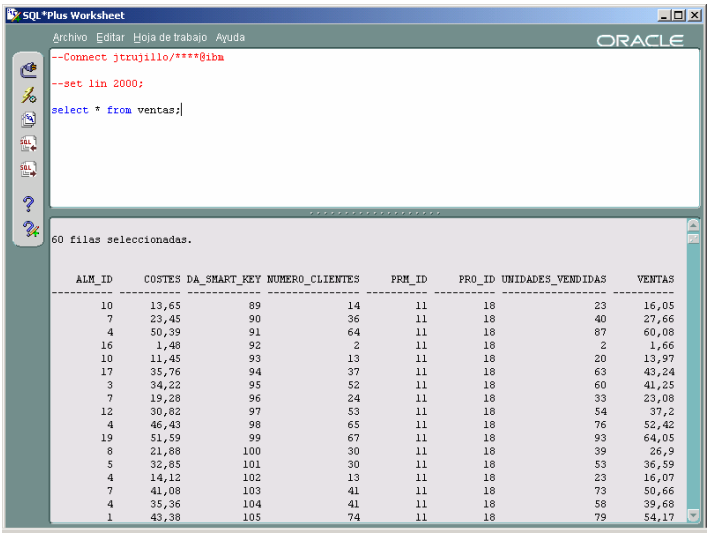


CliKey	ProductoKey	AlmacénKey	Tiempo_key	Sale Amount
1	1	1	1	100€
1	2	1	2	120€
1	3	1	3	200€
.	.	.	.	

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica

Esquema estrella y variantes (SGBDR)

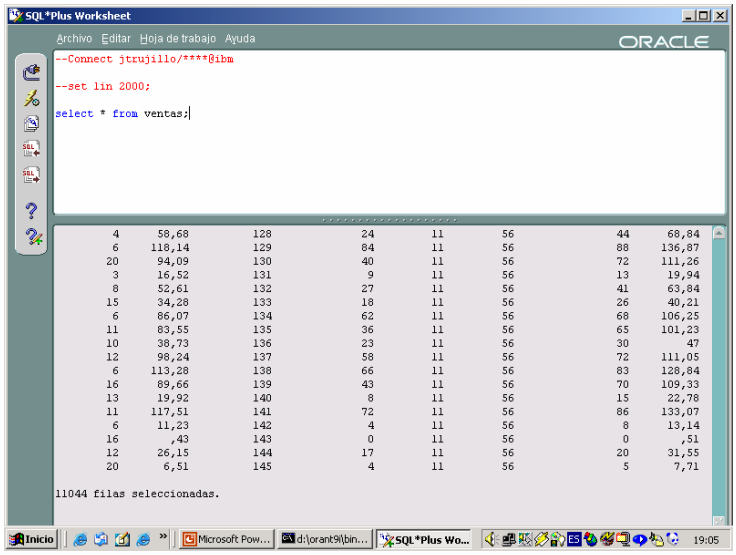


UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Almacenes de Datos

Diseño de AD: visión práctica

Esquema estrella y variantes (SGBDR)



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica
Esquema estrella y variantes (SGBDR)

- Esquema estrella
  - Ventajas
    - Fácil de entender por los usuarios
    - Reduce número de uniones físicas
      - Respuestas rápidas para la mayoría de las consultas
    - Metadatos sencillos
    - Soportado por la inmensa mayoría de aplicaciones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

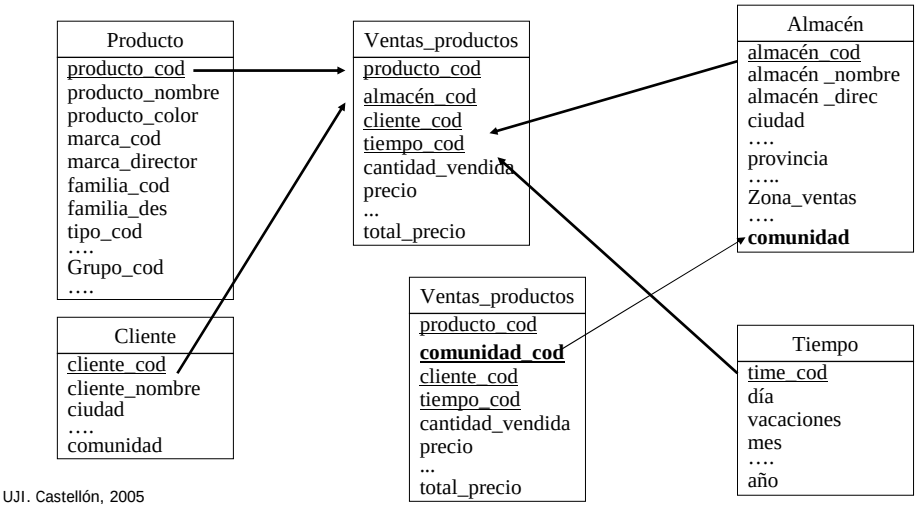
Diseño de AD: visión práctica
Esquema estrella y variantes (SGBDR)

- Esquema estrella
  - Inconvenientes
    - El aumento del tamaño de la tabla de hechos con datos agregados puede empeorar el rendimiento general
      - Por ello se recomienda tablas de hechos agregados al margen
    - Las dimensiones tienen un tamaño enorme
      - Alrededor de 50 atributos (*Kimball*)
    - Es poco robusto o susceptible a cambios

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica

Esquema estrella y variantes (SGBDR). Constelaciones de hechos



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica

Esquema estrella y variantes (SGBDR). Constelaciones de hechos

- Esquema constelaciones de hechos
  - Principal Ventaja
    - Rapidez de respuesta a consultas de datos agregados

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

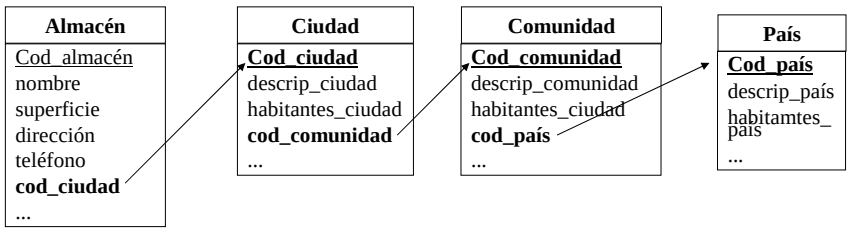
Diseño de AD: visión práctica
Esquema estrella y variantes (SGBDR). Constelaciones de hechos

- Esquema constelaciones de hechos
  - Inconvenientes
    - Un gran número de tablas de agregados
    - Cada tabla de agregados se usa para calcular su nivel
      - Navegar por jerarquías requiere escanear distintas tablas
    - Aumenta el tamaño de los metadatos
    - Dificulta su gestión y mantenimiento ya que para cada carga nueva de datos se han de recalcular todos las tablas de hechos
    - Puede haber requerimientos que necesiten varias tablas

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica
Esquema estrella y variantes (SGBDR). Copos de nieve (snowflake)

- Se normalizan los niveles de jerarquía de dimensiones
  - Tabla dimensión → valores del mínimo nivel de jerarquía



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica
Esquema estrella y variantes (SGBDR). Copos de nieve (snowflake)

- Esquema copos de nieve
  - Ventajas
    - Fácil para definir jerarquías
    - Podría salvar espacio en disco, pero no demasiado
    - Mejora considerablemente el rendimiento cuando un gran número de requerimientos solicita datos agregados o de niveles superiores de jerarquías
      - Los requerimientos escanean un reducido número de filas

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Almacenes de Datos

Diseño de AD: visión práctica
Esquema estrella y variantes (SGBDR). Copos de nieve (snowflake)

- Esquema copos de nieve
  - Inconvenientes
    - Aumenta el número de tablas → aumenta el número de uniones (join)
      - Algunos requerimientos pueden demorarse en exceso
    - Aumenta la complejidad de diseño y mantenimiento
    - Requiere una clave primaria más por cada nivel de jerarquía normalizado
    - No soportado por todas las herramientas del mercado

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica

Esquema estrella y variantes (SGBDR). Copos de nieve (snowflake)

- Recomendaciones
  - Normalmente no se recomienda
    - Realmente cuando el espacio en disco es un problema
  - Normalmente se recomienda normalizar una o dos de las dimensiones más grandes
    - Existen un gran número de filas
  - Suele aplicarse cuando muchos atributos caracterizan a los niveles más altos de las jerarquías

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Almacenes de Datos

Diseño de AD: visión práctica

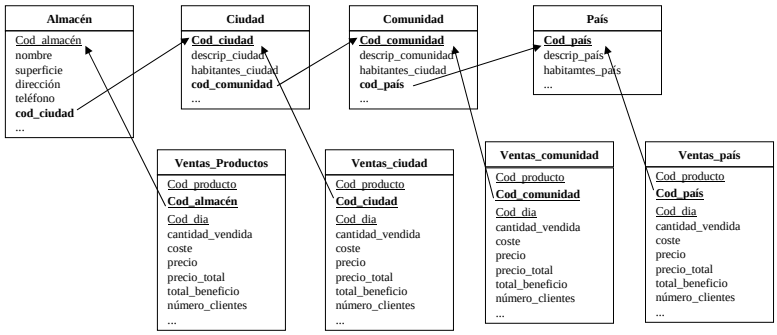
Esquema estrella y variantes (SGBDR). Copos de nieve (snowflake)

- Recomendaciones
  - Utilizar sólo cuando las ventajas son muy explícitas:
    - Ahorro en disco significativo
    - Muchos atributos en los niveles más altos de jerarquías
  - Estadísticamente, el espacio en disco ahorrado utilizando *snowflake schemas* es del 1% del espacio total en disco

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica
Esquema estrella y variantes (SGBDR). Híbrido

- Utilizar copos de nieve con constelaciones de hechos (agregados)



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica
Esquema estrella y variantes (SGBDR). Copos de nieve (snowflake)

- Más particularidades
  - Relaciones *muchos a muchos* hechos-dimensión
  - Hechos que no son hechos *(Factless fact tables)*
  - Dimensiones degeneradas
  - Hechos degenerados
  - Dimensiones que cambian lentamente
  - Hechos y dimensiones comunes
  - Etc.

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

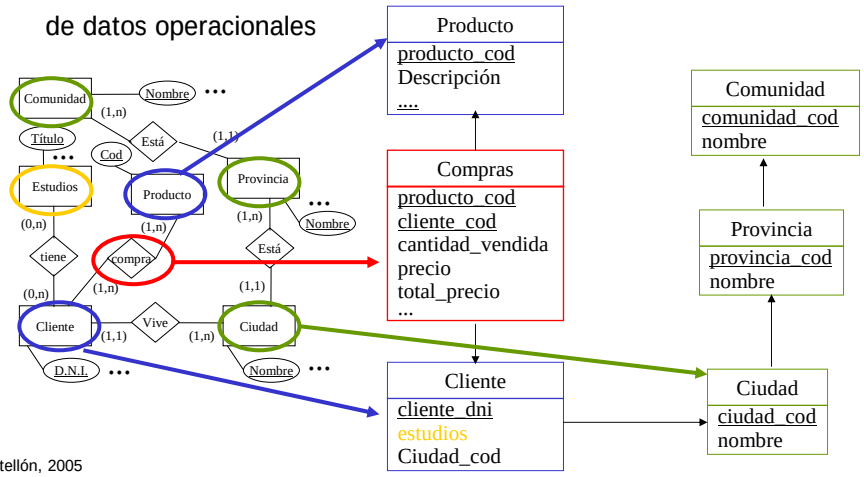
Diseño de AD: visión práctica
Esquema estrella y variantes (SGBDR). Copos de nieve (snowflake)

- Diseño guiado por requerimientos de usuarios (user requirement driven)
  - Análisis requerimientos → Modelado MD
- Diseño guiado por datos
  - A partir de fuentes de datos
- Aproximación híbrida

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica
Esquema estrella y variantes (SGBDR). Un apunte metodológico

- Si partimos de fuentes de datos operacionales



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Indice

- Introducción
- Diseño de almacenes de datos: visión práctica
  - Bases de datos transaccionales vs. almacenes de datos
  - Modelado Multidimensional (MD)
  - Esquema estrella y variantes (en SGBDR)
  - Dimensión tiempo
  - Algunos datos del mercado
- Integración de fuentes: Procesos ETL
- Diseño de almacenes de datos: UML
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

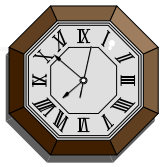
Almacenes de Datos

Indice

- Introducción
- Diseño de almacenes de datos: visión práctica
  - Bases de datos transaccionales vs. almacenes de datos
  - Modelado Multidimensional (MD)
  - Esquema estrella y variantes (en SGBDR)
  - Dimensión tiempo
  - Algunos datos del mercado
- Integración de fuentes: Procesos ETL
- Diseño de almacenes de datos: UML
- Conclusiones

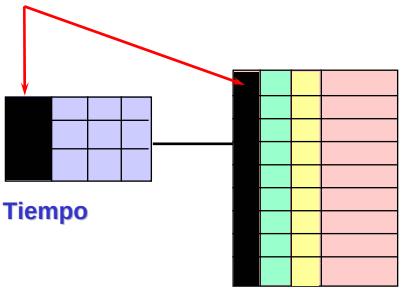
UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica
La dimensión Tiempo



- Tabla de dimensión tiempo
  - Día laborable
  - Período fiscal
  - Principal evento
  - Mes
  - Vacaciones
- Permite un análisis más flexible

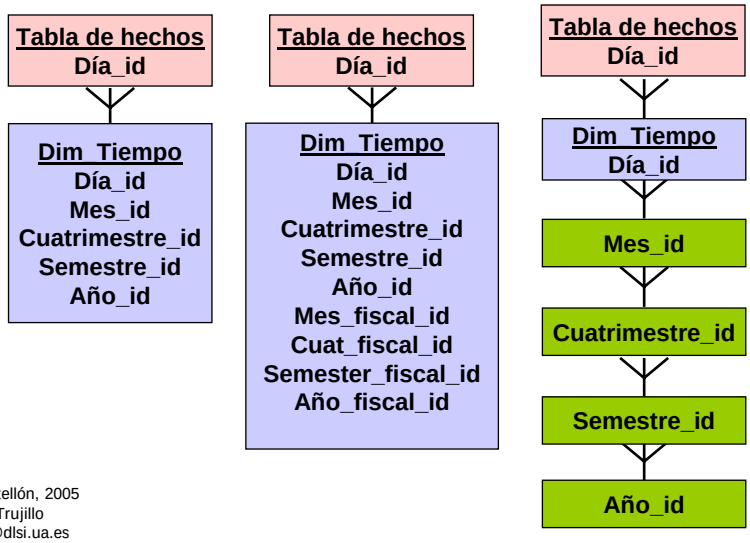
Clave de la Tabla tiempo
Una columna simple: generl. auto generada



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica
La dimensión Tiempo

- Normalizada



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Indice

- Introducción
- Diseño de almacenes de datos: visión práctica
  - Bases de datos transaccionales vs. almacenes de datos
  - Modelado Multidimensional (MD)
  - Esquema estrella y variantes (en SGBDR)
  - Dimensión tiempo
  - Algunos datos del mercado
- Integración de fuentes: Procesos ETL
- Diseño de almacenes de datos: UML
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

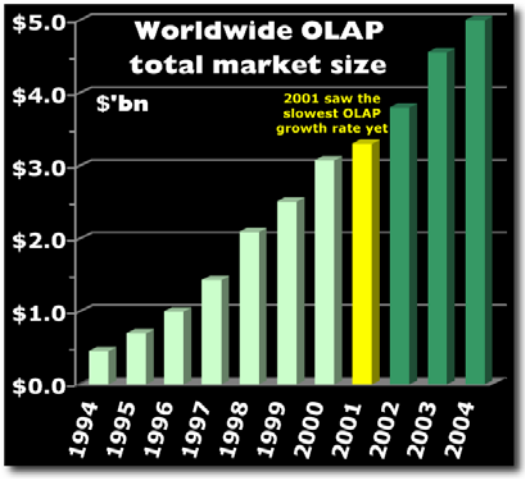
Almacenes de Datos

Indice

- Introducción
- Diseño de almacenes de datos: visión práctica
  - Bases de datos transaccionales vs. almacenes de datos
  - Modelado Multidimensional (MD)
  - Esquema estrella y variantes (en SGBDR)
  - Dimensión tiempo
  - Algunos datos del mercado
- Integración de fuentes: Procesos ETL
- Diseño de almacenes de datos: UML
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica
Algunos datos del mercado



Ratio de crecimiento: bajo

- Mercado grande → dificil crecer
- Cierto grado de saturación
- Media de costes drásticamente disminuida
- Fuente: *The OLAP Report*

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Almacenes de Datos

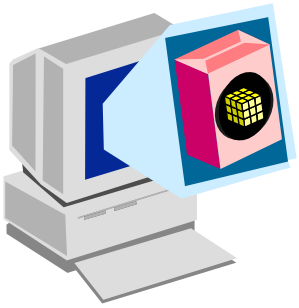
Diseño de AD: visión práctica
Algunos datos del mercado

Hyperion · Microsoft · COGNOS · ORACLE · MicroStrategy · Business Objects · Applix · CARTESIS · COMSHARE · IBM · ADAYTUM · BRIO · sas · crystal decisions · ca

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

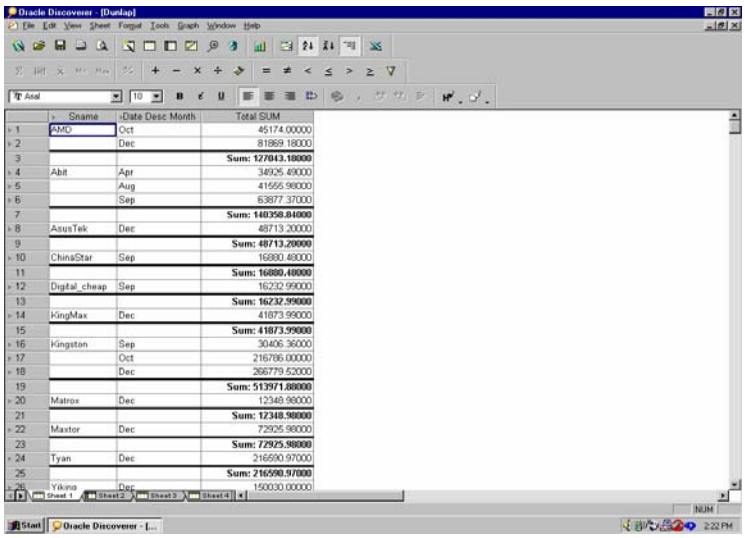
Diseño de AD: visión práctica
Algunos datos del mercado. Oracle

- Almacén de datos
  - Arquitectura del Oracle 9i/10g
  - Oracle Warehouse Builder (OWB)
- Acceso a datos
  - Discoverer para “managers”
  - Express para “analistas”
- Soporta tecnología Web



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica
Algunos datos del mercado. Oracle Discoverer

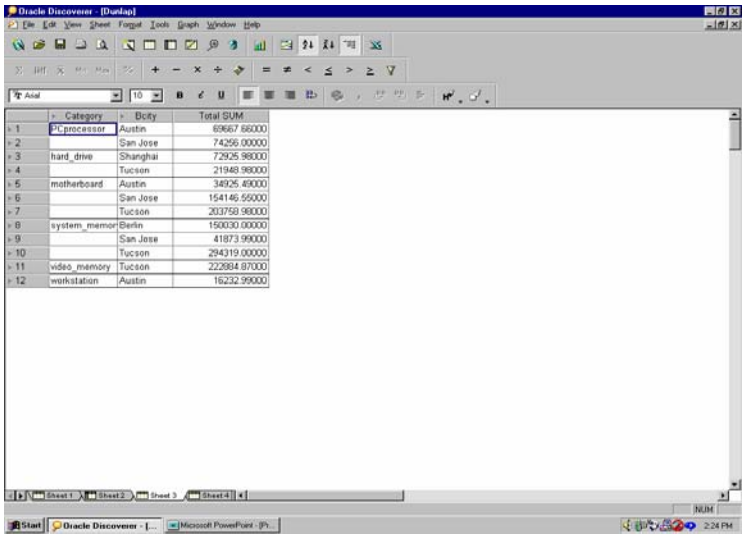


	Sname	Date Desc Month	Total SUM
1	AMD	Oct	45174.00000
2		Dec	81869.18000
3			**Sum: 127043.18000**
4	Abit	Apr	34925.49000
5		Aug	41555.98000
6		Sep	63877.37000
7			**Sum: 140358.84000**
8	AsusTek	Dec	48713.20000
9			**Sum: 48713.20000**
10	ChinaStar	Sep	16880.48000
11			**Sum: 16880.48000**
12	Digital_cheap	Sep	16232.99000
13			**Sum: 16232.99000**
14	KingMax	Dec	41873.99000
15			**Sum: 41873.99000**
16	Kingston	Sep	30406.36000
17		Oct	216786.60000
18		Dec	266779.52000
19			**Sum: 513971.88000**
20	Matrox	Dec	12348.98000
21			**Sum: 12348.98000**
22	Maxtor	Dec	72925.98000
23			**Sum: 72925.98000**
24	Tyan	Dec	216590.97000
25			**Sum: 216590.97000**
26	Yikino	Dec	150030.00000

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica

Algunos datos del mercado. Oracle Discoverer

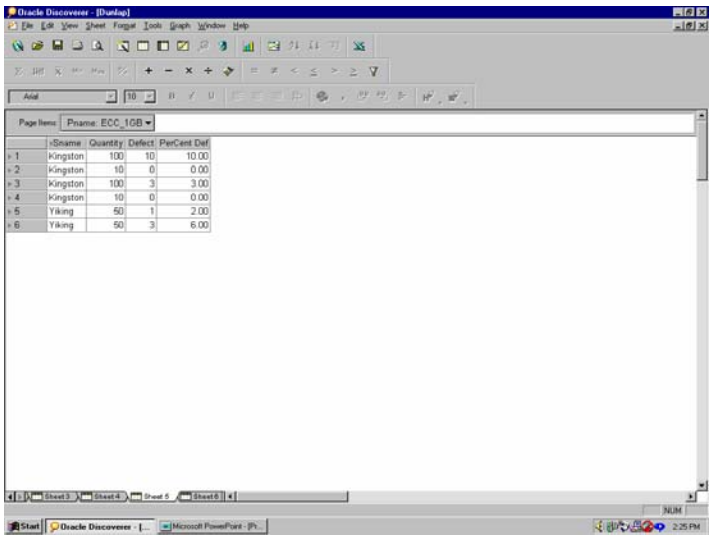


UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Almacenes de Datos

Diseño de AD: visión práctica

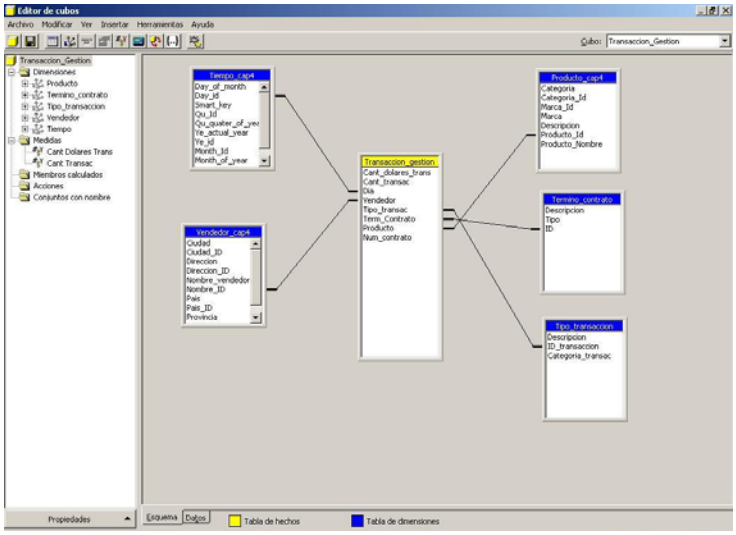
Algunos datos del mercado. Oracle Discoverer



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica

Algunos datos del mercado. SQL Server

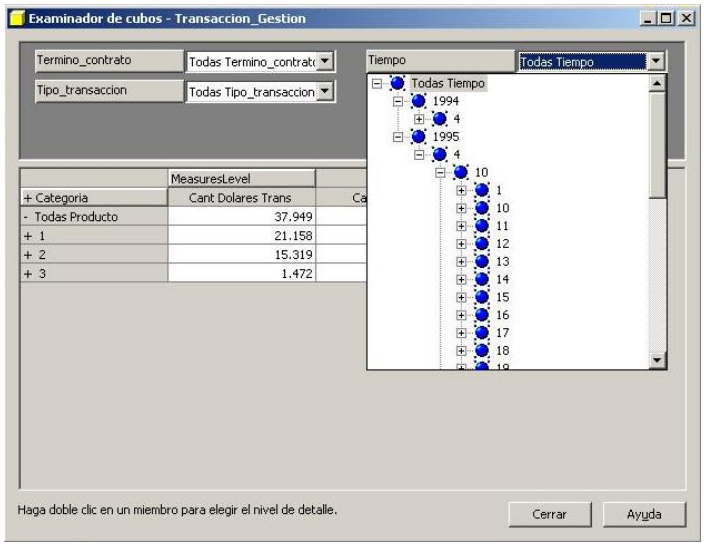


UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Almacenes de Datos

Diseño de AD: visión práctica

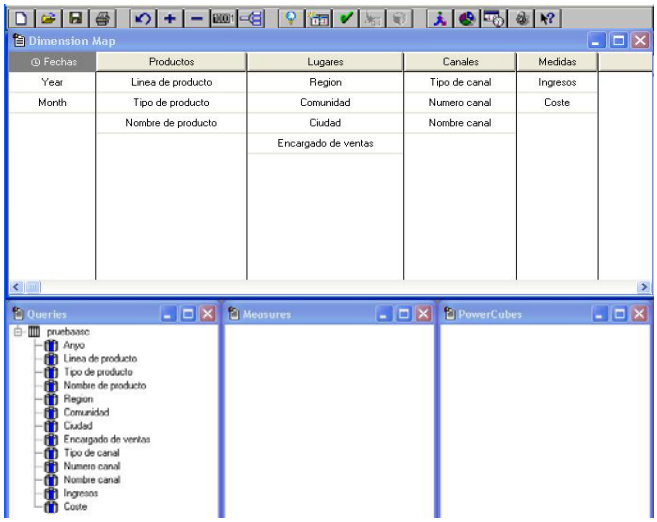
Algunos datos del mercado. SQL Server



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica

Algunos datos del mercado. Cognos Transformer

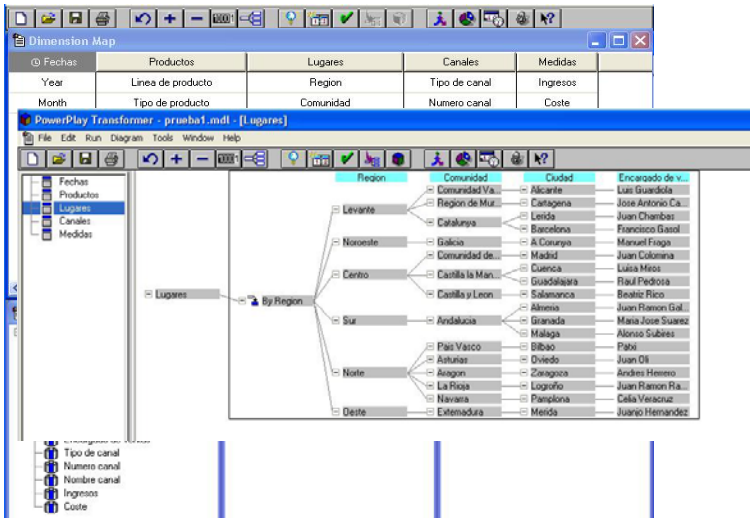


UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Almacenes de Datos

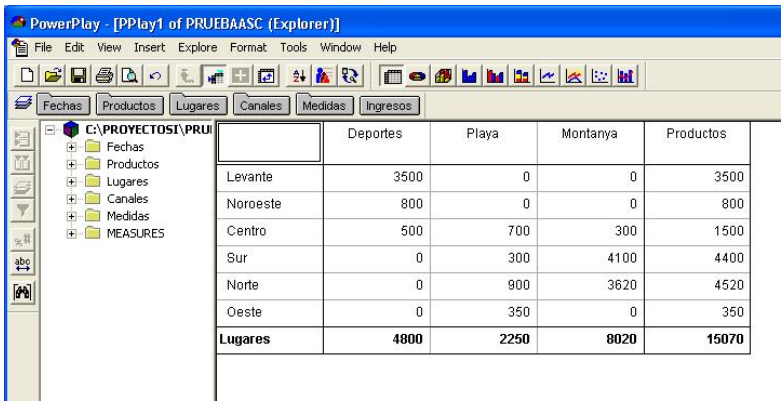
Diseño de AD: visión práctica

Algunos datos del mercado. Cognos Transformer



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: visión práctica
Algunos datos del mercado. Cognos Power Play



	Deportes	Playa	Montanya	Productos
Levante	3500	0	0	3500
Noroeste	800	0	0	800
Centro	500	700	300	1500
Sur	0	300	4100	4400
Norte	0	900	3620	4520
Oeste	0	350	0	350
Lugares	**4800**	**2250**	**8020**	**15070**

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Indice

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Indice

- Introducción
- Diseño de almacenes de datos: visión práctica
- Integración de fuentes: Procesos ETL
- Diseño de almacenes de datos: UML
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Almacenes de Datos

Indice

- Introducción
- Diseño de almacenes de datos: visión práctica
- Integración de fuentes: Procesos ETL
  - Introducción
  - Extracción
  - Transformación
  - Carga
- Diseño de almacenes de datos: UML
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Indice

- Introducción
- Diseño de almacenes de datos: visión práctica
- Integración de fuentes: Procesos ETL
  - Introducción
  - Extracción
  - Transformación
  - Carga
- Diseño de almacenes de datos: UML
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Almacenes de Datos

Indice

- Introducción
- Diseño de almacenes de datos: visión práctica
- Integración de fuentes: Procesos ETL
  - Introducción
  - Extracción
  - Transformación
  - Carga
- Diseño de almacenes de datos: UML
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Introducción

- Algunas tareas comunes de procesos ETL:
  - Datos de distintas fuentes se tienen que unir (join)
  - Datos se tienen que agregar
  - Datos se han de convertir a un formato común
  - Generar claves auto generadas
  - Verificar la calidad de los datos
  - Etc.

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Introducción

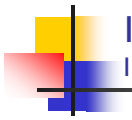
- Definir una estrategia de calidad de datos para la empresa según política de toma de decisiones
- Definir el nivel de calidad óptimo de los datos
- Considerar el modificar reglas de las fuentes de datos operacionales
- Básico → documentar las fuentes
- Diseñar los procesos de limpieza (y sus tareas) de forma muy cuidadosa
- Los procesos de limpieza iniciales puedes variar de los procesos de refresco posteriores

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Introducción

- Cuidado → datos incorrectos o engañosos producirán decisiones estratégicas erróneas
- El mercado de herramientas de ETL en 2001: sobre $667 millones in USA
- Esfuerzo en ETL: 30% del presupuesto total de los proyectos de DW
- Actualmente el diseño y mantenimiento de procesos ETL es todavía un asunto “pendiente”
- Aunque varias herramientas en mercado, no disponemos de modelo o metodología estándar para su diseño desde primeros pasos de un proyecto de DW

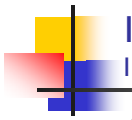
UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es



Integración de fuentes: Procesos ETL
Introducción

- Rutinas convencionales COBOL, 4GL
- Herramientas especializadas
- Proceso de conversión personalizada
- Expertos de negocio

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

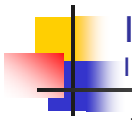


Integración de fuentes: Procesos ETL
Introducción

- Seis pasos detallados:

1. Seleccionar las fuentes para extraer datos
2. Transformar las fuentes
3. Unir las fuentes
4. Seleccionar las estructuras destino a cargar datos (hechos, dimensiones, etc.)
5. Mapear los atributos de las fuentes en los destinos
6. Cargar los datos

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es



Integración de fuentes: Procesos ETL
Introducción

- El paso de transformación también puede incluir limpieza de datos (detectar y borrar errores e inconsistencias)
- La creación manual y mantenimiento de los procesos ETL aumenta el coste de los DW
- CUIDADO: Documentación con gran cantidad de páginas con código de programas ETL

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Indice

- Introducción
- Diseño de almacenes de datos: visión práctica
- Integración de fuentes: Procesos ETL
  - Introducción
  - Extracción
  - Transformación
  - Carga
- Diseño de almacenes de datos: UML
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Almacenes de Datos

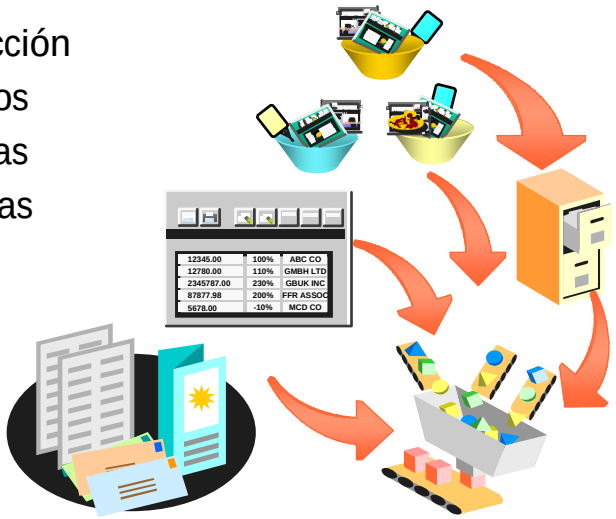
Indice

- Introducción
- Diseño de almacenes de datos: visión práctica
- Integración de fuentes: Procesos ETL
  - Introducción
  - Extracción
  - Transformación
  - Carga
- Diseño de almacenes de datos: UML
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Extracción

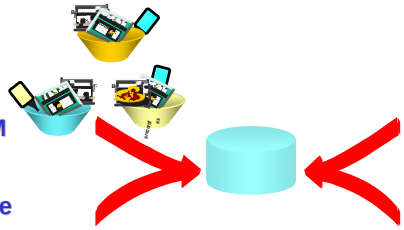
- Producción
- Archivos
- Internas
- Externas



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Extracción

IMS
DB2
VSAM
SQL
Oracle
Sybase
Rdb

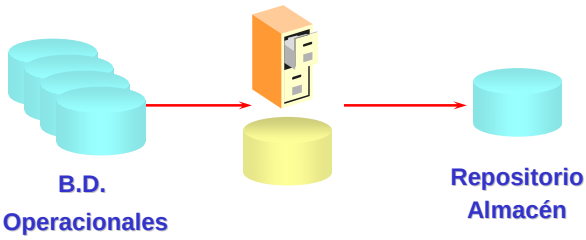


SAP
Sistemas médicos
Predicción financiera
Oracle Financial

- Distintas plataformas de S.O.
- Plataformas Hardware
- Sistemas de ficheros
- Sistemas de bases de datos

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Extracción



- Datos históricos ya almacenados
- Útiles para análisis de largos periodos de tiempo
- Útiles para primera carga
- Generalmente requerirán transformaciones
- Datos estructurados y no estructurados

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Extracción

- Información desde fuera de la organización

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Extracción

- Programas - C, COBOL, PL/SQL
- Gateways – acceso a b.d. transparentes
- Herramientas
  - Coste inicial muy alto
  - Automatización
  - Limpieza de datos

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Indice

- Introducción
- Diseño de almacenes de datos: visión práctica
- Integración de fuentes: Procesos ETL
  - Introducción
  - Extracción
  - Transformación
  - Carga
- Diseño de almacenes de datos: UML
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Indice

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Transformación. Algunas transformaciones comunes

- Anomalías existen en fuentes operacionales
  - Limpiar
  - Valores incoherentes
    - Universidad de Alicante
    - Univ. Alicante
    - U. de Alicante
  - Anomalías de instancias y codificado
  - Valores nulos para algunos campos
  - ¿¿ Violación de reglas de integridad ?? (ETL)

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

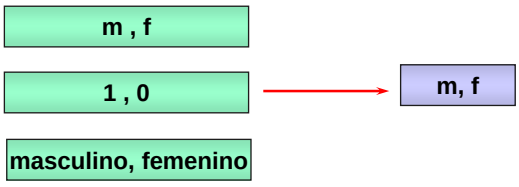
Integración de fuentes: Procesos ETL
Transformación. Algunas transformaciones comunes

- Wrapper: transformar fuentes de datos nativos en fuentes de datos basadas en registros

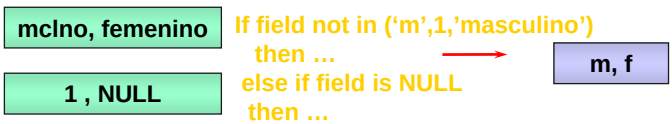
UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Transformación. Algunas transformaciones comunes

- Codificación múltiple



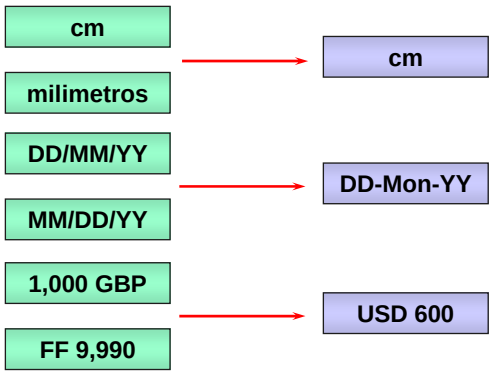
- Detectar datos erróneos



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Transformación. Algunas transformaciones comunes

- Varios formatos válidos y estándares
- Herramientas o filtros para pre-procesar



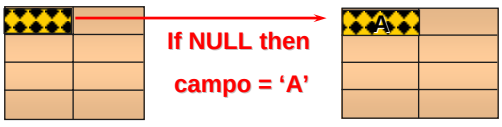
Conversor
A→B

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Transformación. Algunas transformaciones comunes

- NULL y valores que faltan
  - Ignorar
  - Esperar
  - Marcar las filas
  - Extraer bajo condiciones establecidas

Filter



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Transformación. Algunas transformaciones comunes

- Valores duplicados

Join

```
SELECT ...
FROM   table_a, table_b
WHERE table_a.key (+) = table_b.key
UNION
SELECT ...
FROM   table_a, table_b
WHERE table_a.key = table_b.key (+)
```

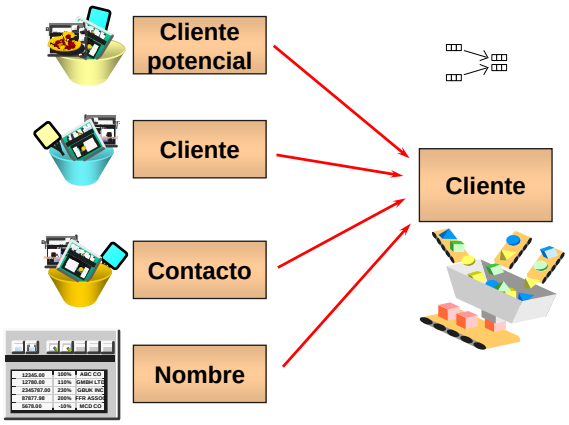
ACME Inc
ACME Inc
ACME Inc

→ ACME Inc

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Transformación. Algunas transformaciones comunes

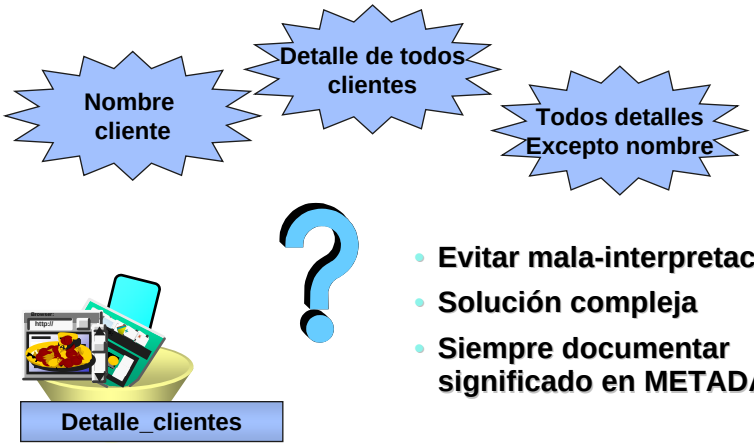
- Atributos compatibles



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Transformación. Algunas transformaciones comunes

- Significado correcto de cada elemento



- **Evitar mala-interpretación**
- **Solución compleja**
- **Siempre documentar significado en METADATA**

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Transformación. Algunas transformaciones comunes

- No hay clave única
- Valores que faltan
- Nombres personales y comerciales mezclados
- Diferentes direcciones para el mismo miembro
- Diferentes nombres y ortografía para el mismo miembro
- Muchos nombres en la misma linea
- Un nombre en dos líneas

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Transformación. Algunas transformaciones comunes. Fusión

- Transacciones operacionales no son un mapeo 1-to-1 con los datos del DW.
- Datos del DW son "fusionados/unidos" para proporcionar información para el análisis
  - Ejemplo:
    - Compra de un producto
    - Devolución mismo producto

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Transformación. Algunas transformaciones comunes

- Permitir análisis del tiempo
- Añadir datos de tiempo en los datos de hechos y dimensiones
  - Añadir *triggers*
  - Aplicaciones de "código"
  - Comparar tablas

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Transformación. Claves autogeneradas.

Surrogate
123→

#1	Sale	1/2/98	12:00:01 Ham Pizza	$10.00
#2	Sale	1/2/98	12:00:02 Cheese Pizza	$15.00
#3	Sale	1/2/98	12:00:02 Anchovy Pizza	$12.00
#4	Return	1/2/98	12:00:03 Anchovy Pizza	- $12.00
#5	Sale	1/2/98	12:00:04 Sausage Pizza	$11.00

Valores de datos o claves artificiales

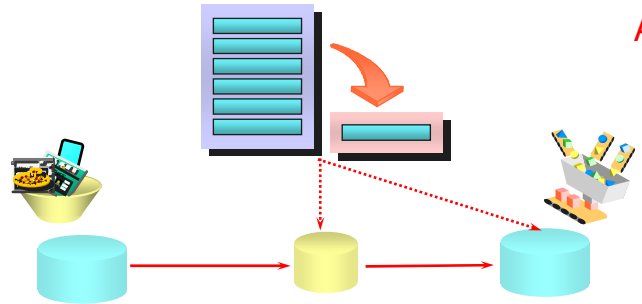
#dw1	Sale	1/2/98	12:00:01 Ham Pizza	$10.00
#dw2	Sale	1/2/98	12:00:02 Cheese Pizza	$15.00
#dw3	Sale	1/2/98	12:00:04 Sausage Pizza	$11.00

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Transformación. Datos agregados/sumados.

- Durante extracción o tratamiento (staging)
- Después de cargar los datos en el DW

Aggregate
→+→



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Indice

- Introducción
- Diseño de almacenes de datos: visión práctica
- Integración de fuentes: Procesos ETL
  - Introducción
  - Extracción
  - Transformación
  - Carga
- Diseño de almacenes de datos: UML
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Almacenes de Datos

Indice

- Introducción
- Diseño de almacenes de datos: visión práctica
- Integración de fuentes: Procesos ETL
  - Introducción
  - Extracción
  - Transformación
  - Carga
- Diseño de almacenes de datos: UML
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Carga. Objetivos.

- Identificar el transporte de datos para la primera vez y refrescos siguientes
- Describir consideraciones estratégicas e implementar el refresco de datos
- Identificar métodos empleados para capturar cambios en los datos y, aplicarlos en el DW
- Describir técnicas de transporte
- Identificar las tareas que se llevan a cabo después de que los datos se cargan

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Carga. Planteamiento general.

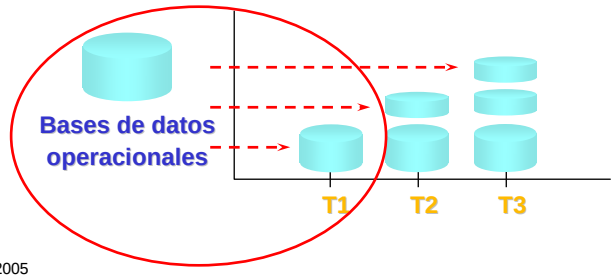
- Carga (*loading*) lleva los datos al DW.

- Carga puede necesitar mucho tiempo:
  - Considerar la ventana de carga
  - Planificar → intentar automatizar todos los procesos
- Carga inicial mueve grandes volumen.
- Cargas posteriores mueven volumen de datos más pequeños (en teoría).
- El “negocio” determina el ciclo de las cargas.

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Carga. Primera carga.

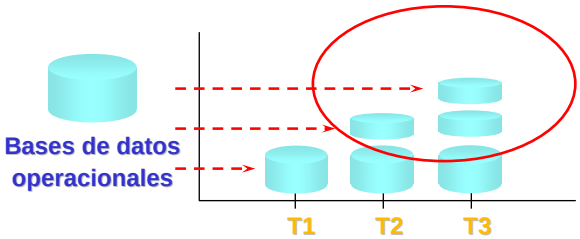
- Primera carga del DW con datos históricos
- Requiere grandes volúmenes de datos
- Puede emplear distintas tareas ETL
- Requiere grandes cantidades de procesamiento después de la primera carga.



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Carga. Refresco.

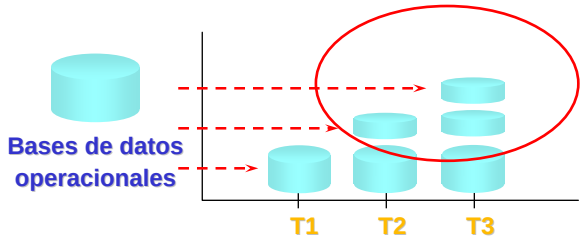
- Realizados de acuerdo al ciclo del negocio.
- Es una tarea más simple
- Menos datos para la carga
- ETL menos complejos
- Menos rutinas de procesamiento después de la carga



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Carga. Estrategia de refresco.

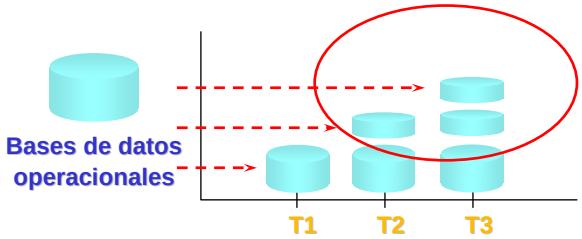
- Considerar la ventana de carga
- Identificar los volúmenes de datos
- Identificar ciclos
- Conocer la infraestructura técnica
- Planificar un área de “trastienda” (staging)
- Determinar cómo detectar cambios



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Carga. Utilizar requerimientos.

- Usuarios definen también el ciclo de refresco
- Documentar todas las tareas y procesos
- Consultar usuarios expertos



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

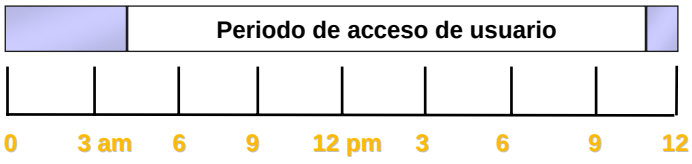
Integración de fuentes: Procesos ETL
Carga. Procesos de transporte.

- Especificar
  - Técnicas y herramientas
  - Métodos de transferencia de ficheros
  - La ventana de carga
  - Ventana de tiempo para otras tareas
  - Volúmenes de primera carga y refresco
  - Frecuencia del ciclo de refresco
  - Ancho de banda de conectividad

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Carga. Ventana de carga.

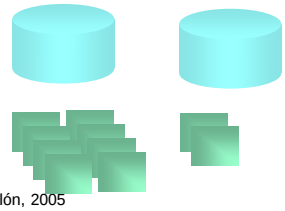
- Planificar y construir procesos de acuerdo a una estrategia.
- Considerar volúmenes de datos
- Identificar infraestructura técnica
- Asegurar la actualidad de los datos
- Considerar en primer lugar los requerimientos de acceso de usuarios
- Muchos requerimientos puede significar una ventana de carga pequeña



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Carga. Granularidad.

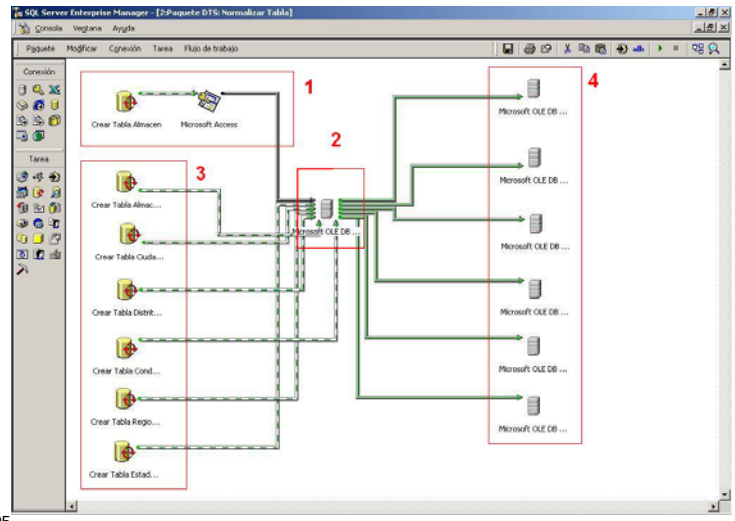
- Importante diseñarla
- Requerimientos de espacio
  - Almacenamiento
  - Copias
  - Recuperación
  - Particionamiento
  - Carga



- Nivel de granularidad bajo
  - Caro, alto nivel de procesamiento, más disco, detalle,
- Nivel de granularidad alto
  - Más barato, menos procesamiento, menos disco, poco detalle

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL
Ejemplo con SQL Server.

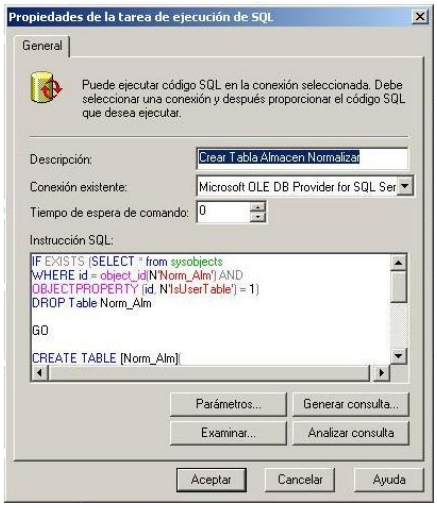


UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Integración de fuentes: Procesos ETL

Ejemplo con SQL Server.

- 1.- Tarea de ejecución SQL
- 2.- Conexión M. OLE DB Provider para SQL Server
- 3.- Tablas destino normalizadas
- 4.- Conexión M. OLE DB Provider para SQL Server para cada tabla normalizada



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Almacenes de Datos

Integración de fuentes: Procesos ETL

Ejemplo con SQL Server.

- Ejemplo de transformación

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Indice

- Introducción
- Diseño de almacenes de datos: visión práctica
- Integración de fuentes: Procesos ETL
- Diseño de almacenes de datos: UML
  - Método global
  - Diseño del almacén de datos y procesos ETL con UML
  - Herramientas CASE
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Almacenes de Datos

Indice

- Introducción
- Diseño de almacenes de datos: visión práctica
- Integración de fuentes: Procesos ETL
- Diseño de almacenes de datos: UML
  - Método global
  - Diseño del almacén de datos y procesos ETL con UML
  - Herramientas CASE
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: UML
Método global

- Objetivo: Un método completo de diseño de AD

- Principios de nuestra aproximación:
  - Abordar el diseño conceptual de los A.D.
  - Notación estándar → UML
  - Completo → Incluye las principales fases de diseño
  - Potente pero fácil de entender
    - Diferentes niveles de detalle para diferentes usuarios (técnicos y usuarios finales) → Empleo de paquetes
  - Método flexible → Punto de inicio, pero no un esquema estricto
  - Aplicable → Soportado por herramientas CASE

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

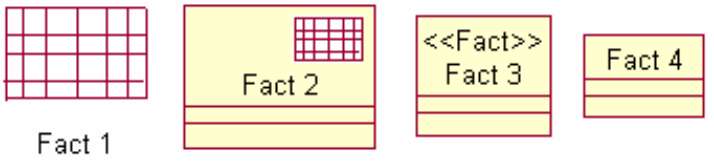
Diseño de AD: UML
Método global

- UML es un lenguaje de modelado visual de **propósito general**
- Mecanismos de extensión permiten adaptarlo a dominios específicos
- Mecanismos:
  - Stereotypes → Nuevos elementos de construcción
  - Tagged values → Nuevas propiedades
  - Constraints → Nuevas semánticas

Profile

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

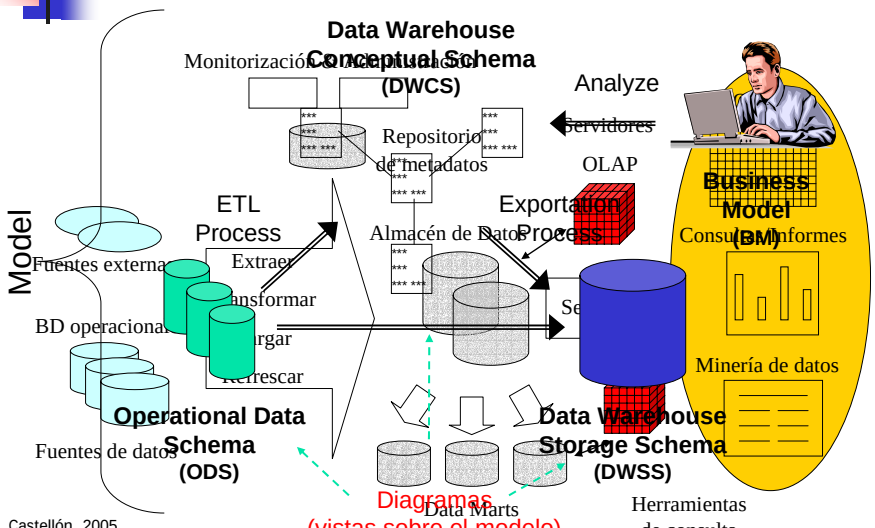
Diseño de AD: UML
Método global



Icono Adorno Etiqueta Ninguno

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: UML
Método global



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: UML
Método global

- Proponemos *Data Warehouse Engineering Process* (DWEP), basado en UP pero adaptado a la construcción de almacenes de datos
- Seis flujos:
  - Requisitos
  - Análisis
  - Diseño
  - Implementación
  - Test
  - Revisión posterior

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: UML
Método global

- Etapas:
  - Origen
  - Integración
  - Almacén de datos
  - Adaptación
  - Cliente
- Niveles:
  - Conceptual
  - Lógico
  - Físico
- Diagramas: las 5 etapas y los 3 niveles originan 15 diagramas (no se tienen que definir todos los diagramas en todos los proyectos)

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: UML
Método global

	Source (S) (OLTP, external data, ...)	Integration	Data Warehouse (DW)	Customization	Client (C) (OLAP, data mining, ...)
Conceptual	**SCS** Class diagram *Standard UML*	**DM** Class diagram *Data Mapping Profile*	**DWCS** Class diagram *Standard UML*	**DM** Class diagram *Data Mapping Profile*	**CCS** Class diagram *Standard UML Multidimensional Profile*
Logical	**SLS** Class diagram *Different data modeling profiles*	**ETL Process** Class diagram *ETL Profile*	**DWLS** Class diagram *Different data modeling profiles*	**Exporting Process** Class diagram *ETL Profile*	**CLS** Class diagram *Different data modeling profiles*
Physical	**SPS** Comp. & deploy. diagrams *Database Deployment Profile*	**Transportation Diagram** Deployment diagram *Database Deployment Profile*	**DWPS** Comp. & deploy. diagrams *Database Deployment Profile*	**Transportation Diagram** Deployment diagram *Database Deployment Profile*	**CPS** Comp. & deploy. diagrams *Database Deployment Profile*

LEGEND: CS: Conceptual Schema, LS: Logical Schema, PS: Physical Schema, Comp. & deploy: Component and deployment

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Indice

- Introducción
- Diseño de almacenes de datos: visión práctica
- Integración de fuentes: Procesos ETL
- Diseño de almacenes de datos: UML
  - Metodología global
  - Diseño del almacén de datos y procesos ETL con UML
  - Herramientas CASE
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Indice

- Introducción
- Diseño de almacenes de datos: visión práctica
- Integración de fuentes: Procesos ETL
- Diseño de almacenes de datos: UML
  - Metodología global
  - Diseño del almacén de datos y procesos ETL con UML
  - Herramientas CASE
- Conclusiones

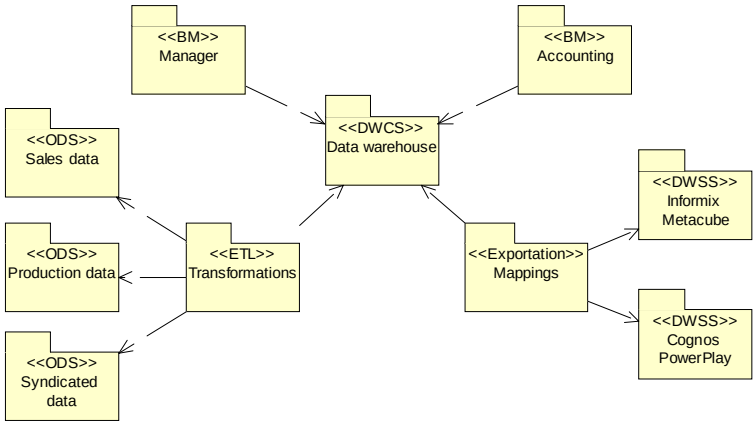
UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: UML

Diseño de almacenes de datos con UML (ER'02, UML'02,...)

Diagrama general (nivel 0)

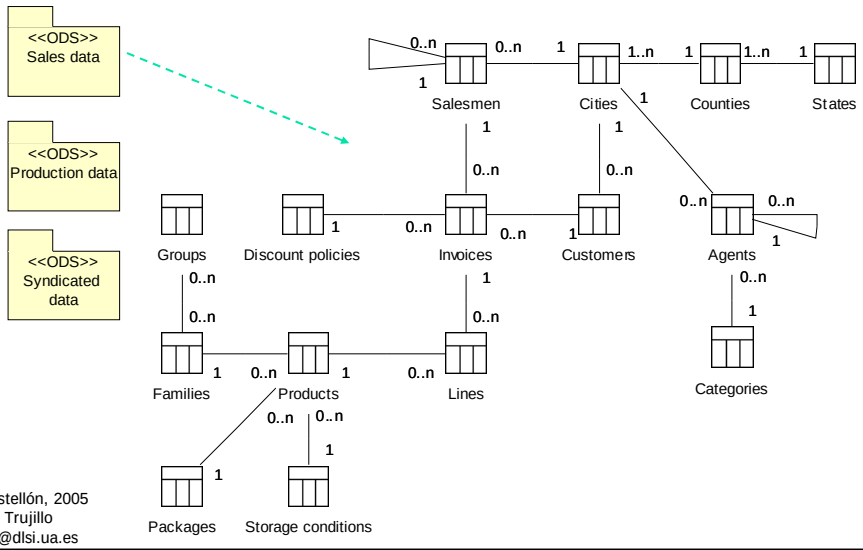
<<ODS>>, <<DWCS>>, <<DWSS>>, <<BM>>, <<ETL>>, <<Exportation>>



UJI. Cas
Juan C.
Jtrujillo@dlsi.ua.es

Diseño de AD: UML

Diseño de almacenes de datos con UML (ER'02, UML'02,...)



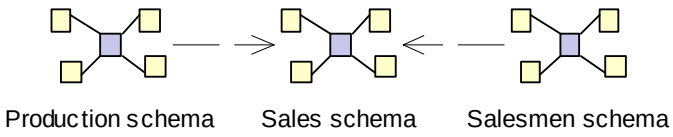
UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: UML

Diseño de almacenes de datos con UML (ER'02, UML'02,...)

Definición del modelo (nivel 1)

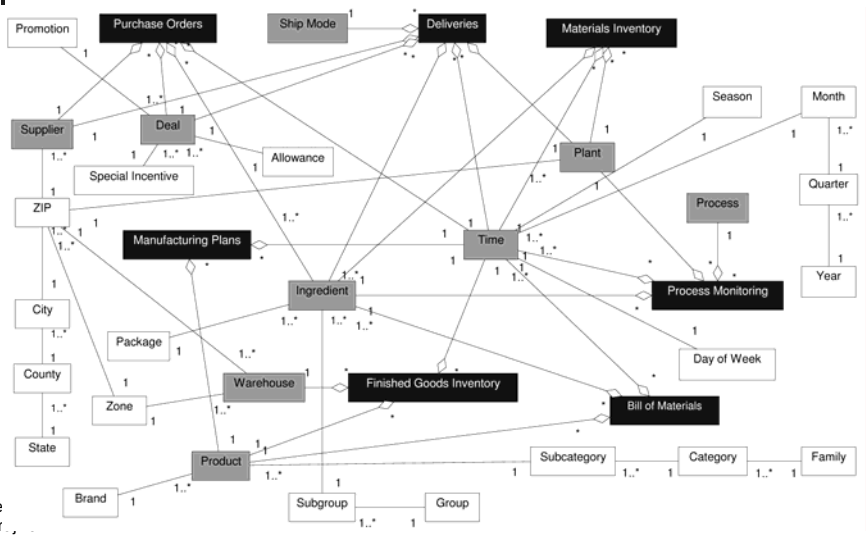
<<StarPackage>>



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: UML

Diseño de almacenes de datos con UML (ER'02, UML'02,...)



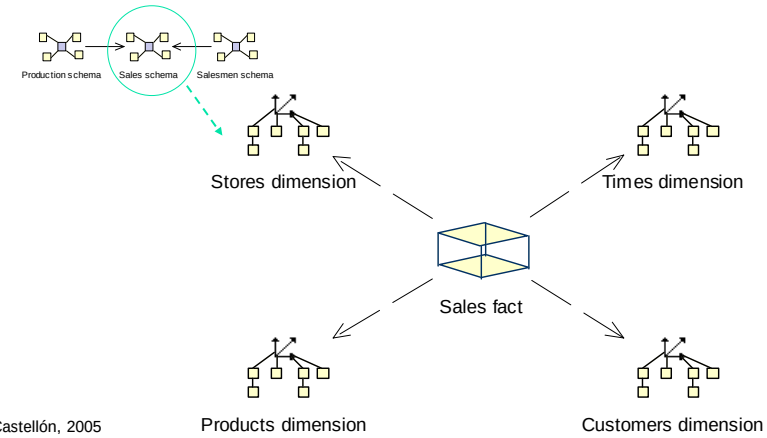
UJI. Caste
Juan C. Tr
Jtrujillo@dlsi.ua.es

Diseño de AD: UML

Diseño de almacenes de datos con UML (ER'02, UML'02,...)

Definición del esquema estrella (nivel 2)

<<FactPackage>>, <<DimensionPackage>>



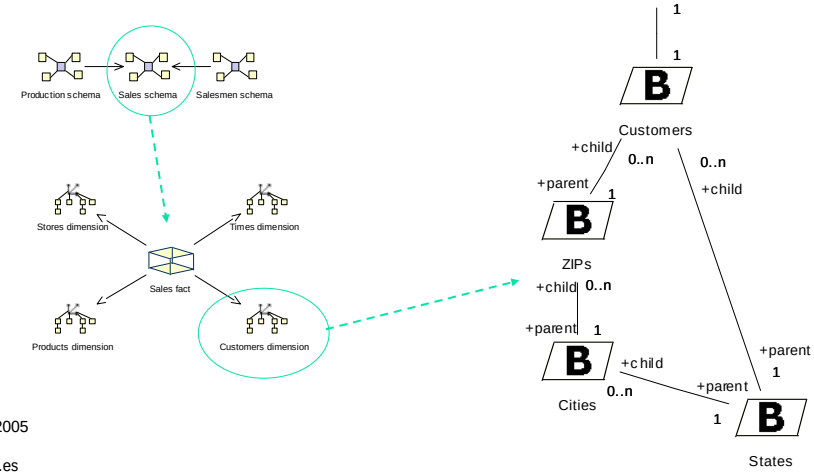
UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: UML

Diseño de almacenes de datos con UML (ER'02, UML'02,...)

Definición de hechos/dimensiones (nivel 3)

<<Fact>>, <<Dimension>>, <<Base>>

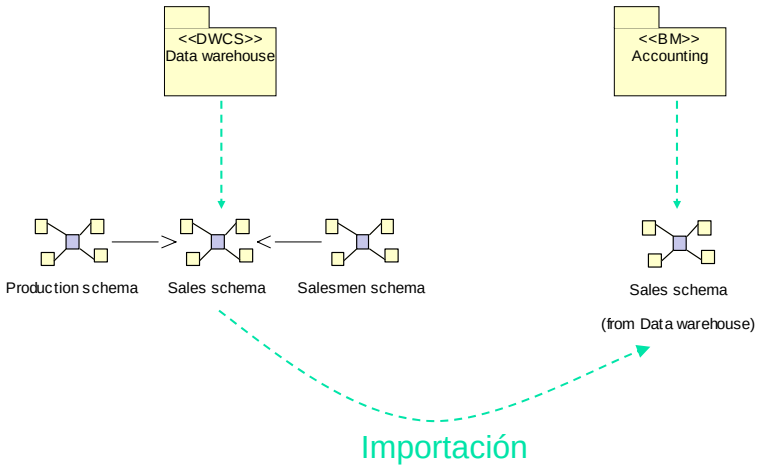


UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Almacenes de Datos

Diseño de AD: UML

Diseño de almacenes de datos con UML (ER'02, UML'02,...)



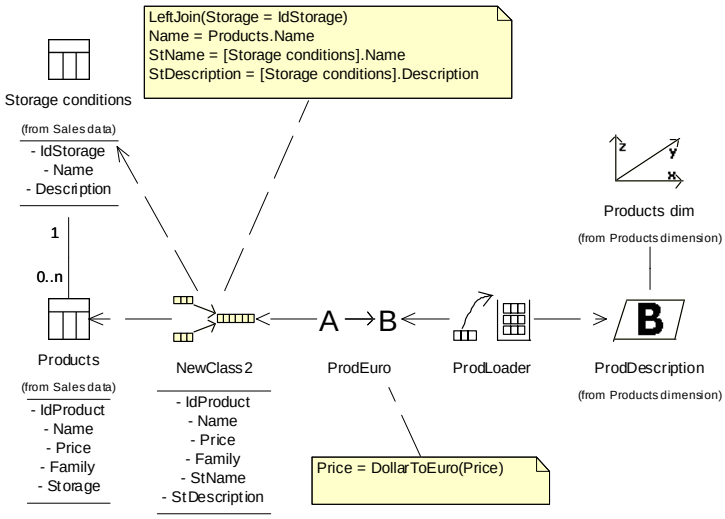
UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: UML
Procesos ETL (ER'03, ER'04,...)

- Un conjunto reducido pero potente para reflejar las principales operaciones de procesos ETL:
  - Aggregation
  - Conversion
  - Filter
  - Incorrect
  - Join
  - Loader
  - Log
  - Merge
  - Surrogate, and
  - Wrapper

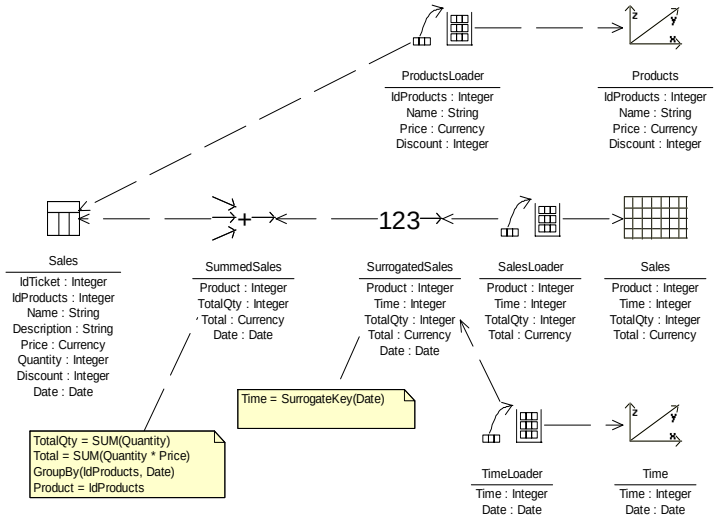
UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: UML
Procesos ETL. Ejemplo.



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: UML
Procesos ETL. Ejemplo.



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Indice

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Indice

- Introducción
- Diseño de almacenes de datos: visión práctica
- Integración de fuentes: Procesos ETL
- Diseño de almacenes de datos: UML
  - Metodología global
  - Diseño del almacén de datos con UML
  - Herramientas CASE
- Conclusiones

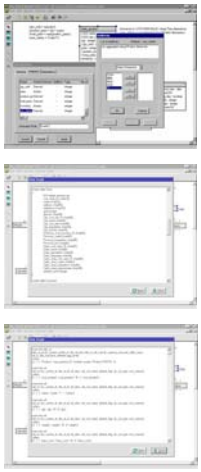
UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

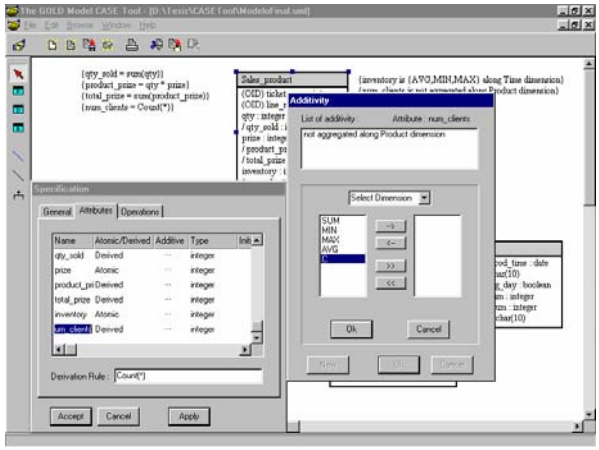
Diseño de AD: UML
Herramientas CASE

Herramientas CASE

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: UML
Herramientas CASE

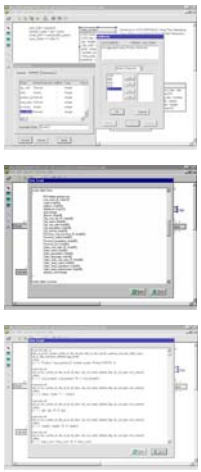


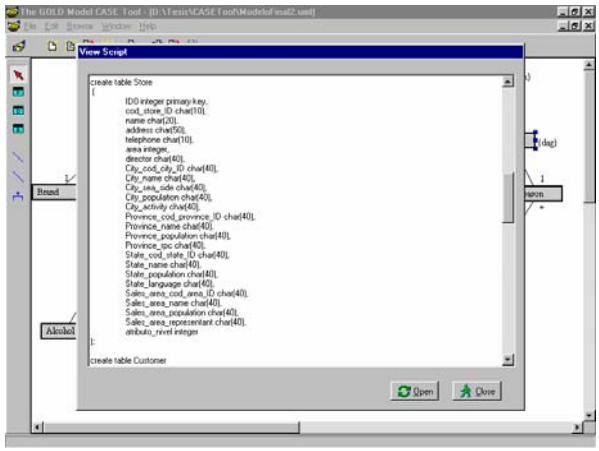


UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Almacenes de Datos

Diseño de AD: UML
Herramientas CASE

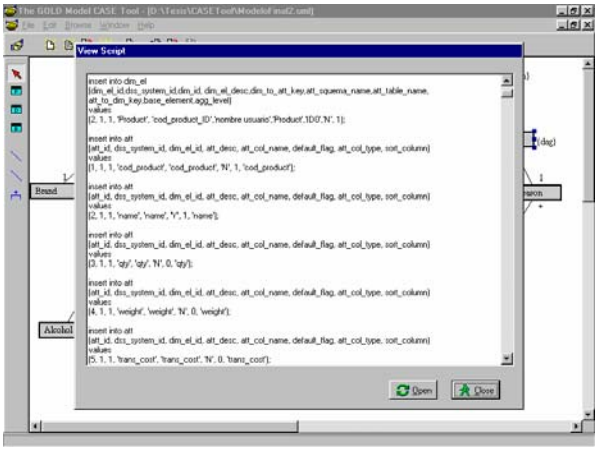




UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

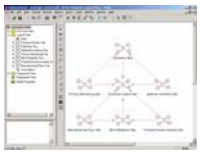
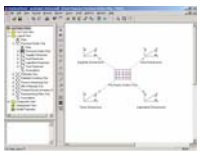
Diseño de AD: UML
Herramientas CASE



UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

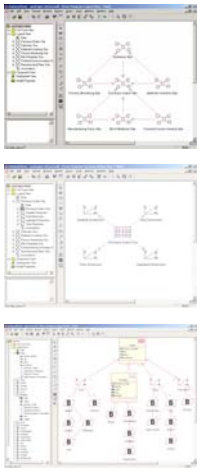
Diseño de AD: UML
Herramientas CASE

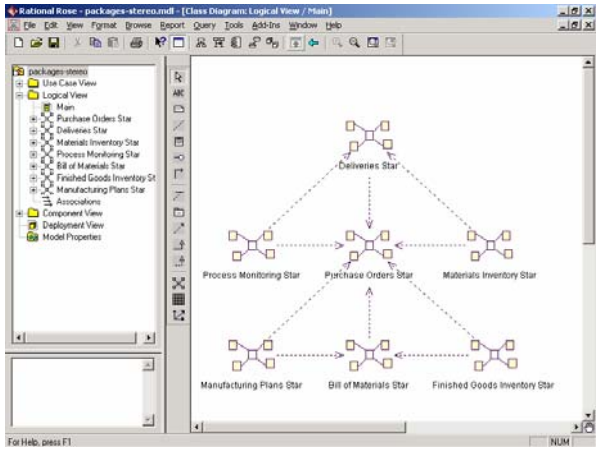




Add-in para Rational Rose

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: UML
Herramientas CASE

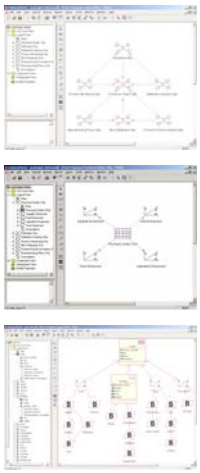


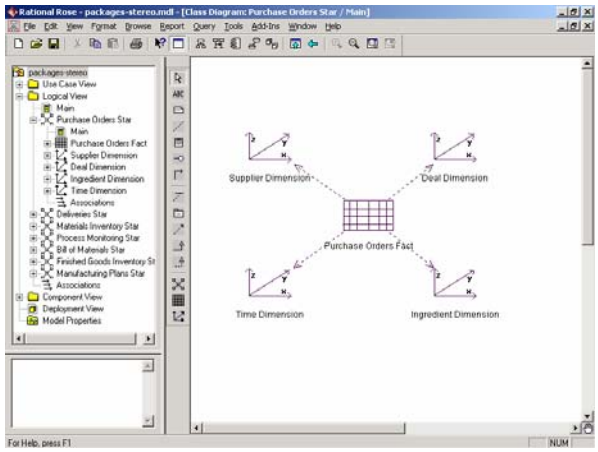


UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Almacenes de Datos

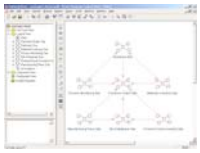
Diseño de AD: UML
Herramientas CASE

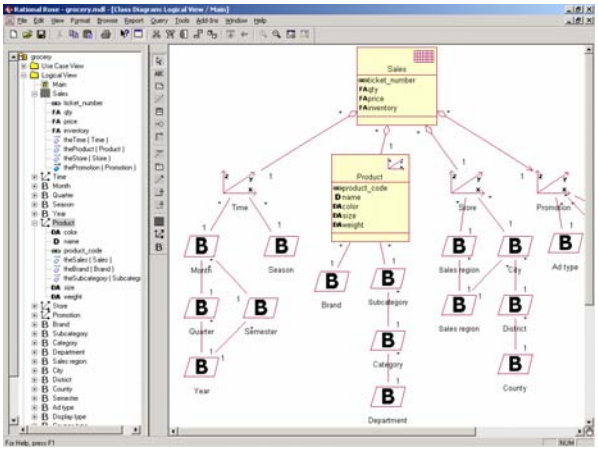




UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Diseño de AD: UML
Herramientas CASE





UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Indice

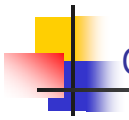
- Introducción
- Diseño de almacenes de datos: visión práctica
- Integración de fuentes: Procesos ETL
- Diseño de almacenes de datos: UML
  - Metodología global
  - Diseño del almacén de datos con UML
  - Herramientas CASE
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Indice

- Introducción
- Diseño de almacenes de datos: visión práctica
- Integración de fuentes: Procesos ETL
- Diseño de almacenes de datos: UML
- Conclusiones

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es



Conclusiones

- Presentado los almacenes de datos como el principal núcleo de los SAD actuales
- Breve descripción de una arquitectura de AD
- Diferencias principales entre sistemas transaccionales y de almacenes de datos
- Diferentes técnicas y modelos para el diseño de los almacenes de datos

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Conclusiones

- Nuestra propuesta: método de diseño global para AD
- Principales ventajas:
  - Misma notación estándar (UML) para todos los modelos
  - Perfil para modelado MD
    - Definición formal → Object Constraint Language (OCL)
  - Integración de diferentes fases de diseño en marco único
  - Se adapta a ADs grandes y complejos
    - Diagrama de paquetes
- Herramienta CASE y Rational Rose → Add-in

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Trabajos actuales/futuros

- Metodología
  - Unified Process → Model Driven Architecture (MDA)
- Métricas para modelos de AD (con grupo *ALARCOS* de Ciudad Real)
- Integrar seguridad en el modelado MD
- Integrar Data Mining y modelado MD
- Web warehouses → ¿ Cómo ?
  - Definición de consultas OLAP basadas en documentos XML
  - Web services para AD y OLAP

UJI. Castellón, 2005
Juan C. Trujillo
Jtrujillo@dlsi.ua.es

Almacenes de datos

Juan Carlos Trujillo Mondéjar

IWAD: Ingeniería del Web y Almacenes de Datos

Dpto. Lenguajes y Sistemas Informáticos
(Language and Information Systems)
Universidad de Alicante