

Plan

- **Contexto - Ciclo de vida de un dw**
 - Carga inicial
 - Problemática del proceso de refreshes
 - **Carga inicial**
 - **Herramientas ETL**
 - Laboratorio
 - **Conclusión**

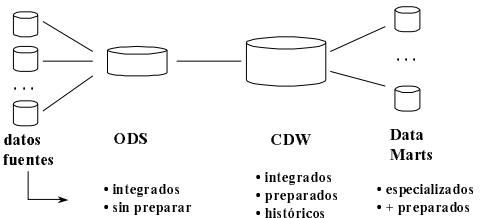
7/5/2001

InCo - Aplicación de WF para la definición de procesos de refresque de DWs

1

Contexto / Definición de dw

- El dw puede ser definido como una jerarquía de depósitos de datos [BFM99].



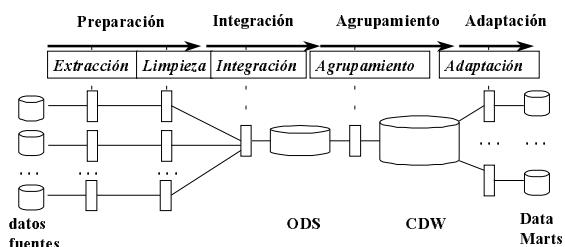
7/5/2001

InCo - Aplicación de WF para la definición de procesos de refresque de DWs

2

Contexto / Un enfoque de construcción de un dw

- Se distinguen 4 actividades en la construcción de la jerarquía de depósitos [JLLV00].



3/5/2001

Instituto de Pesquisas da WEIZSACK, definido como o desempenho da referência de DWE.

8

Herramientas ETL

Extraction, Transformation and Loading

- Características generales
 - Presentación de Hummingbird Genio
 - Presentación de Microsoft DTS

3/5/2001

¹ In *Ciencia y cultura en la definición de la cultura de referencia del DWI*, 1999.

1

Características generales

- **Objetivo principal**
 - facilitar el desarrollo y la *reutilización* de aplicaciones que migran datos aplicando *transformaciones*.
 - En este tipo de aplicaciones, los objetos típicos a definir:
 - conexiones
 - estructuras de los depósitos de datos
 - correspondencias y transformaciones entre los depósitos
 - excepciones
 - planificaciones de las transformaciones
 - informes de resultados de ejecución

3/5/2001

InG: Antecedentes de IWE para la definición de los roles de enfermería del DWE

5

Características generales (1)

- Las herramientas ETL son *ambientes especializados* que permiten la definición y manipulación de objetos típicos en aplicaciones de intercambios de datos.
 - Facilidades para la modificación y mantenimiento de las aplicaciones.
 - Generación de meta información.

7/5/2001

¹InG1: Até que o fundo WEF tenha a definição das normas de referência de DWI.

6

Características generales (2)

- Una componente importante en la arquitectura de estas herramientas es el *diccionario de datos*
 - Permite almacenar la información de los procesos de migración ofreciendo una organización centralizada.
 - De esta manera, se ofrece acceso a toda la información asociada a los procesos de migración (los objetos definidos anteriormente).
- En estas herramientas, el **data warehouse** y/o los **data marts** son vistos como depósitos adonde migrar datos transformados.

7/3/2001

InCo - Aplicación de WF para la definición de procesos de refresco de DWs

7

Características generales (3)

- En general, ETLs *NO* ofrecen funcionalidades específicas para:
 - la captura de cambios en los datos,
 - la integración de esquemas y datos
- ETLs son "pobres" en cuanto al manejo de excepciones.
 - No significa que no se puedan manejar sino que su manejo es aún "engorroso".
- Las herramientas pueden clasificarse en 3 categorías
 - "Loaders"
 - Generadores de código
 - Ambientes especializados

7/3/2001

InCo - Aplicación de WF para la definición de procesos de refresco de DWs

8

"Loaders"

- Importadores/Exportadores convencionales entre archivos ascii y Rdbms.
 - E.g. SQL*Loader de Oracle.
- Ofrecen parametrización mediante archivos de control.
 - E.g. delimitador, formato de fechas, ...
- Adecuado para cargas sin demasiadas transformaciones en los datos a partir de archivos de texto simples.
- No adecuado
 - diferentes fuentes de datos (no sólo texto)
 - transformaciones complejas
 - planificación de diferentes procesos de carga

7/3/2001

InCo - Aplicación de WF para la definición de procesos de refresco de DWs

9

Generadores de código

- Editores gráficos permitiendo definir
 - conexiones a fuentes de datos
 - transformaciones entre los datos
- Generan programas en lenguajes como Cobol, C, RPG, ABAP, ...
 - Pueden ser afinados posteriormente.
- Orientados particularmente a extracción directa en mainframes.
- El inconveniente es la gestión y coordinación de una gran cantidad de programas.
- E.g. Passport (Carleton), Warehouse Manager (Prism).

7/3/2001

InCo - Aplicación de WF para la definición de procesos de refresco de DWs

10

Ambientes especializados

- Editores gráficos para definición y planificación de procesos de carga.
 - Importar información de diccionarios de datos
 - Integración de mecanismos tales como middleware existentes (odbc, oledb, "nativo"), transferencia de archivos y envío de mails.
- Lenguajes de programación para definir las transformaciones.
 - Proveen el motor de ejecución de los programas escritos en estos lenguajes.
 - Ofrecen funciones predefinidas y permiten el agregado de funciones definidas por el usuario.
- Mecanismos para el control del flujo de los procesos.

7/3/2001

InCo - Aplicación de WF para la definición de procesos de refresco de DWs

11

Ambientes especializados (2)

- Fuertemente basados en el uso de diccionario de datos contenido los objetos involucrados en un proceso de extracción, transformación y carga.
- Inconvenientes:
 - Soporte limitado de accesos nativos a fuentes de datos en mainframes.
 - Soporte limitado de acceso a sistemas de bases de datos multidimensionales.
 - No ofrecen soporte específico para la captura de cambios en los datos.
 - Manejo elemental de errores y excepciones
- E.g. Genio (Hummingbird), DTS (Microsoft).

7/3/2001

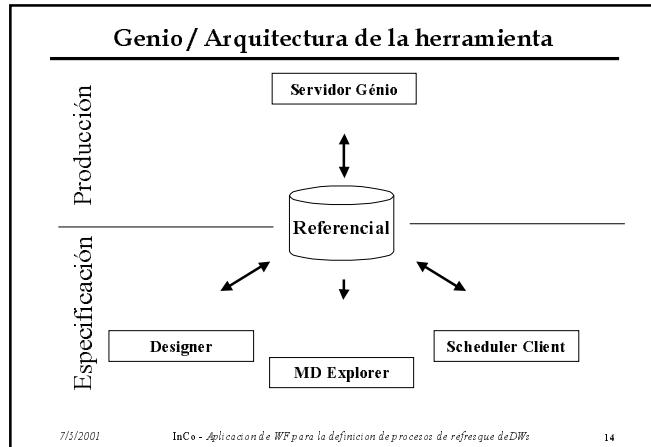
InCo - Aplicación de WF para la definición de procesos de refresco de DWs

12

Hummingbird Genio

- La base de la herramienta consiste en 4 grandes componentes software
 - Designer
 - Asistente gráfico para la definición de los procesos encargados de la transformación de datos.
 - Scheduler
 - Asistente gráfico para la planificación de los procesos definidos.
 - Metadato Explorer
 - Permite visualizar la info. en el diccionario de datos
 - Server
 - Servicios encargados de la ejecución de los procesos, control del orden de la planificación y del acceso a las fuentes de datos.
- Usa un referencial para almacenar todas las definiciones.
 - Instalado sobre un RDBMS estandard

7/3/2001 InCo - Aplicacion de WF para la definicion de procesos de refresque de DWs 13



Genio / Conexión e importación de estructuras

7/3/2001 InCo - Aplicacion de WF para la definicion de procesos de refresque de DWs 15

Genio / Módulo

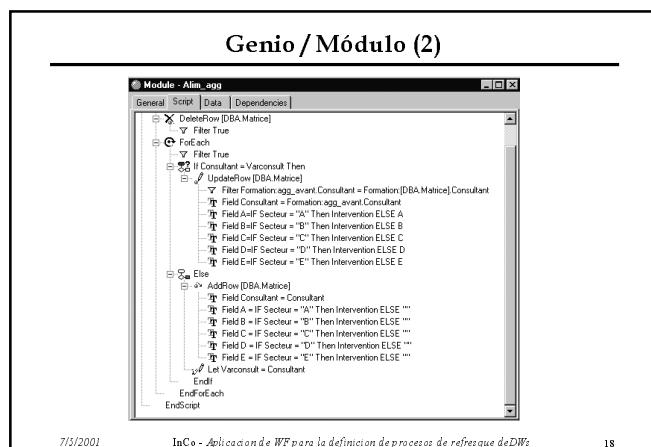
- Permite definir correspondencias y transformaciones entre la fuente de datos origen y las fuentes de datos destino.
- Genio provee un lenguaje de programación procedural parecido a Basic:
 - dim, let
 - if, for, while
 - foreach, forall
 - execute procedure
 - call module
 - write
 - raiseexception

7/3/2001 InCo - Aplicacion de WF para la definicion de procesos de refresque de DWs 16

Genio / Módulo (1)

- Instrucciones específicas para operaciones sobre datos de las fuentes destino.
 - Relacionales
 - » AddRow, SmartInsert, UpdateRow, DeleteRow.
 - Multidimensionales
 - » ClearDimension, AddMember, DeleteRow(elimina las celdas de un cubo), SetCell, RefreshCube.

7/3/2001 InCo - Aplicacion de WF para la definicion de procesos de refresque de DWs 17



Genio / Proceso

- Un proceso define una cadena de transformaciones de datos.
 - Un proceso es una secuencia de reglas de la forma:
(nro, cond, accion1, accion2)
que se lee:
*if (cond) then accion1
else accion2*
donde las acciones pueden ser:
 - » run modulo, run ejecutable (.exe, .bat),
 - » send mensaje, ftp archivo,
 - » commit/rollback,
 - » trigger eventos,
 - » goto nro | exit

7/5/2001

InCo - Aplicación de WF para la definición de procesos de refresque de DWs

19

Genio / Proceso (1)

7/3/2001

InCo - Aplicación de WF para la definición de procesos de refresque de DWs

20

Genio / Otros objetos

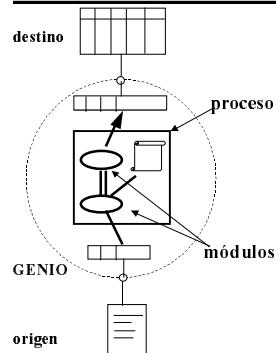
- Datasets
 - Variables globales y del sistema
 - Excepciones
 - Funciones
 - Tablas de lookup
 - Procedimientos

7/3/2001

InCo - Aplicación de WF para la definición de procesos de refresque de DWs

21

Genio / Etapas en la definición



1. Conexión
(Acceso a las fuentes de datos)
 2. Tablas Genio
(Importación de estructuras)
 3. Módulos
 4. Procesos
 5. Activación de procesos
 - evento
 - tiempo
 - presencia / modif. archivo

7/3/2001

InCo - Aplicación de WF para la definición de procesos de refresque de DWs

22

Genio / Otras funcionalidades

- Análisis de impacto frente a cambios en
 - esquemas de las fuentes,
 - objetos Genio
 - Mecanismo de sincronización entre las estructuras de tablas importadas en el referencial Genio y los servidores de BDs.
 - Histórico de cambios.
 - Logs de ejecución en el referencial Genio.

7/3/2001

InCo - Aplicación de WF para la definición de procesos de refresque de DWs

23