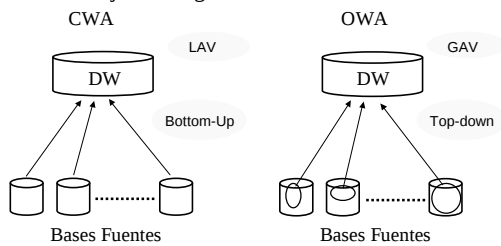


Integración en el diseño de Data Warehouses

Data Warehouse

- Abordaje de integración: materializado



Interoperabilidad - 2002

Data Warehouse

- BD con características particulares
 - información
 - para toma de decisiones
 - proviene de otras bases
 - sufrió transformaciones
 - funcionalidad
 - consultas complejas
 - grandes volúmenes de datos
 - mantenimiento de la información, no OLTP

➡ Técnicas de diseño específicas para DW

Interoperabilidad - 2002

Diseño de DW

- Estructuras que optimicen o faciliten
 - Obtención de la información de las bds operacionales en forma: *resumida, totalizada, cruzada por distintos campos y a distintos niveles*. [Kim96]
- Enfoques de diseño de DW existentes
 - 2 grupos:
 - Vistas Materializadas
 - Transformación de Esquemas ←

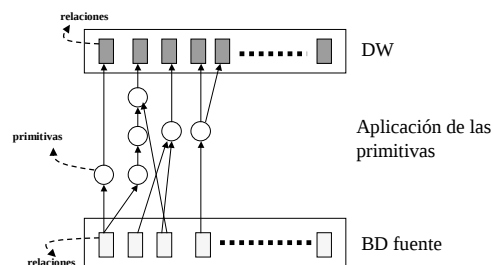
Interoperabilidad - 2002

Una propuesta desarrollada en el InCo

- Herramienta para diseñar un esquema lógico de DW a través de transformaciones. [Mar00]
 - Conjunto de **primitivas de transformación** de esquemas
 - Se obtiene la traza del diseño
 - Modelo Relacional + clasificación de sus elementos en conceptos dimensionales
 - Definición de propiedades de consistencia de un esquema de DW
 - Reglas y estrategias para aplicación de las primitivas

Interoperabilidad - 2002

Arquitectura de la Transformación



Interoperabilidad - 2002

El conjunto de Primitivas

- | | |
|--------------------------|-------------------------------|
| P1 – Identity | P8 – Hierarchy Roll Up |
| P2 – Data Filter | P9 – Aggregate Generation |
| P3 – Temporalization | P10 – Data Array Creation |
| P4 – Key Generalization* | P11 – Partition by Stability* |
| P5 – Foreign Key Update | P12 – Hierarchy Generation* |
| P6 – DD-Adding* | P13 – Minidimension Break Off |
| P7 – Attribute Adding | P14 – New dimension Crossing |

Interoperabilidad - 2002

Especif. de una Primitiva

P 9 - Aggregate Generation

- ENTRADA

- esquema origen : $R(A_1, \dots, A_n) \in Rel_C$
- Z conjunto de atributos / $card(Z) = k$ (medidas)
- $\{e_1, \dots, e_k\}$, expresiones de agregaciones
- $Y / Y \subset \{A_1, \dots, A_n\} \wedge Y \subset (Att_D(R) \cup Att_M(R))$
(atributos a eliminarse)

- ESQUEMA RESULTADO

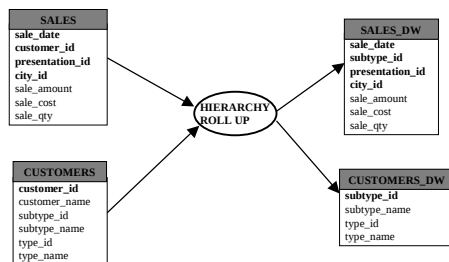
- $R'(A'_1, \dots, A'_n) \in Rel_C /$
 $\{A'_1, \dots, A'_n\} = \{A_1, \dots, A_n\} - Y \cup Z$

- TRANSFORMACION DE LOS DATOS

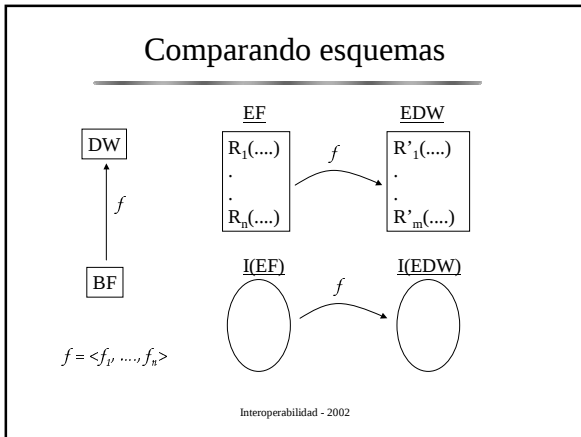
- $r' = \text{select } (\{A'_1, \dots, A'_n\} - Z) \cup \{e_1, \dots, e_k\}$
from R
group by $\{A'_1, \dots, A'_n\} - Z$

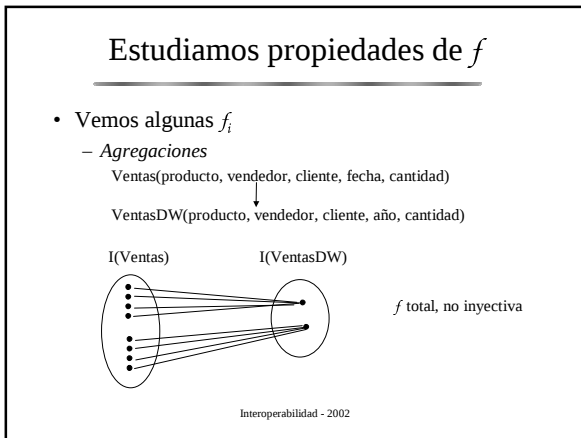
Interoperabilidad - 2002

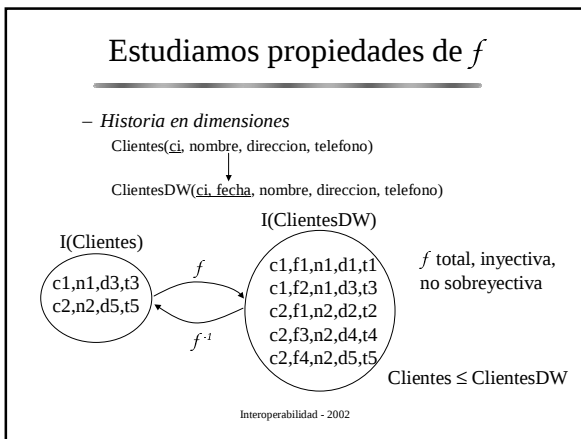
Aplicación de una Primitiva



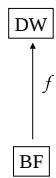
Interoperabilidad - 2002







Entonces...



- f es total, no inyectiva.
- BF y DW no son equivalentes.
- No puedo decir que DW domina a BF, ni que BF domina a DW.

Interoperabilidad - 2002

Construcción del DW a partir de múltiples fuentes

- Lo visto hasta ahora...
 - Una propuesta para diseño de DW
 - No se tiene en cuenta aspecto integración
- En adelante...
 - Como tener en cuenta en el diseño de DW el hecho de que proviene de múltiples fuentes

Interoperabilidad - 2002

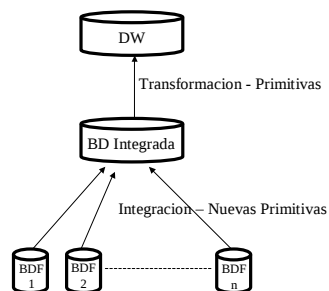
Enfoques posibles

- Enfoque 1

2 etapas

LAV

Bottom-Up



Interoperabilidad - 2002

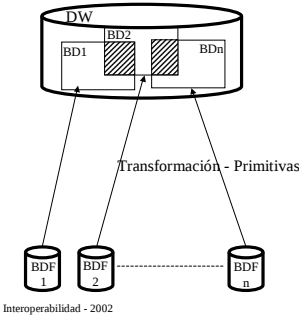
Enfoques posibles

Enfoque 2

1 etapa

GAV

Top-down



Interoperabilidad - 2002

Enfoque elegido

	Problemas integración / diseño separados	Esquema lógico de DW a priori	Se integra información completa de las fuentes	Nuevo conj. de primitivas con otro objetivo
Enfoque 1	si	no	si	si
Enfoque 2	no	si	no	no



Interoperabilidad - 2002

Ejemplo 1

BF1:
Ventas-Monteideo

Vendedor	Articulo	Fecha	Cant
V1	A1	20/2/01	20
V2	A1	14/2/01	5
V3	A2	3/3/01	30

BF2:
Ventas-Interior

Vendedor	Articulo	Fecha	Cant
V4	A1	4/2/01	10
V5	A2	5/3/01	20
V6	A2	10/3/01	4

DW:
Ventas

Articulo	Mes	Cantidad
A1	2/01	35
A2	3/01	54

Interoperabilidad - 2002

Ejemplo 1

BF1 → DW:

Artículo	Mes	Cantidad
A1	2/01	25
A2	3/01	30

BF2 → DW:

Artículo	Mes	Cantidad
A1	2/01	10
A2	3/01	24

DW:
Ventas

Artículo	Mes	Cantidad
A1	2/01	35
A2	3/01	54

✓

Interoperabilidad - 2002

Ejemplo 2

BF1:

Temperaturas1

Ciudad	Fecha	Temp
Montevideo	2/2/01	20
Montevideo	7/2/01	15
Durazno	3/3/01	30

BF2:

Temperaturas2

Ciudad	Fecha	Temp
Montevideo	20/2/01	25
Durazno	7/3/01	28
Durazno	20/3/01	30

DW:

Temperaturas

Ciudad	Mes	Temperatura
Montevideo	2/01	20
Durazno	3/01	29, 3

Interoperabilidad - 2002

Ejemplo 2

BF1 → DW:

Ciudad	Mes	Temp
Montevideo	2/01	17, 5
Durazno	3/01	30

BF2 → DW:

Ciudad	Mes	Temp
Montevideo	2/01	25
Durazno	3/01	29

DW:

Temperaturas

Ciudad	Mes	Temperatura
Montevideo	2/01	21, 25
Durazno	3/01	29, 5

X

Interoperabilidad - 2002

Ejemplo 3

BF1:

Ventas			Articulos	
Articulo	Fecha	Cant	Articulo	Tipo-art
A1	5/2/02	100	A1	T1
A1	7/2/02	50	A2	T1
A2	1/3/02	130	A3	T2
A3	2/3/02	80	A4	T2
A4	3/3/02	40		

BF2:

Articulos	
Articulo	Nac-o-Imp
A1	N
A2	I
A3	N
A4	N

DW:

Vtas-art-nacionales		
Tipo-art	Mes	Cantidad
T1	2/02	150
T2	3/02	120

Interoperabilidad - 2002

Ejemplo 3

BF1 → DW:

Se necesitan datos de BF2 para saber si el articulo es nacional *antes de agrupar*.

BF2 → DW:

No se llevan datos de BF2 directamente al DW

No puedo agrupar las ventas de articulos por tipo de articulo antes de controlar si el articulo es nacional.

⇒ No puedo dejar la integración para el final.

Interoperabilidad - 2002

Casos

- Casos posibles para un sub-esquema del DW:

- Caso 1
 - Transformacion de una sola fuente
- Caso 2
 - Transformacion de una fuente, pero con la "contribucion" de otra fuente
- Caso 3
 - Transformacion de varias fuentes
 - 2 sub-casos: se integra al final o "en el camino"

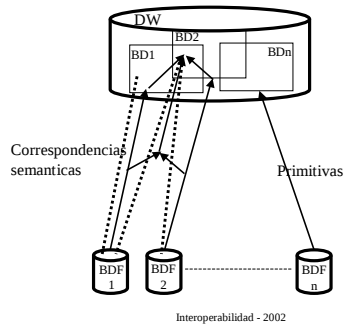
Ej. 1

Ej. 2

Ej. 3

Interoperabilidad - 2002

Generacion DW multiples fuentes



Correspondencias semanticas

- Tipos [Mot98]
 - equivalencia
 - inclusion
 - solapamiento
 - particion vertical/horizontal
 - agrupamiento
- Ejemplo 3
 - $DW.Vtas-art-nacionales \gg BF1.\{Ventas, Articulos\}$
 - $DW.Vtas-art-nacionales \subset BF1.\{Ventas, Articulos\}$
 - $f: A \rightarrow I(DW.Vtas-art-nacionales)$
 - $A = \{ e \in I(BF1.\{Ventas, Articulos\}) / g(e).Nac-o-imp = "N" \}$
 - $g: I(BF1.\{Ventas, Articulos\}) \rightarrow I(BF2.Articulos)$

Interoperabilidad - 2002

Correspondencias semanticas

- Correspondencias - Casos
 - Caso 1
 - Todas las correspondencias son con relaciones de una misma fuente.
 - Caso 2
 - Todas las correspondencias son con relaciones de una misma fuente. Para algunos calculos intervencion de otra fuente.
 - Caso 3
 - Para una misma relacion existen correspondencias con relaciones de distintas fuentes.

Interoperabilidad - 2002

Modificacion de primitivas

- Modificamos algunas primitivas
 - DD-Adding
 - Hierarchy Roll-Up
 - Hierarchy Generation
 - New dimension crossing
- Creamos nueva primitiva
 - Relation Integration

Interoperabilidad - 2002

Ej.: DD-Adding N-1

Input:

- source schema: $SDB_1.R_1(A_1, \dots, A_n), SDB_2.R_2(B_1, \dots, B_m) \in Rel$
- $f(C_1, \dots, C_k) / \{C_1, \dots, C_k\} \subseteq \{A_1, \dots, A_n\} \cup \{B_1, \dots, B_m\}$
- $A / A \in \{A_1, \dots, A_n\} \wedge B / B \in \{B_1, \dots, B_m\}$, join attributes
- is_fk, boolean argument
- $R_3 \in Rel$, relation to which A_{n+1} is a foreign key (optional)
- source instance : r_1, r_2

Resulting schema:

- $R'_1(A_1, \dots, A_n, A_{n+1}) \in Rel$ / A_{n+1} represents $f(C_1, \dots, C_k) \wedge$
if is_fk then $A_{n+1} = Att_{FK}(R'_1, R_3)$

Generated Instance:

- $r'_1 =$

```

select A1, ..., An, f(C1, ..., Ck)
from SDB1.R1 SDB2.R2
where II-Match (SDB1.R1.A, SDB2.R2.B)
            
```

Interoperabilidad - 2002

P15 – Relation Integration

Input:

- source schema: $SDB_1.R_1(A_1, \dots, A_n), SDB_2.R_2(B_1, \dots, B_m) \in Rel$
- $\langle X_1, Y_1, Z_1 \rangle, \dots, \langle X_p, Y_p, Z_p \rangle / X_i \subseteq \{A_1, \dots, A_n\} \wedge$
 $Y_i \subseteq \{B_1, \dots, B_m\} \wedge$
 $X_i \equiv Y_i$
- ii-type, instance integration type ($\cup$ or $\cap$)
- ai-type, attribute integration type ($\cup$ or $\cap$)
- source instance : r_1, r_2

Resulting schema:

- R' / if ai-type = " $\cup$ " then
 $Att(R') = (Att(SDB_1.R_1) - \cup_{i=1..p} X_i) \cup$
 $(Att(SDB_2.R_2) - \cup_{i=1..p} Y_i) \cup (\cup_{i=1..p} Z_i)$
 else if ai-type = " $\cap$ " then
 $Att(R') = \cup_{i=1..p} Z_i$

Interoperabilidad - 2002

P15 – Relation Integration

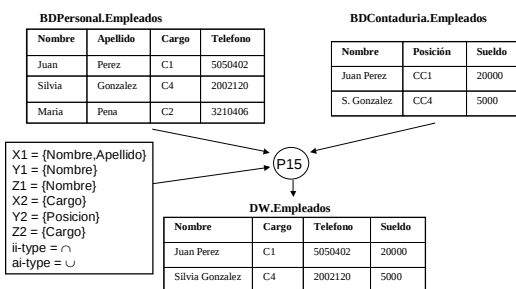
Generated Instance:

```

if ai-type = "∪" then
  if ii-type = "∪" then
    select (Att(SDB1,R1) - ∪i=1..p Xi) ∪ (Att(SDB2,R2) - ∪i=1..p Yi) ∪
      { II-Reconcile(X1,Y1), ..., II-Reconcile(Xp,Yp) }
    from SDB1,R1 SDB2,R2
    where outer-join ( II-Match (SDB1,R1, X1, SDB2,R2,Y1) and
      ..... and II-Match (SDB1,R1, Xp, SDB2,R2,Yp) )
  else if ii-type = "∩" then
    select (Att(SDB1,R1) - ∪i=1..p Xi) ∪ (Att(SDB2,R2) - ∪i=1..p Yi) ∪
      { II-Reconcile(X1,Y1), ..., II-Reconcile(Xp,Yp) }
    from SDB1,R1 SDB2,R2
    where II-Match (SDB1,R1, X1, SDB2,R2,Y1) and
      ..... and II-Match (SDB1,R1, Xp, SDB2,R2,Yp)
  else if ai-type = "∩" then
    .....
  
```

Interoperabilidad - 2002

Ejemplo: Aplicación de P15



Interoperabilidad - 2002

Integracion de Instancias

- Resolver conflictos de datos
 - Caso 2
 - Que valores “matchean”
 - Caso 3
 - Que valores “matchean” y “reconciliar” datos
- Propuesta
 - Clasificación dimensional
 - Funciones [Cal99]
 - II-Match*, *II-Convert*, *II-Reconcile*

Interoperabilidad - 2002

Bibliografía

- [Kim96] R. Kimball. *The Data Warehouse Toolkit*. J. Wiley & Sons, Inc. 1996
- [Mar00] Adriana Marotta. *Data Warehouse Design and Maintenance through Schema Transformations*. Tesis de Maestría. Oct. 2000. InCo - Pedeciba, Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay.
- [Cal99] D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi, R. Rosati. *A principled approach to data integration and reconciliation in Data Warehouses*. International Workshop on Design and Management of Data Warehouses (DMDW'99), 1999.
- [Mot98] R. Motz, P. Frankhauser. *Propagation of Semantic Modifications to an Integrated Schema*. Proc. COOPIS'98, New York, USA. August 1998.

Interoperabilidad - 2002