

Modelli logici

13.05

A.GORI - DWDM 2011/2012

1

DW – Modelli logici

- La struttura multidimensionale di un DW può essere rappresentata utilizzando tre distinti modelli logici:
 - **ROLAP**: Relational OLAP
 - **MOLAP**: Multidimensional OLAP
 - **HOLAP**: Hybrid OLAP (gestione sia di dati relazionali che multidimensionali) ← Express Server (Oracle)
 - Dati di dettaglio in ROLAP; aggregazioni in MOLAP

13.05

A.GORI - DWDM 2011/2012

2

DW – Rolap

- **ROLAP**: Relational OLAP
 - Viene usato il modello relazionale per la rappresentazione dei dati multidimensionali.
 - Scelta frequente per:
 - L'elevata diffusione e conoscenza del modello relazionale
 - L' elevata affidabilità, flessibilità del modello relazionale
 - Le prestazioni più che adeguate del modello relazionale

13.05

A.GORI - DWDM 2011/2012

3

DW – Rolap

- **ROLAP**: Relational OLAP
 - Estensione dei RDBMS
 - Lo schema dei dati Multidimensionali supportato all'interno dei RDBMS
 - Il linguaggio di accesso ai dati e le performance delle query sono ottimizzati per il trattamento di dati multidimensionali
 - Supporto per database molto grandi

13.05

A.GORI - DWDM 2011/2012

4

DW – Rolap

- Occorre riadattare il modello per gestire alcuni concetti (dimensione, misura e gerarchia) non presenti nativamente nel modello relazionale

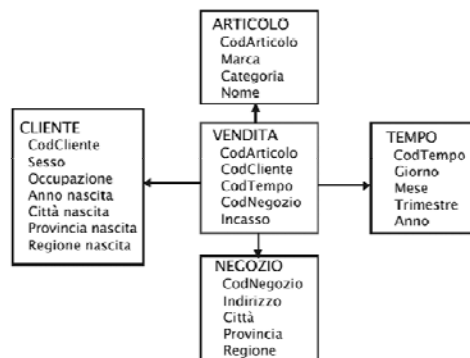
- Schema a stella (Star Schema)
- Schema a fiocco di neve (Snowflake Schema)
- Schema a Costellazione (Constellation Schema)

13.05

A.GORI - DWDM 2011/2012

5

Star Schema



13.05

5

DW – Rolap

- Problema principale: **prestazioni**
 - necessità di eseguire join tra tabelle di grosse dimensioni
- **denormalizzazione**: violazione consapevole della terza forma normale per massimizzare le prestazioni
- **ridondanza**: materializzazione di un certo numero di tabelle intermedie e derivate (viste) che offrono un valido supporto per le interrogazioni OLAP fortemente aggregate.

13.05

A.GORI - DWDM 2011/2012

7

DW – Rolap Architettura

- Un componente particolarmente importante è l'**Aggregate Navigator** che in presenza di viste aggregate si incarica di determinare, tra tutte le viste su cui una data interrogazione può essere risolta, quella che comporta il minimo costo di accesso

13.05

A.GORI - DWDM 2011/2012

8

DW – Molap

- Modello logico ad hoc per i dati e le operazioni multidimensionali
- Uso di speciali strutture dati (array multidimensionali) con accesso posizionale
- Il vantaggio di MOLAP rispetto a ROLAP è che le operazioni multidimensionali sono realizzate in modo semplice e naturale senza ricorrere a complesse operazioni di join
→ prestazioni ottimali

13.05

A.GORI - DWDM 2011/2012

9

DW – Molap

- Svantaggi:
 - Sparsità dei dati (tutte le celle di un cubo MD devono essere rappresentate ma solo una porzione di esse contengono effettivamente informazioni, le rimanenti corrispondono ad eventi non accaduti)
 - scarsa diffusione
 - necessita di molte risorse
 - poco flessibile
 - non esiste ancora uno standard

13.05

A.GORI - DWDM 2011/2012

10

Il Modello Multidimensionale

- Dimensionamento / Sparsità:
 - Le informazioni nel cubo costituiscono una sintesi (es. *il dato sintetico delle vendite di un prodotto in una giornata*) di quelle memorizzate nella base di dati operativa (dove si registrano le singole voci, ad es. le singole vendite)
 - Le informazioni possono essere comunque tantissime

13.05

A.GORI - DWDM 2011/2012

11

Cube Storage Options

- Lo spazio rappresenta un problema
 - I dati possono essere aggregati a più livelli
 - Sono duplicati (ripetuti) all'interno del cubo
 - All'aumentare delle dimensioni e misure le richieste di spazio aumentano esponenzialmente

13.05

A.GORI - DWDM 2011/2012

12

Il Modello Multidimensionale

• Dimensionamento:

– Es. di dimensionamento:

Se il DW si riferisce alle vendite degli ultimi 3 anni:

- Negozi = 50
- Prodotti = 1000
- Date = 1000 (giorni lavorativi in tre anni)

→ Numero totale di eventi = $50 \times 1000 \times 1000 = 5 \times 10^7$

N.B. Il problema della sparsità nel multidimensionale.

13.05

A.GORI - DWDM 2011/2012

13

“Data Sparsity”

Fatturato	P200	PI266	PI300	PI350	PI400	Win98	Office97	WinNT	SOL/Srv
Milano	134		100		80	45		200	200
Roma		100		234			134	120	
Napoli	80		70		30		70	80	
Torino	60	23			10			100	
Venezia	134		100		80	45		200	200
Bologna		100		234			134	120	
Aosta	80		70		30		70	80	
Trieste	60	23			10			100	
Ancona	134		100		80	45		200	200
Firenze		100		234			134	120	
Bari	80		70		30		70	80	
Palermo	60	23			10			100	
Capriari	134		100		80	45		200	200
Genova		100		234			134	120	
Reggio C.	80		70		30		70	80	

In generale, molte celle possono essere vuote, soprattutto scendendo nei dettagli

13.05

A.GORI - DWDM 2011/2012

14

Cube Storage Options

Fatturato	Hardware			Totale	Software		Totale	Totale
	Desktop	Laptop	Server		Italiano	Inglese		
Home	70	30		100	80		80	180
Business	50	15	5	70	25	5	30	100
Totale	120	45	5	170	105	5	110	280

Celle di dettaglio: 10 (di cui 2 vuote) (A)

Celle di sintesi: 14! (B)

Celle complessive: 24 (C)

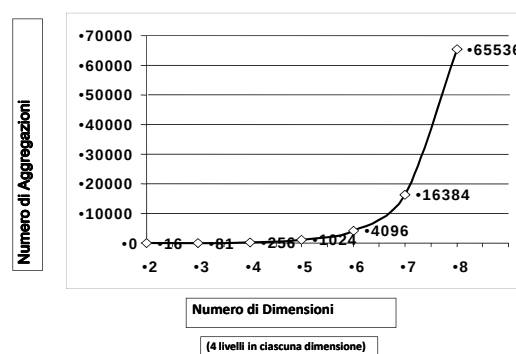
Rapporto (C)/(A): 2.4!

13.05

A.GORI - DWDM 2011/2012

15

Cube Storage Options



13.05

A.GORI - DWDM 2011/2012

16

Cube Storage Options

- **IMPORTANTE !!!**
- Aggregare al 100 % non e' necessario:
 - Aumenta tempo di processo e richieste di spazio
 - Regola 80/20

13.05

A.GORI - DWDM 2011/2012

17

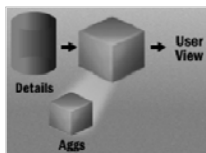
ROLAP vs. MOLAP

Schema	Uses star schema Additional dimensions can be added dynamically.	Uses data cubes Additional dimensions require re-creation of the data cube
Architecture	Client/server Standards-based	Client/server Proprietary
Access	Supports ad hoc requests Unlimited dimensions	Limited to predefined dimensions
Resources	High	Very high
Flexibility	High	Low
Scalability	High	Low
Speed	Good with small data sets; average for medium to large data sets	Faster for small to medium data sets; average for large data sets

13.05

•HOLAP - Hybrid OLAP : il compromesso

- La via di mezzo
 - Mantiene i fatti in RDBMS
 - Le aggregazioni sono nello store MOLAP
- Evita la duplicazione dei dati
- Si perde in termini di prestazioni quando si deve accedere ai dettagli
- Tempi di processo simili a MOLAP



13.05

A.GORI - DWDM 2011/2012

19

Il problema della eterogeneità

13.05

A.GORI - DWDM 2011/2012

20

Eterogeneità

- L'eterogeneità occorre spesso quando vi è la necessità di integrare e rendere interoperabili sistemi che sono preesistenti e che sono nati per vivere autonomamente
- Esempi:
 - i nodi della rete usano DBMS differenti (spesso diversi anche come modello dei dati)
 - Ogni DBMS ha le proprie caratteristiche, le proprie strategie di implementazione (es. gestione della concorrenza, dei blocchi critici, del ripristino)
 -

13.05

A.GORI - DWDM 2011/2012

21

Eterogeneità

- L'eterogeneità quindi può manifestarsi a vari livelli:
 - Piattaforme hardware diverse
 - Sistemi operativi diversi
 - Protocolli di rete diversi
 - Applicazioni scritte in diversi linguaggi
 - il software che crea e manipola i dati nei vari siti è diverso
 - Diversi linguaggi di interrogazione
 - Diversi DBMS – diversi modelli dei dati
 - il formato e la struttura dei dati nei vari siti sono diversi
 - Diversità semantiche

13.05

A.GORI - DWDM 2011/2012

22

Heterogeneity

Database
DBMSs Semantic Heterogeneity
Operating System
File Systems Interprocess communication Level of user access Distributed processing capabilities
Communication
Network topology Networking protocols Physical layer differences
Hardware/System
Instruction Set Data formats and representation ...

13.05

23

Eterogeneità Semantica

Esempi

13.05

A.GORI - DWDM 2011/2012

24

Semantic Heterogeneity Omonimi

Esempio:

- Un DDB è costituito da due DB, DB-A nel nodo A e DB-B nel nodo B.
- Nei Db esiste una relazione RESTAURANT.
- L'attributo MEAL_COST in DB-A descrive il costo medio di un pasto per persona *without* tax.
- L'attributo MEAL_COST in DB-B descrive il costo medio di un pasto per persona *including* tax.

13.05

A.GORI - DWDM 2011/2012

25

Semantic Heterogeneity Omonimi

- Trovare tutti i ristoranti con un price range \$10-\$20.

```
SELECT NAME
FROM RESTAURANTS
WHERE MEAL_COST >= 10
AND MEAL_COST <= 20
```



La query produce risultati di significato diversi.

13.05

A.GORI - DWDM 2011/2012

26

Semantic Heterogeneity Sinonimi

- I ristoranti sono classificati in base al tipo di piatti che servono.
- Nel DB-A questa categoria è gestita nell'attributo TYPE.
- Nel DB-B questa categoria è gestita nell'attributo CLASSIFICATION.

- Trovare tutti i ristoranti greci:

```
SELECT NAME      SELECT NAME
FROM RESTAURANTS FROM RESTAURANTS
WHERE TYPE="Greek" WHERE CLASSIFICATION="Greek"
```

13.05

A.GORI - DWDM 2011/2012

27

Semantic Heterogeneity Granularità

- I ristoranti sono classificati in base alla qualità dei vini serviti.
- In entrambi i DB la categoria è memorizzata nell'attributo QUALITY.
- In DB-A, la qualità è data su una scala di tre valori:
eccellente, buono, scadente.
- IN DB-B, la qualità è data su una scala di 5 valori:
1, 2, 3, 4, 5.

13.05

A.GORI - DWDM 2011/2012

28

Semantic Heterogeneity Granularità

- Produrre la classifica dei ristoranti in base alla qualità del vino.
- Possiamo fare una classifica separata per DB e poi produrne una unica?

13.05

A.GORI - DWDM 2011/2012

29

Heterogeneity Differenze Strutturali

- L'indirizzo di ogni ristorante è memorizzato nel DB.
- In DB-A, l'indirizzo completo è memorizzato nell' attributo ADDRESS della relazione RESTAURANTS.
- In DB-B, l'indirizzo appare in una relazione separata: ADDRESS (NAME, ADDRESS, STATE, ZIPCODE).
- Es.
Trovare tutti i ristoranti della zona con zipcode 08851

13.05

A.GORI - DWDM 2011/2012

30

Heterogeneity Domini differenti

- Nel DB-A lo zipcode può essere letto come una stringa nell'attributo ADDRESS.
- Nel DB-B lo zipcode è memorizzato come dato numerico.

13.05

A.GORI - DWDM 2011/2012

31

Modello dei dati

- Occorrerebbe sviluppare un singolo schema utilizzando un modello dei dati comune
(per esempio, si può scegliere il modello relazionale)
- Relazione: RESTAURANTS
(NAME, MEAL_COST_BEFORE_TAX,
MEAL_COST_AFTER_TAX, TYPE,
QUALITY, ADDRESS)

13.05

A.GORI - DWDM 2011/2012

32

Data dictionary

- Per risolvere i problemi dovuti a:
 - Omonimi
 - Sinonimi
 - Granularità
 - Differenze Strutturali
 - Domini
- Un data dictionary fornisce le informazioni su:
 - Schema unificato
 - Regole di traduzione tra gli schemi

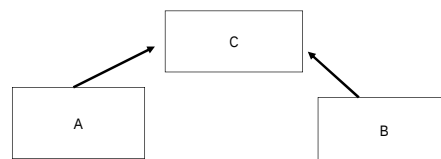
13.05

A.GORI - DWDM 2011/2012

33

Global schema integration

Viene prodotto un nuovo schema (C) frutto della omogeneizzazione dei differenti schemi di A e di B.



13.05

A.GORI - DWDM 2011/2012

34

Global schema integration

Vantaggi

Trasparenza per l'utente finale: tutto appare come un unico sistema

Svantaggi

- Difficile da implementare: richiede l'intervento umano per effettuare l'integrazione
- Perdita di autonomia locale
- Staticità: non evolve automaticamente → gestione della manutenzione

13.05

A.GORI - DWDM 2011/2012

35

Data dictionary

Esempio

- Global schema DB-A:
 - MEAL_COST_BEFORE_TAX=MEAL_COST
 - MEAL_COST_AFTER_TAX=MEAL_COST*1.20
- Global schema DB-B:
 - MEAL_COST_BEFORE_TAX=MEAL_COST * 0.833
 - MEAL_COST_AFTER_TAX =MEAL_COST
 - TYPE=CLASSIFICATION
 - QUALITY={
 - Excellent → 1 or 2
 - Good → 3 or 4
 - Poor → 5}
 - ADDRESS=ADDRESS | STATE | ZIPCODE

13.05

A.GORI - DWDM 2011/2012

36

Il processo di integrazione

13.05

A.GORI - DWDM 2011/2012

37

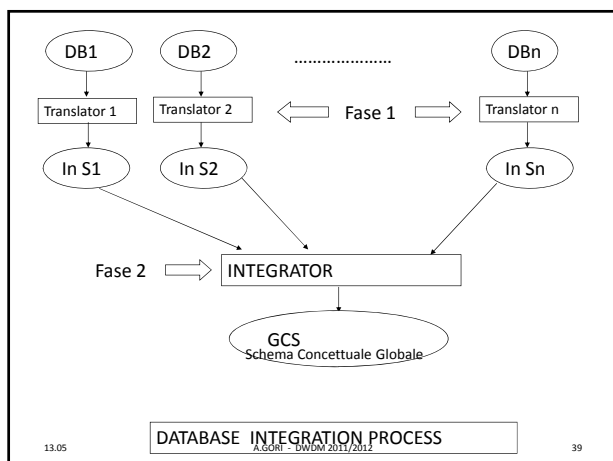
Database integration: 2 passi

Il processo di integrazione avviene in due passi:

1. Schema Translation: gli schemi locali dei db partecipanti sono tradotti in una rappresentazione intermedia canonica.
2. Schema Integration: ogni schema intermedio viene integrato in uno schema concettuale globale.

13.05

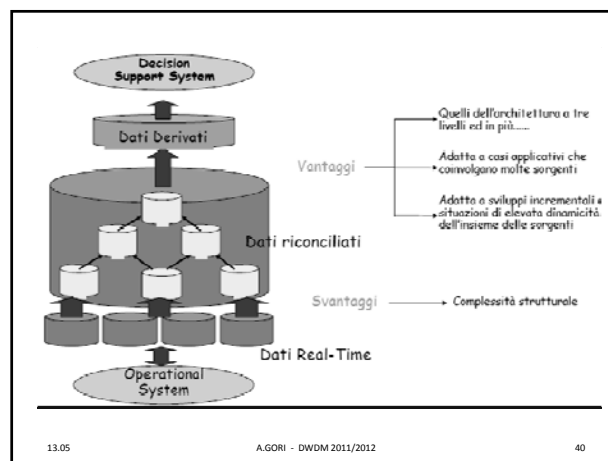
A.GORI - DWDM 2011/2012



13.05

A.GORI - DWDM 2011/2012

39



13.05

A.GORI - DWDM 2011/2012

40

Schema Translation

- ◆ Viene richiesta la specifica di un comune modello di dati per il global conceptual schema.
- ◆ Relational data model e l' E-R data model sono i modelli di dati più frequentemente utilizzati.

13.05

A.GORI - DWDM 2011/2012

Schema Translation

- Molto importante è la scelta del modello canonico ← dovrebbe essere quello che è maggiormente in grado di incorporare i concetti dei singoli DBMS che saranno integrati.
- In realtà, spesso si tende ad utilizzare il modello che è più congeniale o di cui si ha maggior conoscenza

13.05

A.GORI - DWDM 2011/2012

Schema Translation

- Ovviamente questo passo è necessario solo se gli schemi locali dei singoli DBMS sono di un modello diverso.
- Gli schemi dei DB componenti vengono tradotti in una forma canonica intermedia, ossia in una forma che tenda a standardizzare la rappresentazione
→ ciò facilita il processo di traduzione.

13.05

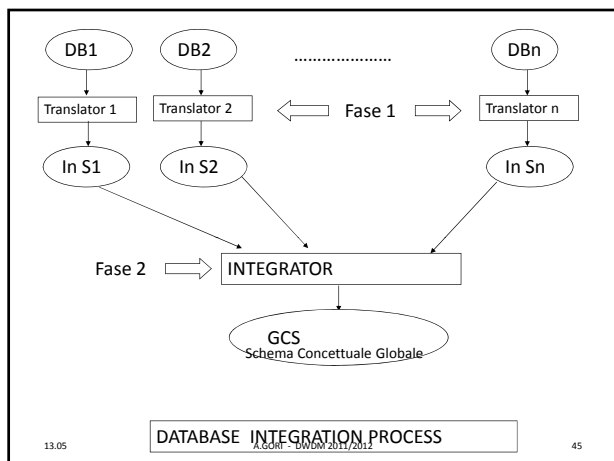
A.GORI - DWDM 2011/2012

Schema Integration

13.05

A.GORI - DWDM 2011/2012

44



Schema Integration

- Fase che segue il processo di traduzione e genera il GCS a partire dagli schemi intermedi.
- Si divide in due sottofasi.....→

13.05

A.GORI - DIWIM 2011/2012

Schema Integration

- Sottofasi:
 - Omogeneizzazione
 - Soluzione dei conflitti strutturali e semantici
 - Integrazione
 - Merging e ristrutturazione degli schemi "omogeneizzati"



Global Conceptual Schema

13.05

A.GORI - DIWIM 2011/2012

Omogeneizzazione

- Con la fase di omogeneizzazione vengono risolti problemi di eterogeneità semantica e strutturale.

13.05

A.GORI - DIWIM 2011/2012

Omogeneizzazione

- Eterogeneità semantica:

- NAMING CONFLICTS:

- Sinonimi: una stessa entità individuata con nomi diversi.
 - Esempio: Employee ed E usati per il nome del dipendente.
- Omonimi: due entità diverse con lo stesso nome.

13.05

A.GORI - DWSIM 2011/2012

Omogeneizzazione

- NAMING CONFLICTS:

- Ci sono vari metodi alternativi per trattare questo problema

- Esempio:

- I sinonimi sono risolti semplicemente rinominando.
- Gli omonimi si risolvono:
 - Prefissando ogni attributo col nome dell'entità a cui appartiene
 - Prefissando ogni entità col nome dello schema cui appartiene.

13.05

A.GORI - DWSIM 2011/2012

Omogeneizzazione

- Conflitti strutturali (1)

- Type conflict (1):

- quando lo stesso oggetto è rappresentato da un attributo in uno schema e da un'entità in un altro

- Type conflict (2):

- quando lo stesso oggetto è rappresentato da attributi di tipologia diversa nei vari schemi

- Dependency conflict:

- Quando sono usate relationship con cardinalità diversa (es. one-to-one vs. many-to-many) per rappresentare la stessa cosa in schemi differenti

13.05

A.GORI - DWSIM 2011/2012

Omogeneizzazione

- Conflitti strutturali (2)

- Key conflict:

- Quando ci sono chiavi candidate diverse e quando sono scelte chiavi primarie diverse nei diversi schemi

- Behavioral conflict:

- Insiti nel meccanismo di modellazione (es. diversità nei vincoli di integrità: l'eliminazione di una tupla può determinare la dissoluzione di relazioni collegate)

13.05

A.GORI - DWSIM 2011/2012

Integrazione

- Fusione degli schemi intermedi in un unico "database schema" e successiva ristrutturazione per ottenere il "migliore" schema integrato.
- La fusione richiede che le informazioni contenute negli schemi partecipanti siano mantenute nello schema integrato.

13.05

A.GORI - DWSIM 2011/2012

Integrazione

- Merging e Ristrutturazione
 1. Tutti gli schemi vengono fusi in un unico schema e poi ristrutturati per creare lo schema integrato migliore.
 2. Requisiti da rispettare:
 - Completezza: tutte le informazioni provenienti dagli schemi sono integrate nello schema comune.
 - Minimalità: nessuna ridondanza viene prodotta nello schema integrato.
 - Comprensibilità: dopo che si è effettuato il merge di tutti gli elementi, la ristrutturazione deve facilitare la creazione di uno schema comprensibile.

13.05

A.GORI - DWSIM 2011/2012

Esempio

13.05

A.GORI - DWSIM 2011/2012

55

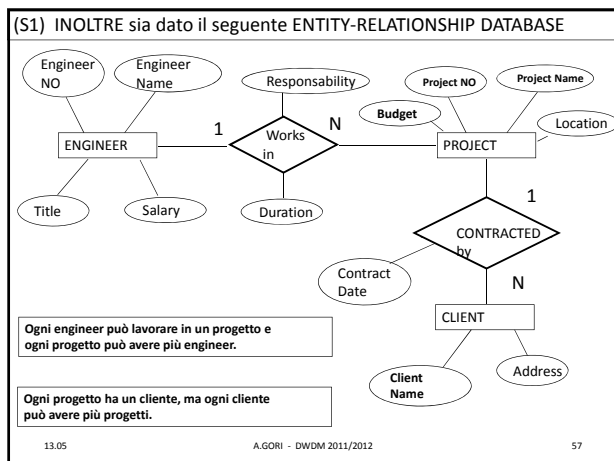
Database Relazionale

(S2) Sia dato il seguente database relazionale:

EMP (ENO, ENAME, TITLE)
 PROJ (PNO, PNAME, BUDGET, LOC, CNAME)
 ASG (ENO, PNO, RESP, DUR)
 PAY (TITLE, SAL)

13.05

A.GORI - DWSIM 2011/2012



Processo

1. Fase: SCHEMA TRANSLATION
 - Traduzione
2. FASE: SCHEMA INTEGRATION
 1. Omogeneizzazione
 2. Integrazione

13.05

A.GORI - DWM 2011/2012

Schema Translation

13.05

A.GORI - DWM 2011/2012

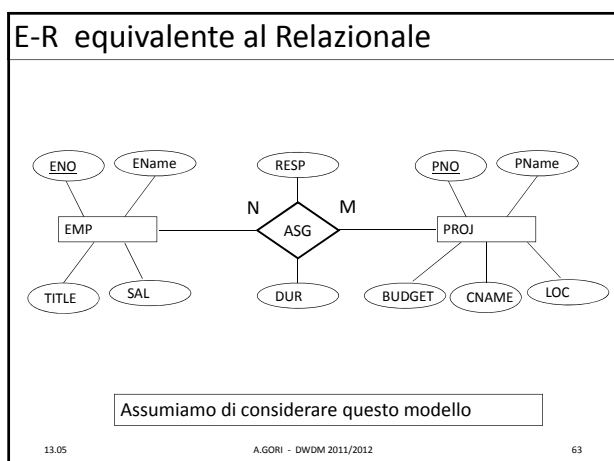
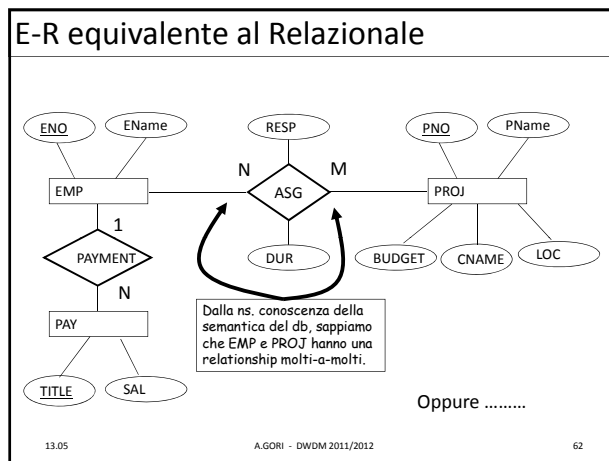
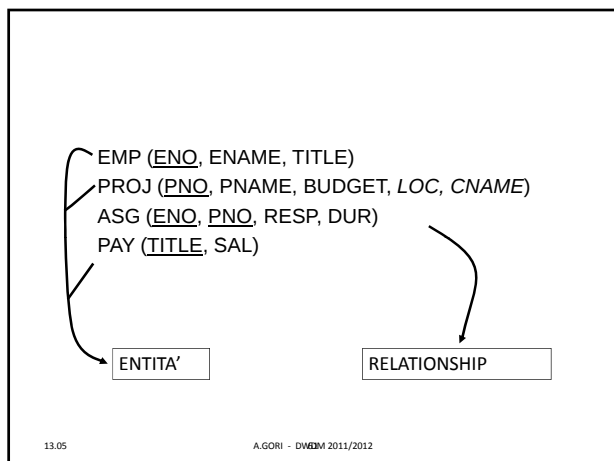
59

Traduzione dello schema relazionale nel modello E-R

- Determinare le relazioni che rappresentano entità
- Determinare le relazioni che rappresentano relationship.
- Identificare la cardinalità delle relationship
- Identificare vincoli sulle relationship
-

13.05

A.GORI - DWM 2011/2012



OMOGENEIZZAZIONE Conflitti semantici

13.05

A.GORI - DWM 2011/2012

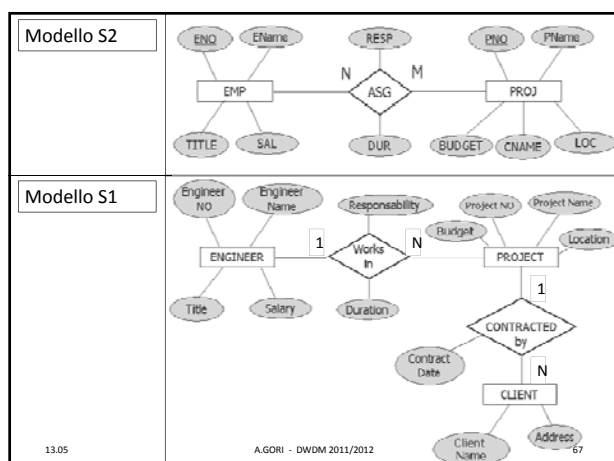
65

Conflitti semantici

- I naming conflicts dell'esempio possono essere risolti rinominando
 - Entità
 - attributi
 - relationships

13.05

A.GORI - DWM 2011/2012



13.05

A.GORI - DWM 2011/2012

A.GORI - DWM 2011/2012

Sinonimi

nello SCHEMA INTERMEDIO

InS ₁	InS ₂
ENGINEER	EMP
Engineer No	ENO
Engineer Name	ENAME
Salary	SAL
WORKS IN	ASG
Responsability	RESP
Duration	DUR
PROJECT	PROJ
Project No	PNO
Project Name	PNAME
Location	LOC

13.05

A.GORI - DWM 2011/2012

Omonimi

nello SCHEMA INTERMEDIO

InS ₁	InS ₂
ENGINEER	EMP
Title	Title

In InS₁, Title si riferisce al titolo degli Engineers mentre in InS₂ si riferisce al titolo di tutti gli impiegati per cui ha un dominio più ampio

→ caso di omonimia: lo stesso nome di attributo si riferisce a due cose diverse.

domain (EMPLOYEE.Title) >> domain (ENGINEER.Title)

13.05

A.GORI - DWM 2011/2012

- Una possibile soluzione alle omonimie è:
 - prefissare ogni attributo col nome dell'entità a cui appartiene
 - prefissare ogni entità col nome dello schema cui appartiene.

Nel ns. esempio potremmo rinominare lo schema relativo al db relazionale (S2) per renderlo conforme alla nomenclatura dello schema del db E-R (S1)

13.05

A.GORI - DWM 2011/2012

OMOGENEIZZAZIONE: Conflitti strutturali

13.05

A.GORI - DWM 2011/2012

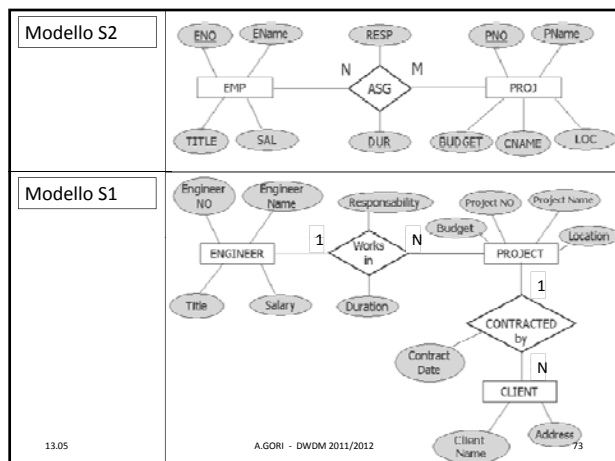
71

Conflitti strutturali

- Un modo di risolverli è tramite le trasformazioni da compiere su:
 - Entità
 - Attributi
 - Relationships

13.05

A.GORI - DWM 2011/2012



Conflitti strutturali (esempio)

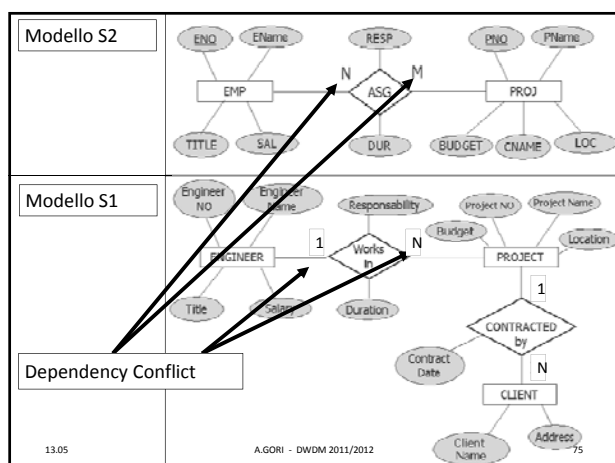
DEPENDENCY CONFLICT:

Nell'esempio, la relationship del modello E-R è multi-a-uno mentre nel relazionale è multi-a-molti.

→

13.05

A.GORI - DWDM 2011/2012



Conflitti strutturali (esempio)

DEPENDENCY CONFLICT:

Nell'esempio, la relationship del modello E-R è multi-a-uno mentre nel relazionale è multi-a-molti.

→ *Decisione di progetto:*
si sceglie di accettare la relationship più generale multi-a-molti.

13.05

A.GORI - DWDM 2011/2012

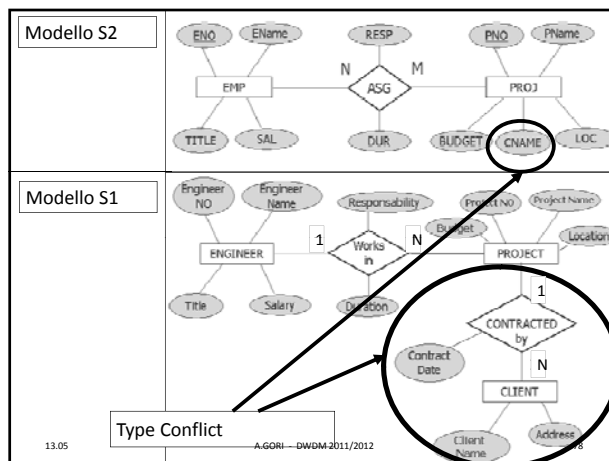
Conflitti strutturali (esempio)

TYPE CONFLICT:

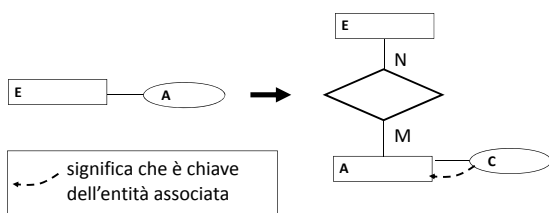
Nell'esempio, nel modello E-R, il Cliente di un progetto è modellato come una entità mentre, nel modello relazionale, è un attributo dell'entità PROJ.

13.05

A.GORI - DWDM 2011/2012



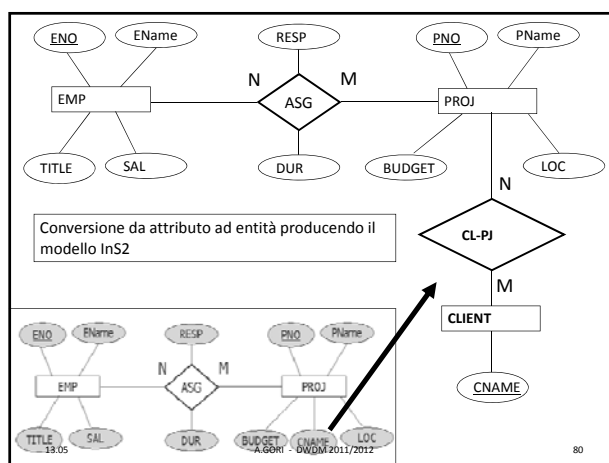
Modello di conversione da attributo ad entità (1)



13.05

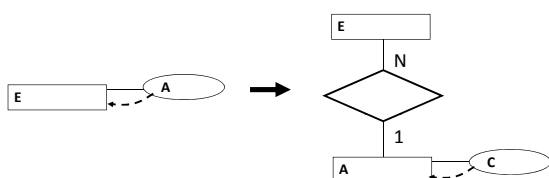
A.GORI - DWDM 2011/2012

Un attributo A, non chiave, può essere trasformato in una nuova entità. Tra la nuova entità e quella di partenza esiste una relationship multi-a-multi su di un nuovo attributo (C).



80

Modello di conversione da attributo ad entità (2)



Un attributo chiave può essere trasformato in una nuova entità; tra la nuova e l'originale entità esiste una relationship multi-a-uno tramite un nuovo attributo (C).

13.05

A.GORI - DVBIM 2011/2012

Come correlare i due schemi

- Come sono correlati tra loro gli schemi intermedi cui siamo giunti?
- Ci sono le seguenti possibilità:
 - Sono identici l'un l'altro
 - Uno è subset dell'altro
 - Hanno alcuni componenti in comune
 - Sono completamente disgiunti
- Nel ns. esempio abbiamo che InS2 è un subset di InS1.

13.05

A.GORI - DVBIM 2011/2012

Come correlare i due schemi

- Abbiamo visto che la fase di omogeneizzazione richiede una buona dose di intervento umano dato che è richiesta la conoscenza semantica di tutti gli schemi intermedi.

13.05

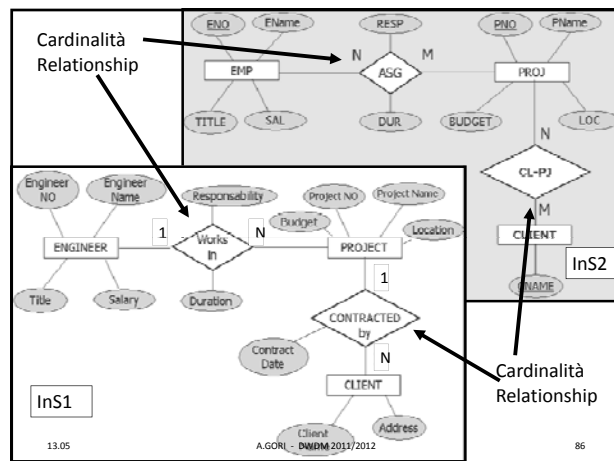
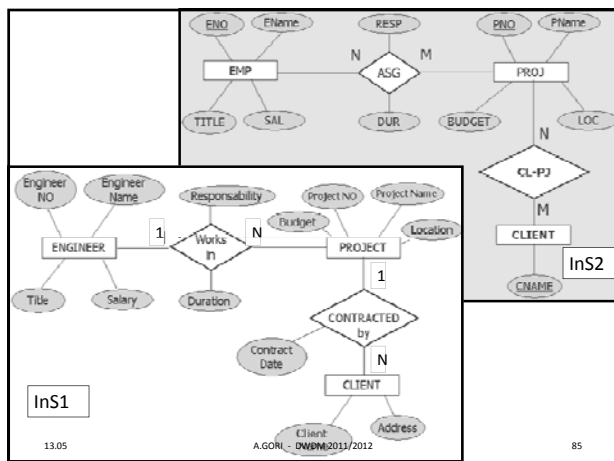
A.GORI - DVBIM 2011/2012

Integrazione

- Fusione degli schemi intermedi in un unico "database schema" e successiva ristrutturazione per ottenere il "migliore" schema integrato.
- La fusione richiede che le informazioni contenute negli schemi partecipanti siano mantenute nello schema integrato.

13.05

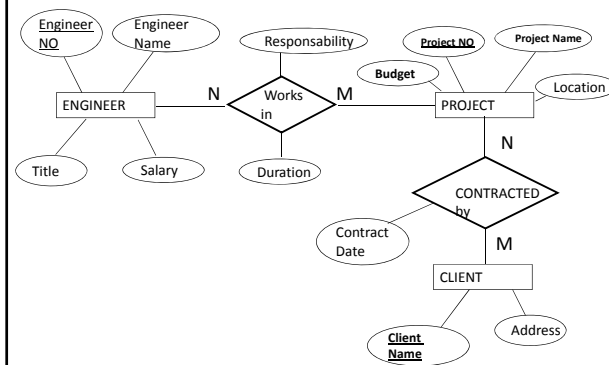
A.GORI - DVBIM 2011/2012



Esempio

- Nel ns. esempio, essendo InS2 subset di InS1
→ InS1 diviene lo schema integrato.
- **Unica complicazione:** le diverse caratteristiche della relationship CONTRACTED-BY (in InS1 è di tipo multi-a-uno; in InS2, è di tipo multi-a-molti che è più generale).
→ La fusione può compiersi accettando InS1 come risultato MA disponendo che la relationship CONTRACTED-BY sia convertita in una di tipo multi-a-molti.

GLOBAL CONCEPTUAL SCHEMA INTEGRATO



Merging e Ristrutturazione

- Requisiti:
 - **Completezza**: tutte le informazioni provenienti dagli schemi sono integrate nello schema comune.
 - **Minimalità**: nessuna ridondanza viene prodotta nello schema integrato.
 - **Comprensibilità**: dopo che si è effettuato il merge di tutti gli elementi, la ristrutturazione deve facilitare la creazione di uno schema comprensibile.