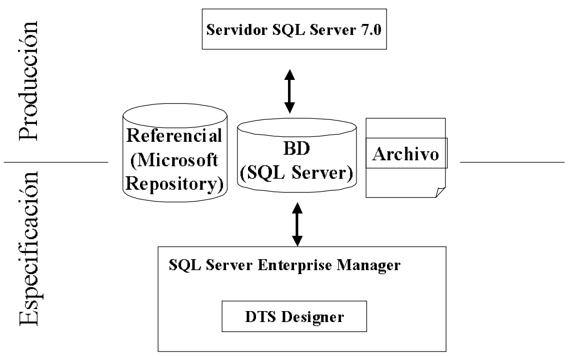
Microsoft DTS

- Se trata de una componente predefinida del RDBMS SQL Server 7.0 de Microsoft.
- Como cliente, esta componente se presenta bajo 3 formas:
  - DTS Designer
    - Asistente gráfico para la definición de los procesos (paquetes) encargados de la transformación de datos.
  - DTS Import y Export wizards
    - Asistente gráfico para la definición de paquetes más simples.
  - DTS programming interfaces (API)
    - Interfaces para ser usadas desde leng. de programación (VBasic, VC++)
- Para la ejecución y planificación
  - ServidorSQL Server 7.0 (incluyendo el servicio Agent)

Microsoft DTS (1)

- Permite el uso de un referencial para almacenar todas las definiciones.
  - Necesariamente: Microsoft Repository
  - Otras formas de almacenar las definiciones:
    - » en archivo con formato específico
    - » en SQL Server (dentro de la bd de nombre *msdb*)

DTS / Arquitectura de la herramienta



DTS / Acceso a los datos

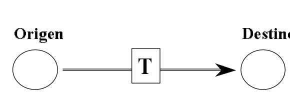
- DTS se apoya fuertemente en el acceso y almacenamiento a través de OLE DB.
  - DTS es un consumidor OLE DB
- Provee conexiones especificas para archivos de texto.

DTS / Tareas

- DTS se basa en la definición de tareas y un orden parcial entre ellas.
- La tarea básica que permite definir correspondencias y transformaciones entre la fuente de datos origen y la fuente de datos destino:
  - Transform data (data pump)
- Transform data
  - Accede y almacena datos a traves de Ole DB o archivos de texto.
  - Transforma datos entre las fuentes.
  - La transformación puede tratarse de (extremos):
    - » una simple copia entre columnas
    - » una invocación de un script (VB Script, JScript)

DTS / Transform data

Simplificando,



T:
- copia
- copia más transformaciones incluyendo funciones en un lenguaje script (VB Script, JScript).

se interpreta como:

for each o . Origen
    columnas(d) = T (columnas(o));
    insert d en Destino;
endfor

DTS / Otras tareas

- Data Driven
  - Permite realizar actualizaciones y borrados además de inserciones.
- Execute SQL
  - Permite definir un conjunto de instrucciones SQL.
- Execute Process
  - Permite invocar a un ejecutable (.exe, .bat)
- Send Mail

DTS / Otras tareas (1)

- Bulk Insert
  - Método rápido para copiar datos en archivos ascii a una bd SQL Server. No permite definición de transformaciones.
- Active X Script
  - Permite invocar un Active X script (VB Script, Perl Script, Java Script)

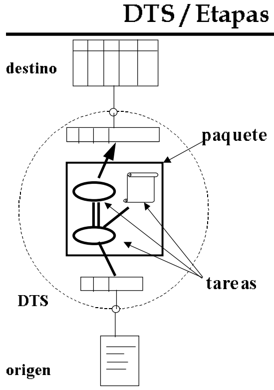
DTS / Paquete

- Un paquete es un "workflow" que define un proceso de transformación.
- Un paquete es un grafo donde:
  - los nodos son *tareas*, y
  - los arcos representan *pasos* que definen el orden en la cual se ejecutarán las tareas.
- Un paso puede tener asociado una *restricción de precedencia* definiendo cómo el resultado de una tarea determina la ejecución de la otra.
  - on success
  - on failure
  - on completion

DTS / Paquete (1)

- Posible confusión "gráfica":
  - Transform data es un nodo (tarea) dentro del workflow.

DTS / Etapas en la definición



1. Conexión
   (Acceso a las fuentes de datos)
2. Tareas
   - Importación de estructuras
3. Paquetes
4. Activación de paquetes
   - tiempo

Resumen

DTS	Genio
• Paquete	• Proceso
– Tarea	– Acción
» Transform data	» Módulo + Foreach/Forall + AddRow + datatsets
» Data driven	» Módulo + Foreach/Forall + AddRow + DeleteRow + UpdateRow + datatsets
» Execute SQL	» Módulo + Execute Procedure
» Execute Process	» Ejecutables
» Send Mail	» Send
» Bulk Insert	» *<no hay correspondiente>*
» Active X Script	» ***<no hay correspondiente>***
– Paso	. (nro, condición, acción 1, acción 2)
» restricción de precedencia (on ...)	

Resumen (1)

- DTS maneja explicitamente la sincronización
  - En Genio, se podría "simular" usando vars globales
- Genio maneja condiciones dentro del proceso
  - DTS provee las restricciones de precedencia pero no restricciones.
- Genio provee 3 mecanismos para la activación de procesos.
  - DTS ofrece activación por tiempo.
- Genio ofrece un análisis de impacto
  - DTS ???

Observación

- Las dos últimas transparencias resumen comparativamente las caracteristicas del "modelo (de WF?)" de Genio y DTS.

Conclusión

- Un ambiente de data warehousing tiene características diferentes a un ambiente de migración.
- Ambiente de data warehousing
  - Se construye el depósito de datos
  - A intervalos, se agregan cambios al depósito
  - Se archivan o se eliminan datos muy viejos
  - Se ejecutan consultas

Conclusión (1)

- A nuestro conocimiento, no hay *una* herramienta que realice o ayude a realizar todas las tareas que requiere instanciar (poblar) un data warehouse relacional.
- Variedad enfatizando algunos aspectos más que otros
  - Análisis del estado de los datos origen
  - Limpieza
  - Extracción, transformación y carga
  - Captura de cambios en los datos

Conclusion (2)

- Notar que ETLs no resuelven todas las tareas.
  - Una solución "comercial hoy" es una combinación de:
    - » herramientas de análisis del estado de los datos (calidad de los datos) a migrar,
    - » herramientas ETL,
    - » herramientas o mecanismos de captura de cambios en los datos.
  - Costo

Conclusion (3)

- La selección de herramientas es una tarea que consume tiempo no despreciable.
- Algunas sugerencias
  - (Como siempre) Importante saber lo que se quiere y se requiere.
  - Fijar un tiempo para evaluar las herramientas y no extenderlo.
  - Entender el diccionario de datos usado y contrastarlo con los ya existentes en su proyecto o los planeados a incorporar.
  - No subestimar el tiempo de analizar que tan "sucios" están los datos origen y la metainformación sobre ellos (documentación, personas).

Conclusión (4)

- **Algunas sugerencias (cont.)**
  - No subestimar el volumen y la especificación de las correspondencias.
    - » En una etapa de análisis, *identificar* las correspondencias de 300 atributos a 10 minutos/atr = 50h . 8 dias de 8h.
  - << Espacio para compartir experiencias >>

Conclusión (5)

- **Dificultad para distinguir la problemática que se resuelve bajo los titulos "Limpieza" e "Integración de datos".**