Data warehouse
Progettazione

Tania Cerquitelli
Politecnico di Torino

Elena Baralis
Politecnico di Torino

Fattori di rischio

- Aspettative elevate degli utenti
  - il data warehouse come soluzione dei problemi aziendali
- Qualità dei dati e dei processi OLTP di partenza
  - dati incompleti o inaffidabili
  - processi aziendali non integrati e ottimizzati
- Gestione “politica” del progetto
  - collaborazione con i “detentori” delle informazioni
  - accettazione del sistema da parte degli utenti finali

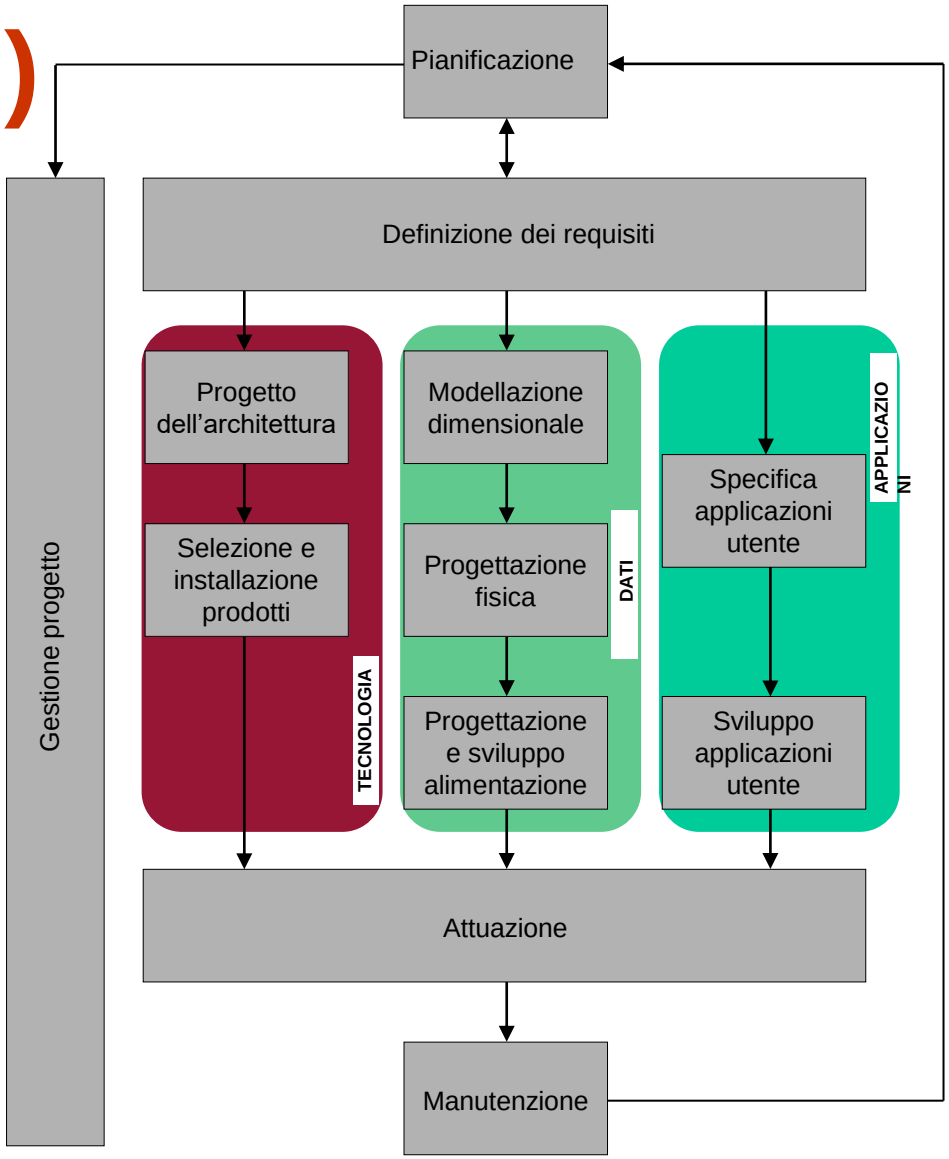
Elena Baralis
Politecnico di Torino

Progettazione di data warehouse

- Approccio top-down
  - realizzazione di un data warehouse che fornisca una visione globale e completa dei dati aziendali
  - costo significativo e tempo di realizzazione lungo
  - analisi e progettazione complesse
- Approccio bottom-up
  - realizzazione incrementale del data warehouse, aggiungendo data mart definiti su settori aziendali specifici
  - costo e tempo di consegna contenuti
  - focalizzato separatamente su settori aziendali specifici

Elena Baralis
Politecnico di Torino

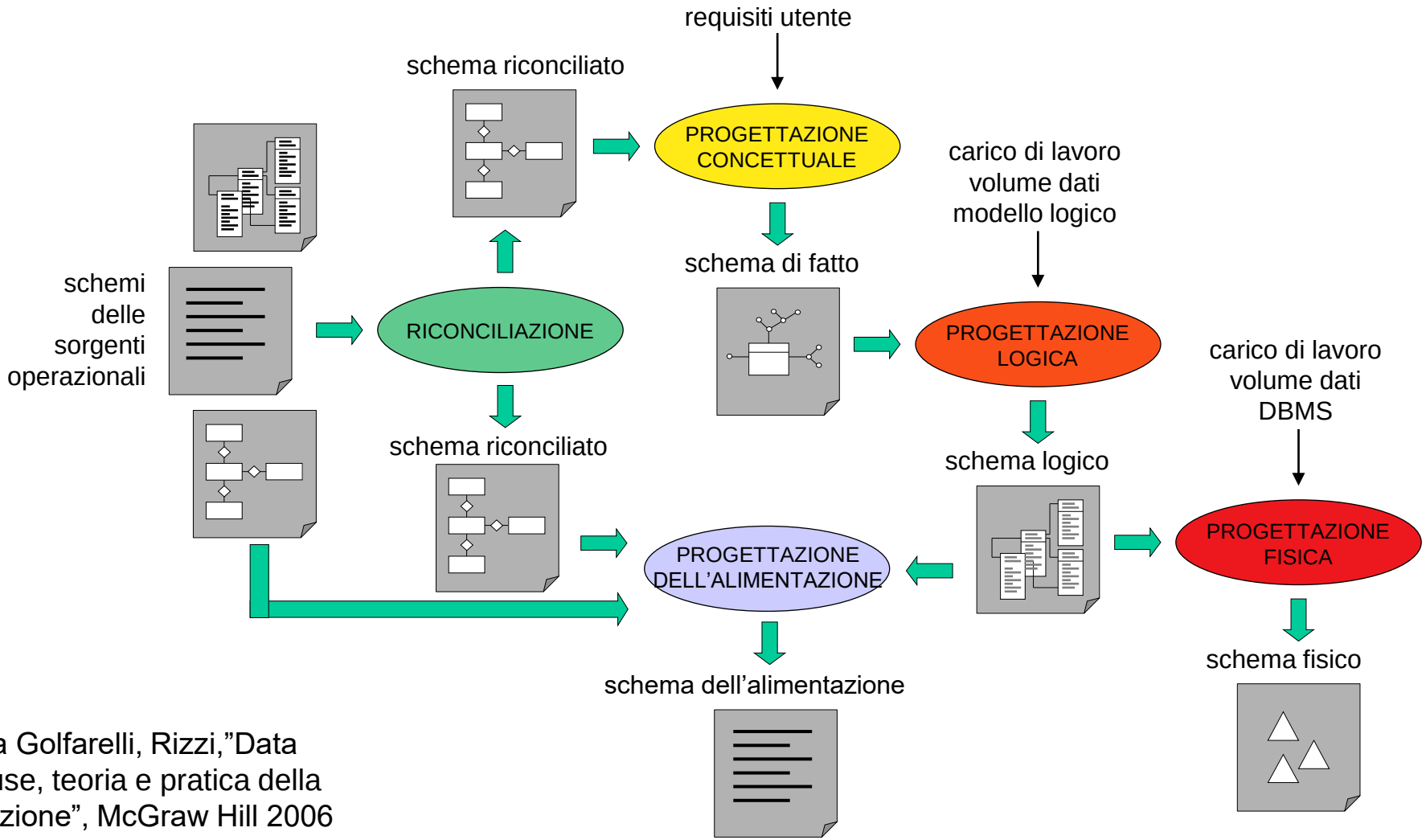
Business Dimensional Lifecycle (Kimball)



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Progettazione di data mart



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Analisi dei requisiti

Tania Cerquitelli
Politecnico di Torino

Elena Baralis
Politecnico di Torino

Analisi dei requisiti

- Raccoglie
  - le esigenze di analisi dei dati che dovranno essere soddifatte dal data mart
  - i vincoli realizzativi dovuti ai sistemi informativi esistenti
- Fonti
  - business users
  - amministratori del sistema informativo
- Il data mart prescelto è
  - strategico per l'azienda
  - alimentato da (poche) sorgenti affidabili

Elena Baralis
Politecnico di Torino

Requisiti applicativi

- Descrizione degli eventi di interesse (fatti)
  - ogni fatto rappresenta una categoria di eventi di interesse per l'azienda
    - esempi: (per il CRM) reclami, servizi
  - caratterizzati da dimensioni descrittive (granularità), intervallo di storicizzazione, misure di interesse
  - informazioni raccolte in un glossario
- Descrizione del carico di lavoro
  - esame della reportistica aziendale
  - interrogazioni espresse in linguaggio naturale
    - esempio: numero di reclami per ciascun prodotto nell'ultimo mese

Elena Baralis
Politecnico di Torino

Requisiti strutturali

- Periodicità dell'alimentazione
- Spazio disponibile
  - per i dati
  - per le strutture accessorie (indici, viste materializzate)
- Tipo di architettura del sistema
  - numero di livelli
  - data mart dipendenti o indipendenti
- Pianificazione del deployment
  - avviamento
  - formazione

Elena Baralis
Politecnico di Torino

Progettazione concettuale

Tania Cerquitelli
Politecnico di Torino

Elena Baralis
Politecnico di Torino

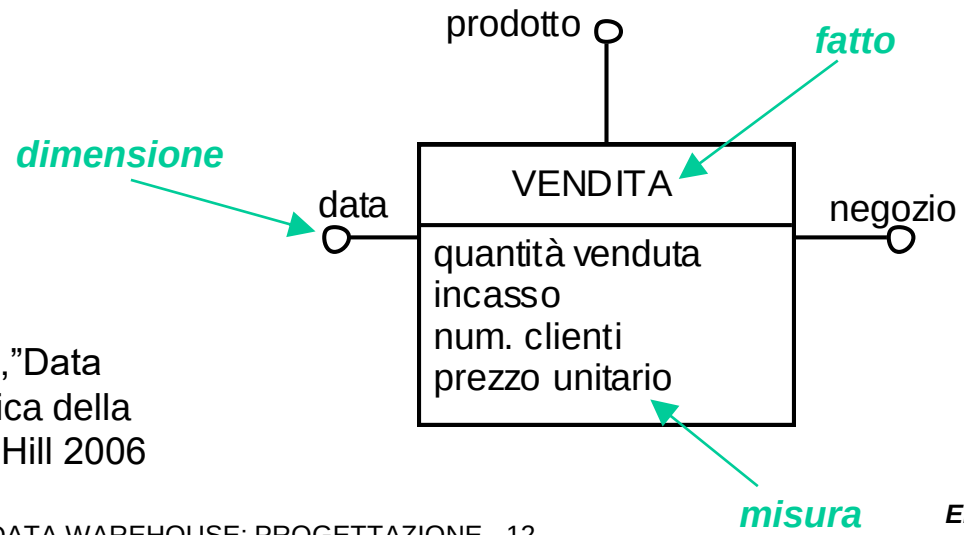
Progettazione concettuale

- Non esiste un formalismo di modellazione comunemente accettato
  - il modello ER non è adatto
- Dimensional Fact Model (Golfarelli, Rizzi)
  - per uno specifico fatto, definisce schemi di fatto che modellano
    - dimensioni
    - gerarchie
    - misure
  - modello grafico a supporto della progettazione concettuale
  - offre una documentazione di progetto utile sia per la revisione dei requisiti con gli utenti, sia a posteriori

Elena Baralis
Politecnico di Torino

Dimensional Fact Model

- Fatto
  - modella un insieme di eventi di interesse (vendite, spedizioni, reclami)
  - evolve nel tempo
- Dimensione
  - descrive le coordinate di analisi di un fatto (ogni vendita è descritta dalla data di effettuazione, dal negozio e dal prodotto venduto)
  - è caratterizzata da numerosi attributi, tipicamente di tipo categorico
- Misura
  - descrive una proprietà numerica di un fatto, spesso oggetto di operazioni di aggregazione (ad ogni vendita è associato un incasso)

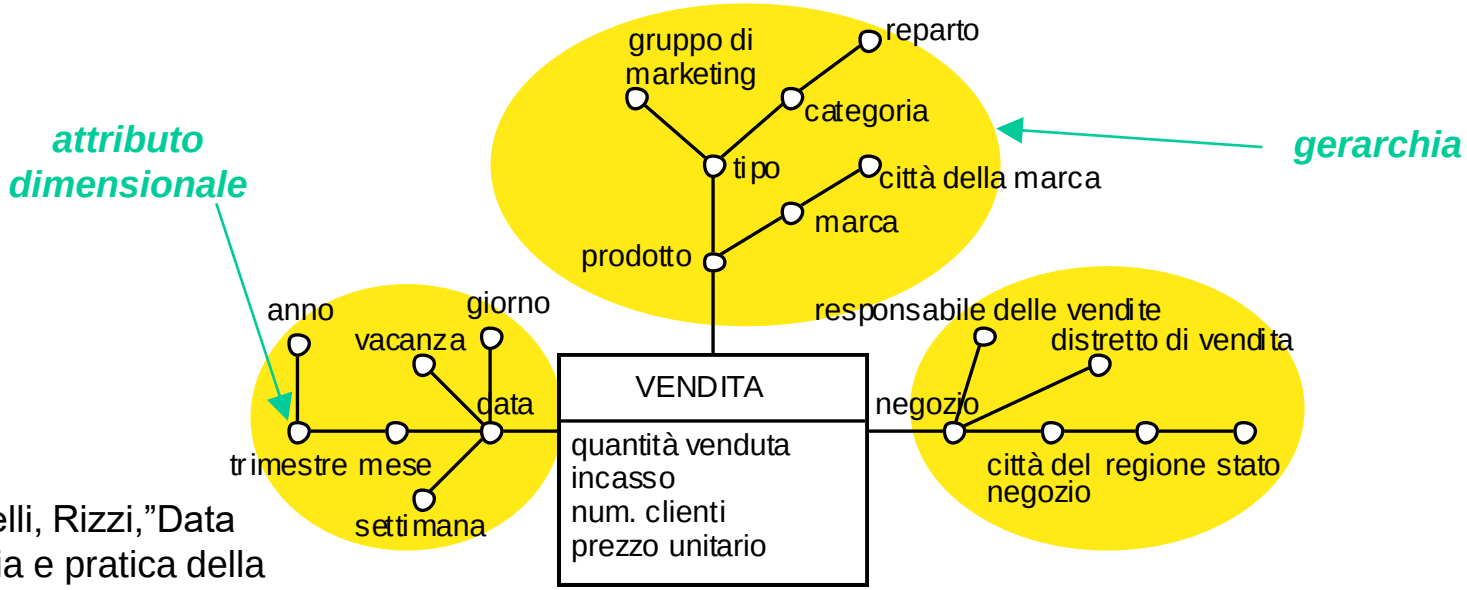


Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Dimensional Fact Model

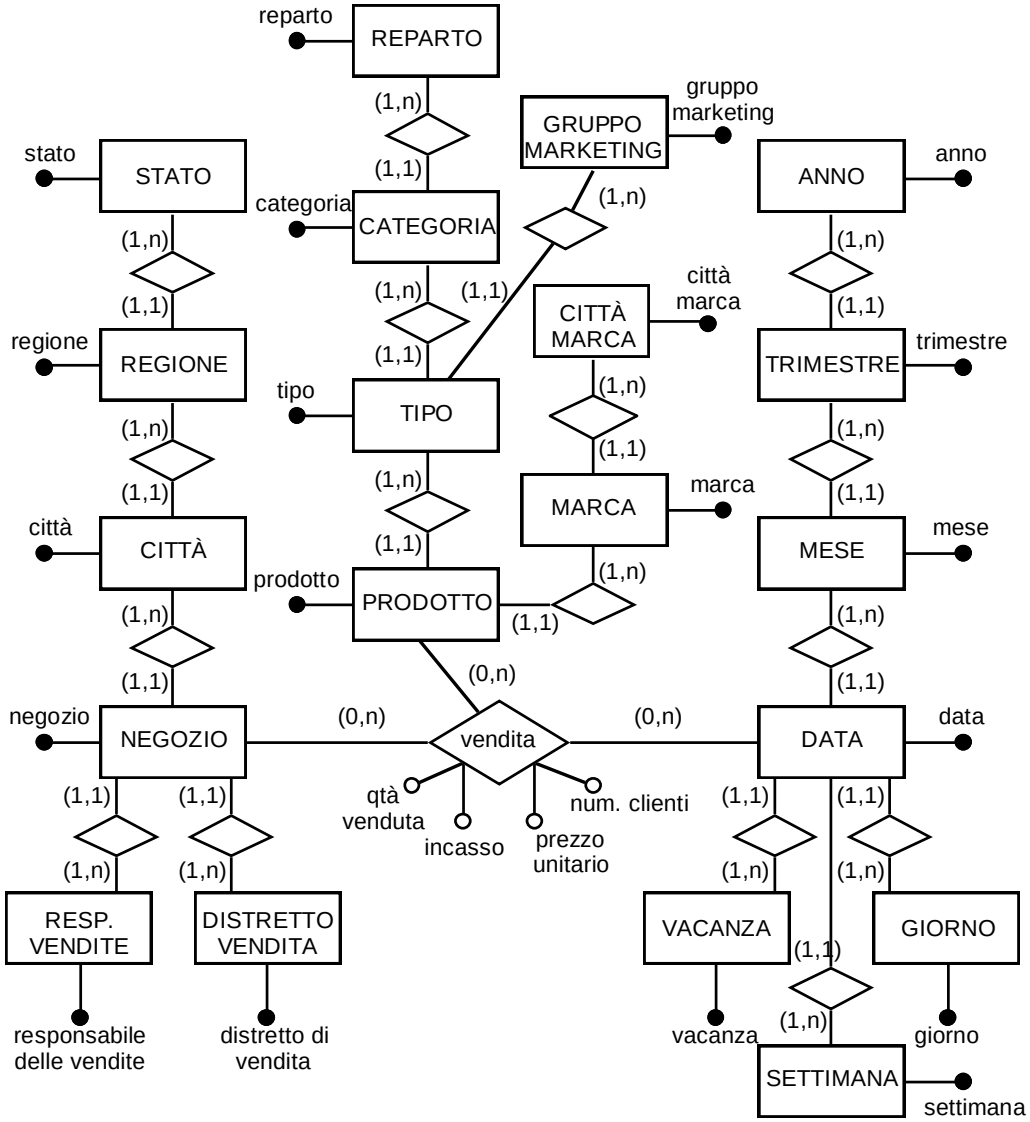
- Gerarchia
  - rappresenta una relazione di generalizzazione tra un sottoinsieme di attributi di una dimensione (gerarchia geografica per la dimensione negozio)
  - è una dipendenza funzionale (relazione 1:n)



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

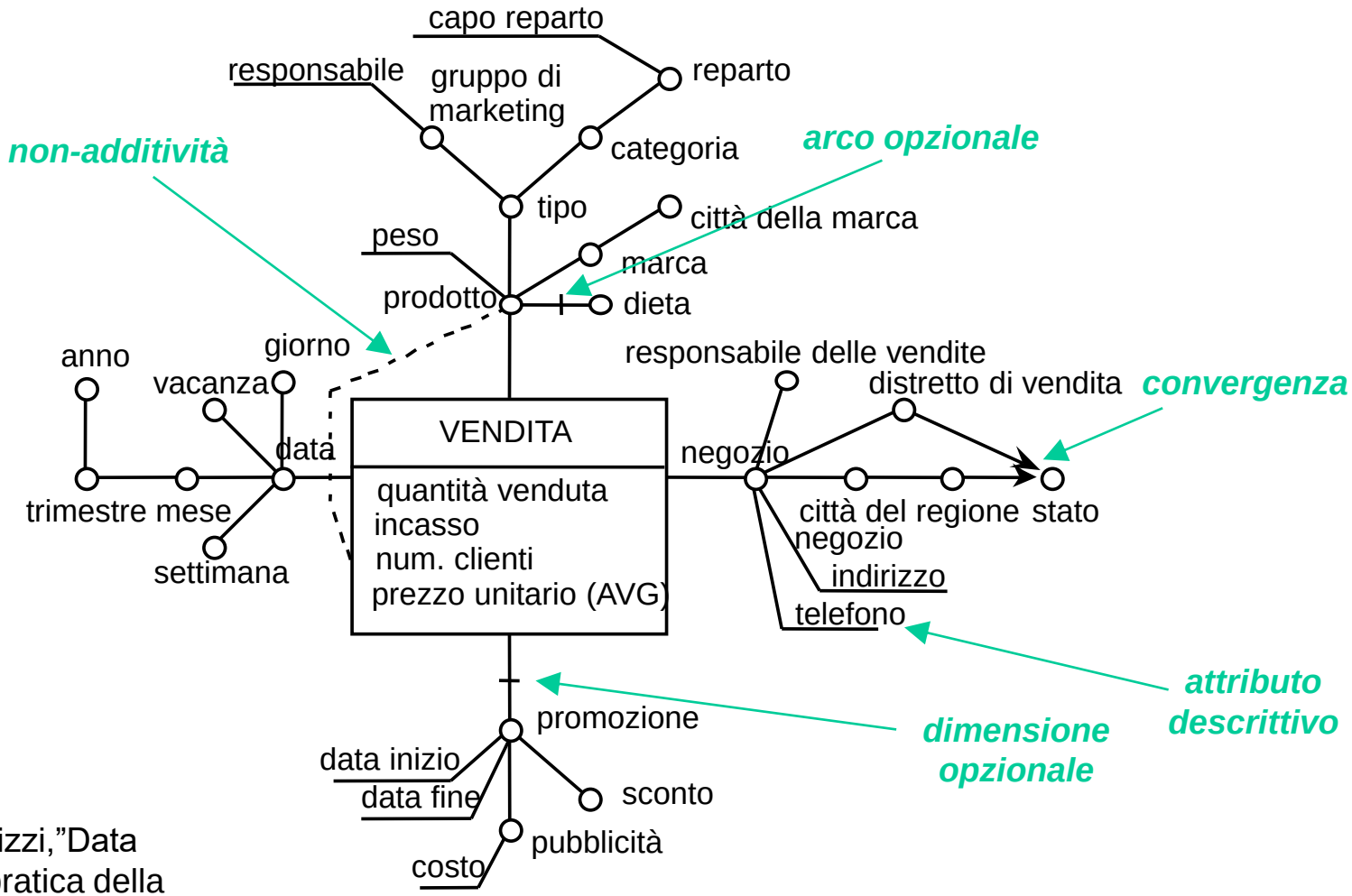
Corrispondenza con l'ER



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

DFM: costrutti avanzati



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Aggregazione

- Processo di calcolo del valore di misure a granularità meno fine di quella presente nello schema di fatto originale
  - la riduzione del livello di dettaglio è ottenuta risalendo lungo una gerarchia
  - operatori di aggregazione standard: SUM, MIN, MAX, AVG, COUNT
- Caratteristiche delle misure
  - additive
  - non additive: non aggregabili lungo una gerarchia mediante l'operatore di somma
  - non aggregabili

Elena Baralis
Politecnico di Torino

Classificazione delle misure

- Misure di flusso
  - possono essere valutate cumulativamente alla fine di un periodo di tempo
  - sono aggregabili mediante tutti gli operatori standard
  - esempi: quantità di prodotti venduti, importo incassato
- Misure di livello
  - sono valutate in specifici istanti di tempo (snapshot)
  - non sono additive lungo la dimensione tempo
  - esempi: livello di inventario, saldo del conto corrente
- Misure unitarie
  - sono valutate in specifici istanti di tempo ed espresse in termini relativi
  - non sono additive lungo nessuna dimensione
  - esempio: prezzo unitario di un prodotto

Elena Baralis
Politecnico di Torino

Operatori di aggregazione

categoria	tipo	prodotto
pulizia casa	detersivo	Brillo
		Sbianco
		Lucido
	sapone	Manipulite
		Scent
alimentari	latticino	Latte F Slurp
		Latte U Slurp
		Yogurt Slurp
	bibita	Bevimi
		Colissima

anno	1999				2000			
trim.	I '99	II '99	III '99	IV '99	I '00	II '00	III '00	IV '00
Brillo	100	90	95	90	80	70	90	85
Sbianco	20	30	20	10	25	30	35	20
Lucido	60	50	60	45	40	40	50	40
Manipulite	15	20	25	30	15	15	20	10
Scent	30	35	20	25	30	30	20	15
Latte F Slurp	90	90	85	75	60	80	85	60
Latte U Slurp	60	80	85	60	70	70	75	65
Yogurt Slurp	20	30	40	35	30	35	35	20
Bevimi	20	10	25	30	35	30	20	10
Colissima	50	60	45	40	50	60	45	40

categoria	anno 1999 trim. I'99	II'99	III'99	IV'99	2000 I'00	II'00	III'00	IV'00
pulizia cas	225	225	220	200	190	185	215	170
alimentari	240	270	280	240	245	275	260	195

categoria	tipo
pulizia casa	detersivo
	sapone
alimentari	latticino
	bibita

anno 1999	2000
670	605
200	155
750	685
280	290

categoria	anno 1999	2000
pulizia casa	870	760
alimentari	1030	975

Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Operatori di aggregazione

- Distributivi
  - sempre possibile il calcolo di aggregati da dati a livello di dettaglio maggiore
  - esempi: sum, min, max

Elena Baralis
Politecnico di Torino

Operatori non distributivi

categoria	tipo	prodotto
pulizia casa	detersivo	Brillo
		Sbianco
		Lucido
	sapone	Manipulite
		Scent

anno / trim.	1999			
	I'99	II'99	III'99	IV'99
	2	2	2,2	2,5
	1,5	1,5	2	2,5
	–	3	3	3
	1	1,2	1,5	1,5
	1,5	1,5	2	–

categoria	tipo
pulizia casa	detersivo
	sapone
	media:

anno / trim.	1999			
	I'99	II'99	III'99	IV'99
	1,75	2,17	2,40	2,67
	1,25	1,35	1,75	1,50
	1,50	1,76	2,08	2,09

categoria
pulizia casa

anno / trim.	1999			
	I'99	II'99	III'99	IV'99
	1,50	1,84	2,14	2,38

Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

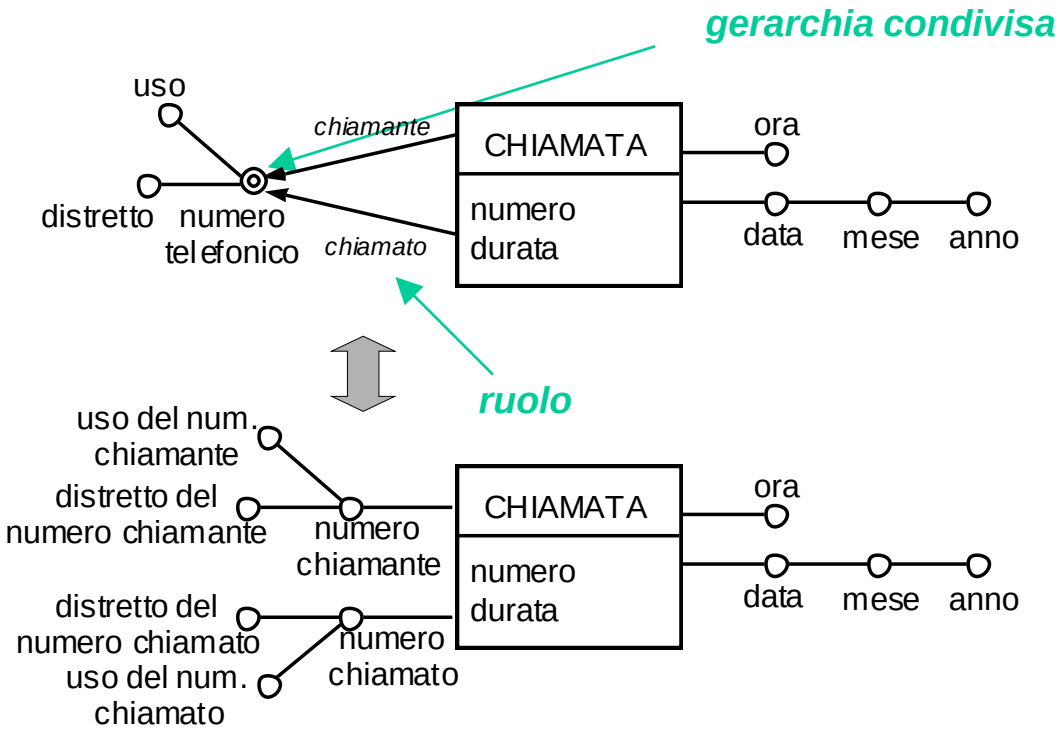
Elena Baralis
Politecnico di Torino

Operatori di aggregazione

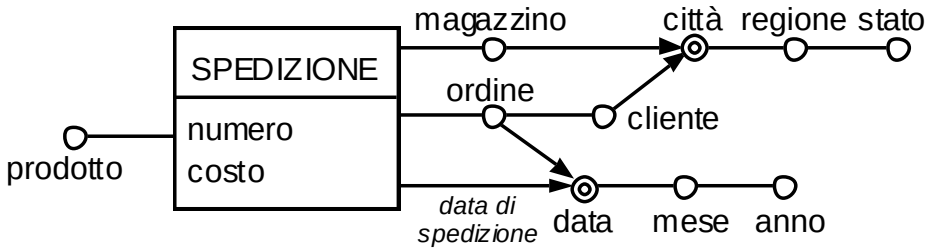
- Distributivi
  - sempre possibile il calcolo di aggregati da dati a livello di dettaglio maggiore
  - esempi: sum, min, max
- Algebrici
  - il calcolo di aggregati da dati a livello di dettaglio maggiore è possibile in presenza di misure aggiuntive di supporto
  - esempi: avg (richiede count)
- Olistici
  - non è possibile il calcolo di aggregati da dati a livello di dettaglio maggiore
  - esempi: moda, mediana

Elena Baralis
Politecnico di Torino

DFM: costrutti avanzati

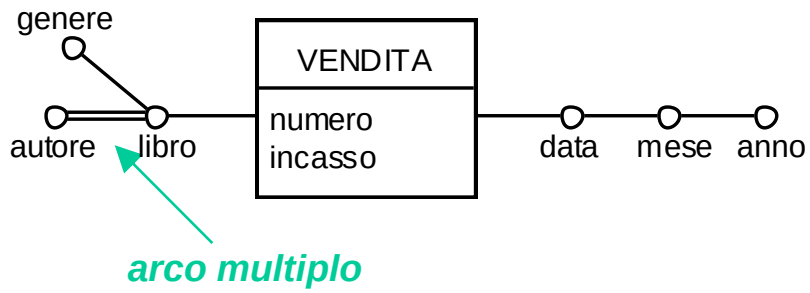


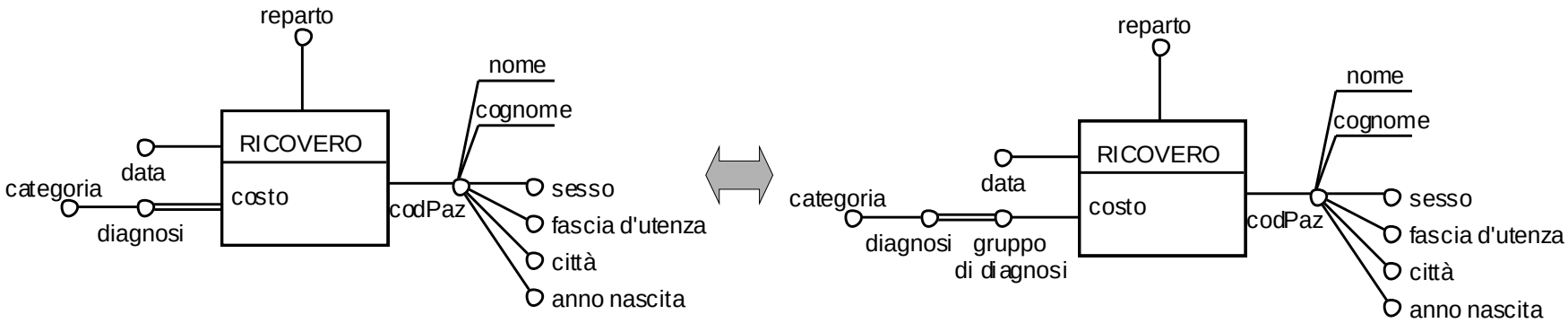
Tratti da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006



Elena Baralis
Politecnico di Torino

DFM: costrutti avanzati



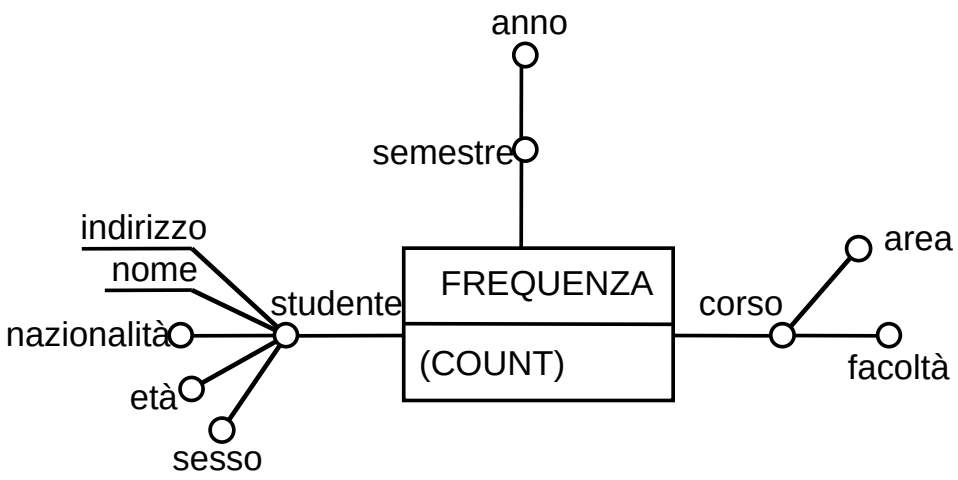


Tratti da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Schemi di fatto vuoti

- L'evento può non essere caratterizzato da misure
  - schema di fatto vuoto
  - registra il verificarsi di un evento
- Utile per
  - conteggio di eventi accaduti
  - rappresentazione di eventi non accaduti (insieme di copertura)



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Rappresentazione del tempo

- La variazione dei dati nel tempo è rappresentata esplicitamente dal verificarsi degli eventi
  - presenza di una dimensione temporale
  - eventi memorizzati sotto forma di fatti
- Possono variare nel tempo anche le dimensioni
  - variazione tipicamente più lenta
    - slowly changing dimension [Kimball]
  - esempi: dati anagrafici di un cliente, descrizione di un prodotto
  - necessario prevedere esplicitamente nel modello come rappresentare questo tipo di variazione

Elena Baralis
Politecnico di Torino

Modalità di rappresentazione del tempo (tipo I)

- Fotografia dell'istante attuale
  - esegue la sovrascrittura del dato con il valore attuale
  - proietta nel passato la situazione attuale
  - utilizzata quando non è necessario rappresentare esplicitamente la variazione
  - Esempio
    - il cliente Mario Rossi cambia stato civile dopo il matrimonio
    - tutti i suoi acquisti sono attribuiti al cliente "sposato"

Elena Baralis
Politecnico di Torino

Modalità di rappresentazione del tempo (tipo II)

- Eventi attribuiti alla situazione temporalmente corrispondente della dimensione
  - per ogni variazione di stato della dimensione
    - si crea una nuova istanza nella dimensione
    - i nuovi eventi sono correlati alla nuova istanza
  - gli eventi sono partizionati in base alle variazioni degli attributi dimensionali
  - Esempio
    - il cliente Mario Rossi cambia stato civile dopo il matrimonio
    - i suoi acquisti sono separati in acquisti attributi a Mario Rossi "celibe" e acquisti attribuiti a Mario Rossi "sposato" (nuova istanza di Mario Rossi)

Elena Baralis
Politecnico di Torino

Modalità di rappresentazione del tempo (tipo III)

- Eventi attribuiti alla situazione della dimensione campionata in uno specifico istante di tempo
  - proietta tutti gli eventi sulla situazione della dimensione in uno specifico istante di tempo
  - richiede una gestione esplicita delle variazioni della dimensione nel tempo
    - modifica dello schema della dimensione
      - introduzione di una coppia di timestamp che indicano l'intervallo di validità del dato (inizio e fine validità)
      - introduzione di un attributo che consenta di identificare la sequenza di variazioni di una specifica istanza (capostipite o master)
    - ogni variazione di stato della dimensione richiede la definizione di una nuova istanza

Elena Baralis
Politecnico di Torino

Modalità di rappresentazione del tempo (tipo III)

– Esempio
- il cliente Mario Rossi cambia stato civile dopo il matrimonio
- la prima istanza conclude il suo periodo di validità il giorno del matrimonio
- la nuova istanza inizia la sua validità nello stesso giorno
- gli acquisti sono separati come nel caso precedente
- esiste un attributo che permette di ricostruire tutte le variazioni associabili a Mario Rossi

Elena Baralis
Politecnico di Torino

Carico di lavoro

- Carico di riferimento definito da
  - reportistica standard
  - stime discusse con gli utenti
- Carico reale difficile da stimare correttamente durante la fase di progettazione
  - se il sistema ha successo, il numero di utenti e interrogazioni aumenta nel tempo
  - la tipologia di interrogazioni può variare nel tempo
- Fase di tuning
  - dopo l'avviamento del sistema
  - monitoraggio del carico di lavoro reale del sistema

Elena Baralis
Politecnico di Torino

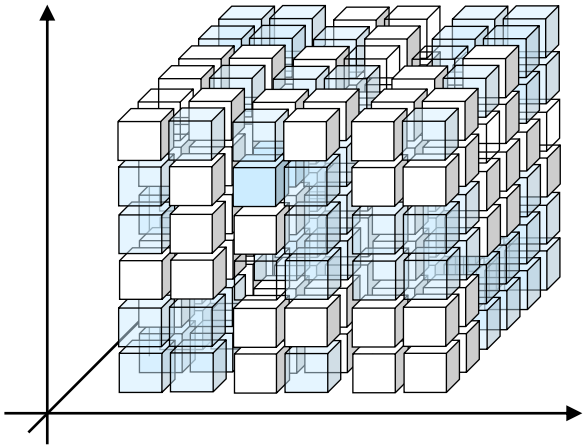
Volume dei dati

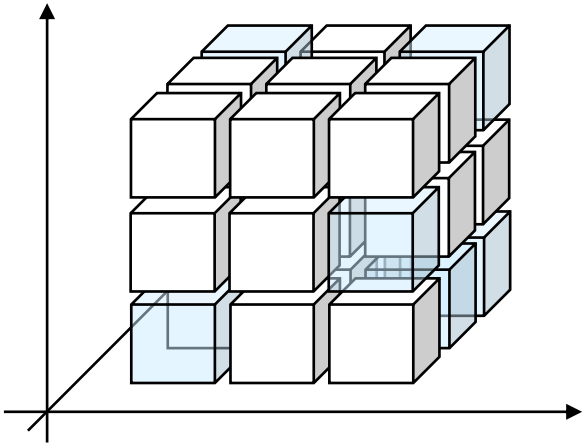
- Stima dello spazio necessario per il data mart
  - per i dati
  - per le strutture accessorie (indici, viste materializzate)
- Si considerano
  - numero di eventi di ogni fatto
  - numero di valori distinti degli attributi nelle gerarchie
  - lunghezza degli attributi
- Dipende dall'intervallo temporale di memorizzazione dei dati
- Valutazione affetta dal problema della sparsità
  - il numero degli eventi accaduti non corrisponde a tutte le possibili combinazioni delle dimensioni
  - esempio: percentuale dei prodotti effettivamente venduti in ogni negozio in un dato giorno pari circa al 10% di tutte le possibili combinazioni

Elena Baralis
Politecnico di Torino

Sparsità

- Si riduce al crescere del livello di aggregazione dei dati
- Può ridurre l'affidabilità della stima della cardinalità dei dati aggregati





Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Progettazione logica

Tania Cerquitelli
Politecnico di Torino

Elena Baralis
Politecnico di Torino

Progettazione logica

- Si considera il modello relazionale (ROLAP)
  - inputs
    - schema (di fatto) concettuale
    - carico di lavoro
    - volume dei dati
    - vincoli di sistema
  - output
    - schema logico relazionale
- Basata su principi diversi rispetto alla progettazione logica tradizionale
  - ridondanza dei dati
  - denormalizzazione delle tabelle

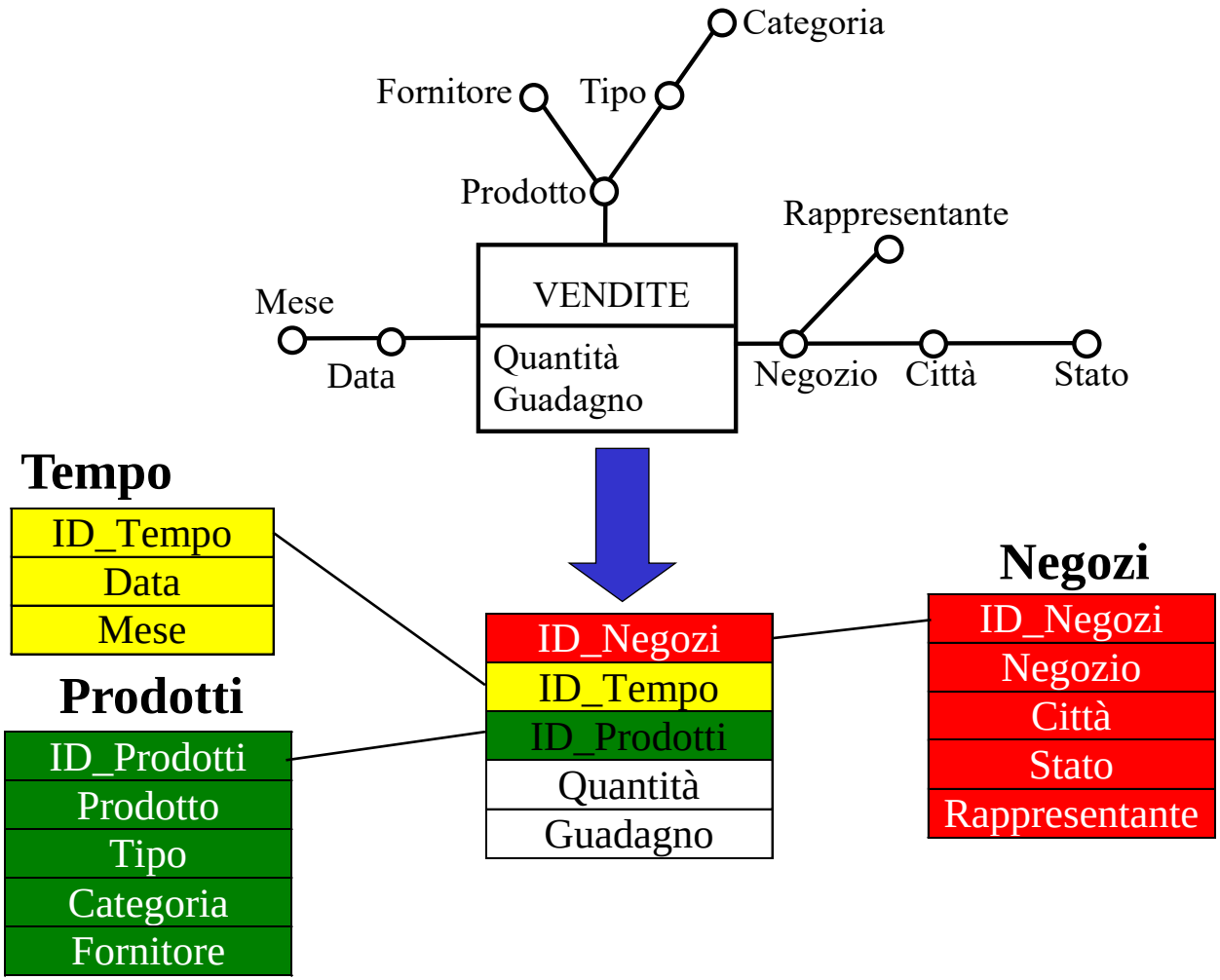
Elena Baralis
Politecnico di Torino

Schema a stella

- Dimensioni
  - una tabella per ogni dimensione
  - chiave primaria generata artificialmente (surrogata)
  - contiene tutti gli attributi della dimensione
  - gerarchie non rappresentate esplicitamente
    - gli attributi della tabella sono tutti allo stesso livello
  - rappresentazione completamente denormalizzata
    - presenza di ridondanza nei dati
- Fatti
  - una tabella dei fatti per ogni schema di fatto
  - chiave primaria costituita dalla combinazione delle chiavi esterne delle dimensioni
  - le misure sono attributi della tabella

Elena Baralis
Politecnico di Torino

Schema a stella



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

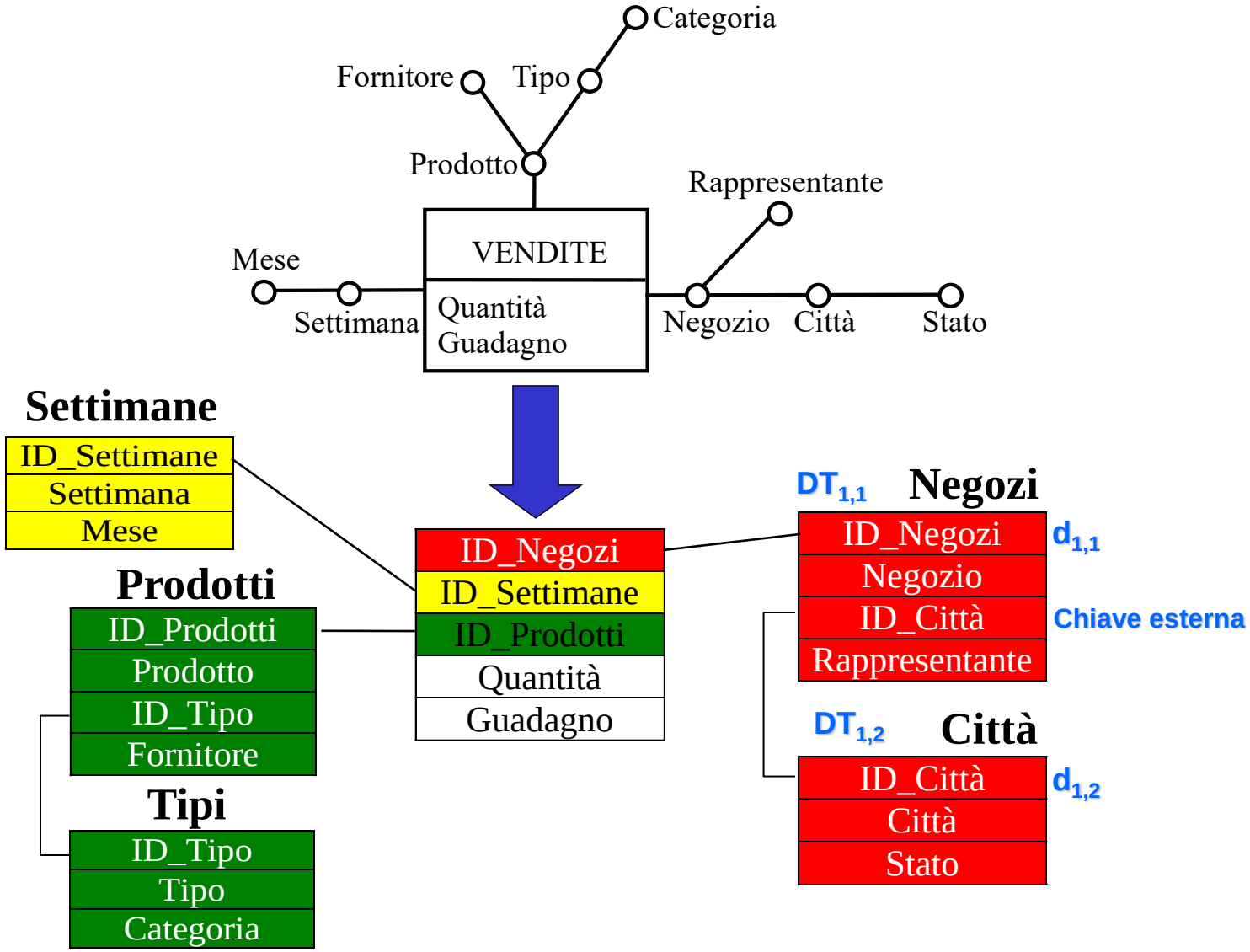
Elena Baralis
Politecnico di Torino

Snowflake schema

- Separazione di (alcune) dipendenze funzionali frazionando i dati di una dimensione in più tabelle
  - si introduce una nuova tabella che separa in due rami una gerarchia dimensionale (taglio su un attributo della gerarchia)
  - una nuova chiave esterna esprime il legame tra la dimensione e la nuova tabella
- Si riduce lo spazio necessario per la memorizzazione della dimensione
  - riduzione non significativa
- Aumenta il costo di ricostruzione dell'informazione della dimensione
  - è necessario il calcolo di uno o più join

Elena Baralis
Politecnico di Torino

Snowflake schema



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

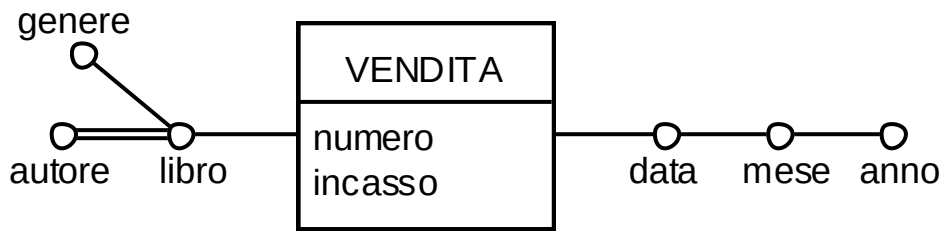
Elena Baralis
Politecnico di Torino

Star o snowflake?

- Lo schema snowflake è normalmente sconsigliato
  - la riduzione di spazio occupato è scarsamente benefica
    - l'occupazione maggiore di spazio è dovuta alla tabella dei fatti (la differenza è pari ad alcuni ordini di grandezza)
  - il costo di eseguire più join può essere significativo
- Lo schema snowflake può essere utile
  - quando porzioni di una gerarchia sono condivise tra più dimensioni (esempio: gerarchia geografica)
  - in presenza di viste materializzate che richiedano una rappresentazione "aggregata" anche della dimensione

Elena Baralis
Politecnico di Torino

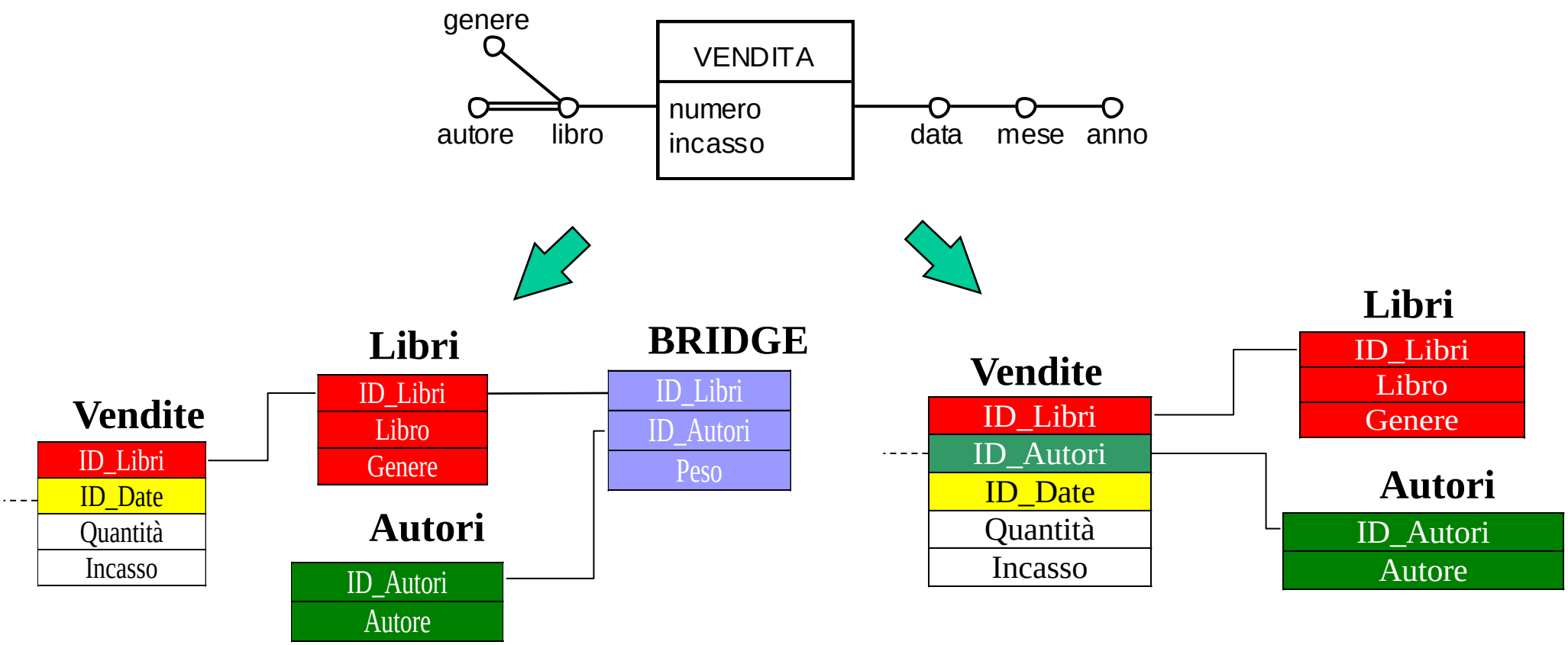
Archi multipli



- Soluzioni realizzative
  - bridge table
    - tabella aggiuntiva che modella la relazione molti a molti
    - nuovo attributo che consenta di pesare la partecipazione delle tuple nella relazione
  - push down
    - arco multiplo integrato nella tabella dei fatti
    - nuova dimensione corrispondente nella tabella dei fatti

Elena Baralis
Politecnico di Torino

Archi multipli



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Archi multipli

- Tipologie di interrogazione
  - pesate: considerano il peso dell'arco multiplo
    - esempio: incasso di ciascun autore
    - con bridge table

```
SELECT ID_Autori, SUM(Incasso*Peso)
...
group by ID_Autori
```

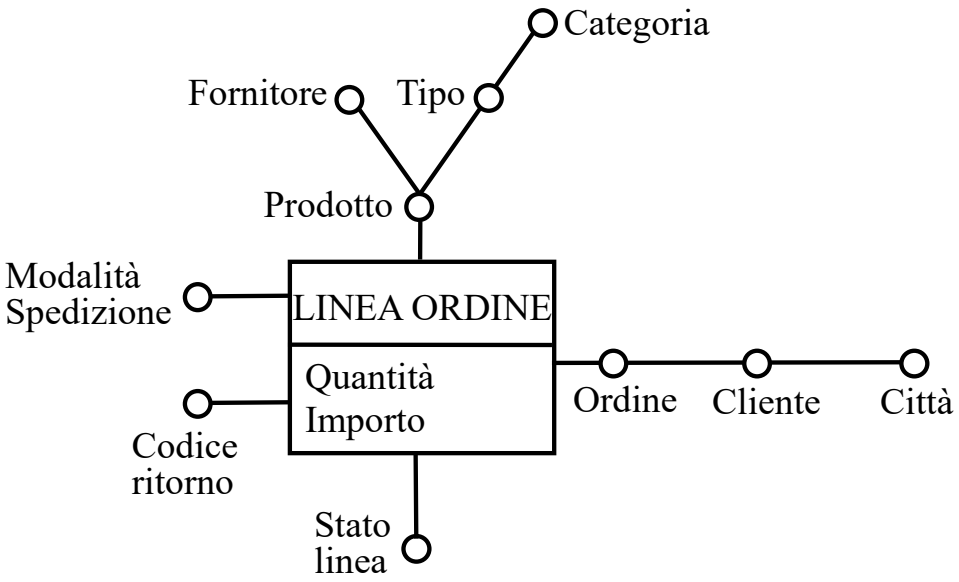
  - di impatto: non considerano il peso
    - esempio: numero di copie vendute per ogni autore
    - con bridge table

```
SELECT ID_Autori, SUM(Quantità)
...
group by ID_Autori
```


Elena Baralis
Politecnico di Torino

Dimensioni degeneri

- Dimensioni rappresentate da un solo attributo



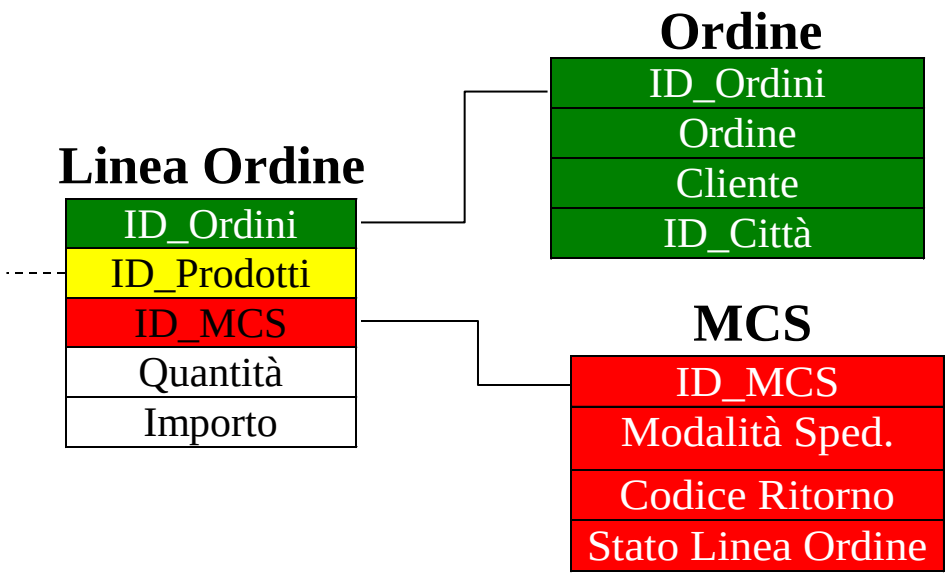
Elena Baralis
Politecnico di Torino

Dimensioni degeneri

- Soluzioni realizzative
  - integrazione nella tabella dei fatti
    - per attributi di dimensione (molto) contenuta
  - junk dimension
    - unica dimensione che integra più dimensioni degeneri
    - non esistono dipendenze funzionali tra gli attributi della dimensione
      - sono possibili tutte le combinazioni
      - attuabile solo per cardinalità limitate del dominio degli attributi

Elena Baralis
Politecnico di Torino

Junk dimension



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

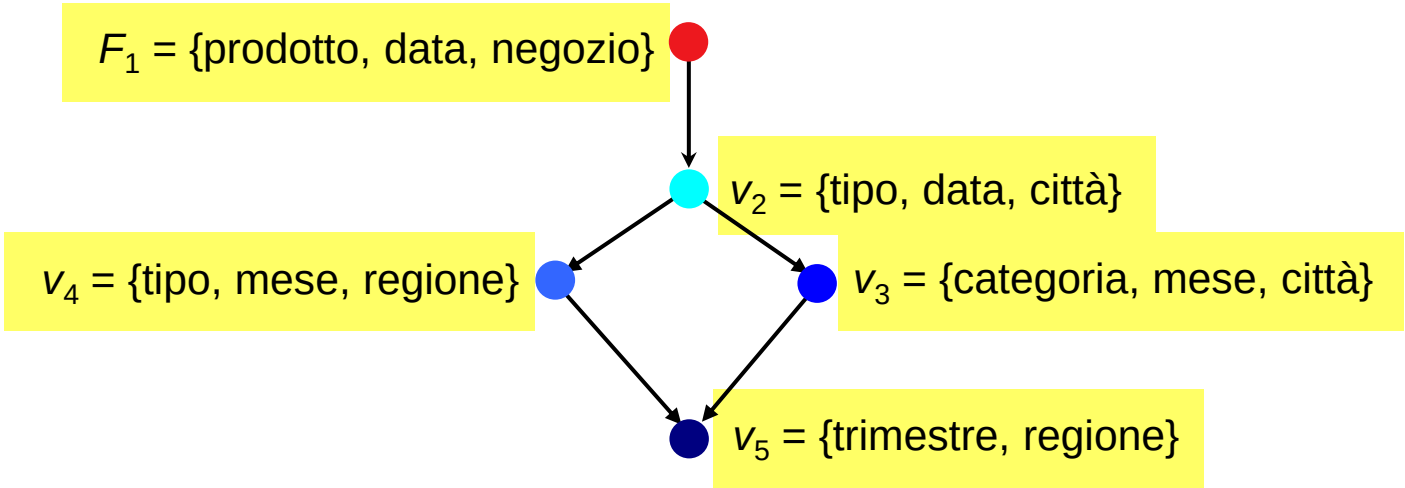
Elena Baralis
Politecnico di Torino

Viste materializzate

Tania Cerquitelli
Politecnico di Torino

Viste materializzate

- Sommari precalcolati della tabella dei fatti
  - memorizzati esplicitamente nel data warehouse
  - permettono di aumentare l'efficienza delle interrogazioni che richiedono aggregazioni



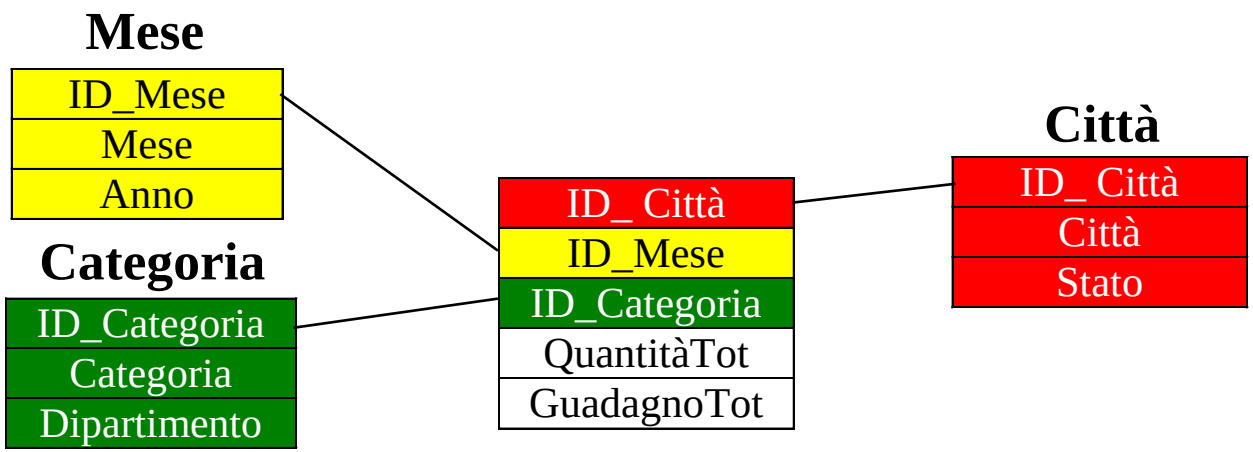
Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Viste materializzate

- Definite da istruzioni SQL
- Esempio: definizione di v_3
  - a partire da tabelle di base o viste di granularità superiore

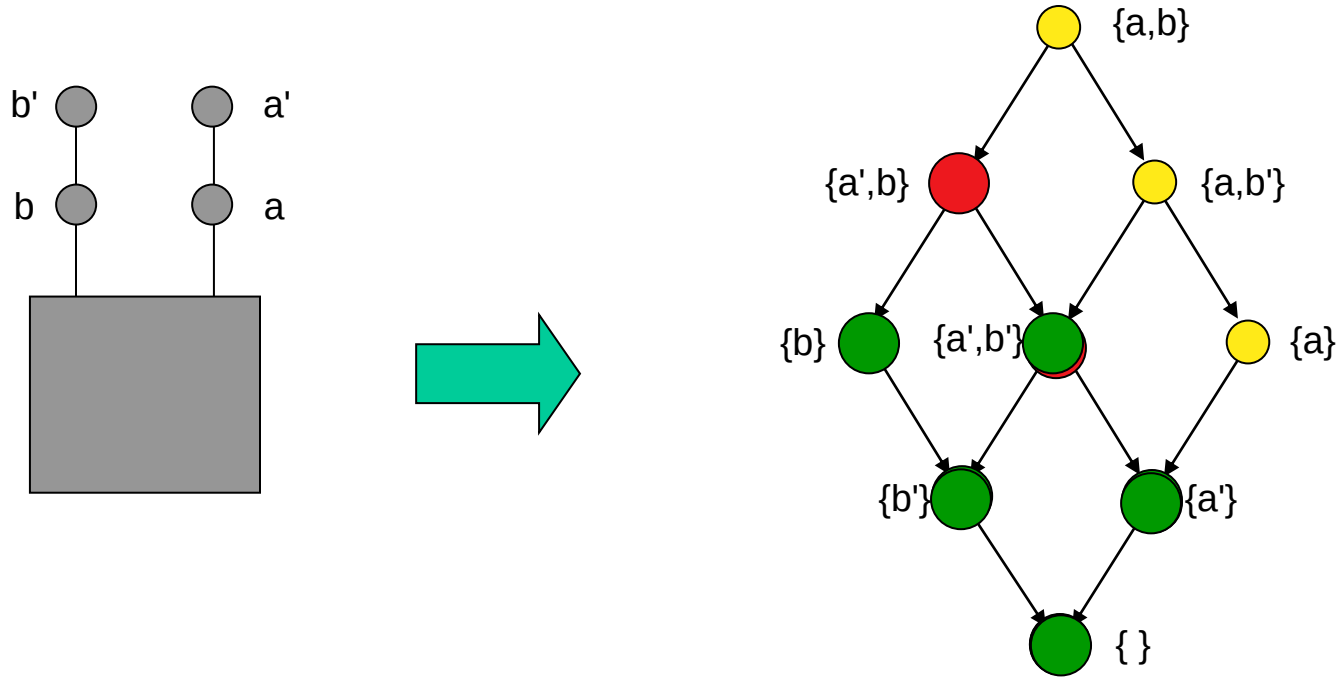
    `group by Città, Mese, Categoria`
  - aggregazione (SUM) sulle misure `Quantità`, `Guadagno`
  - riduzione dettaglio delle dimensioni



Elena Baralis
Politecnico di Torino

Viste materializzate

- Una vista materializzata può essere utilizzata per rispondere a più interrogazioni diverse
  - attenzione al tipo di operatore di aggregazione richiesto



Reticolo multidimensionale

Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

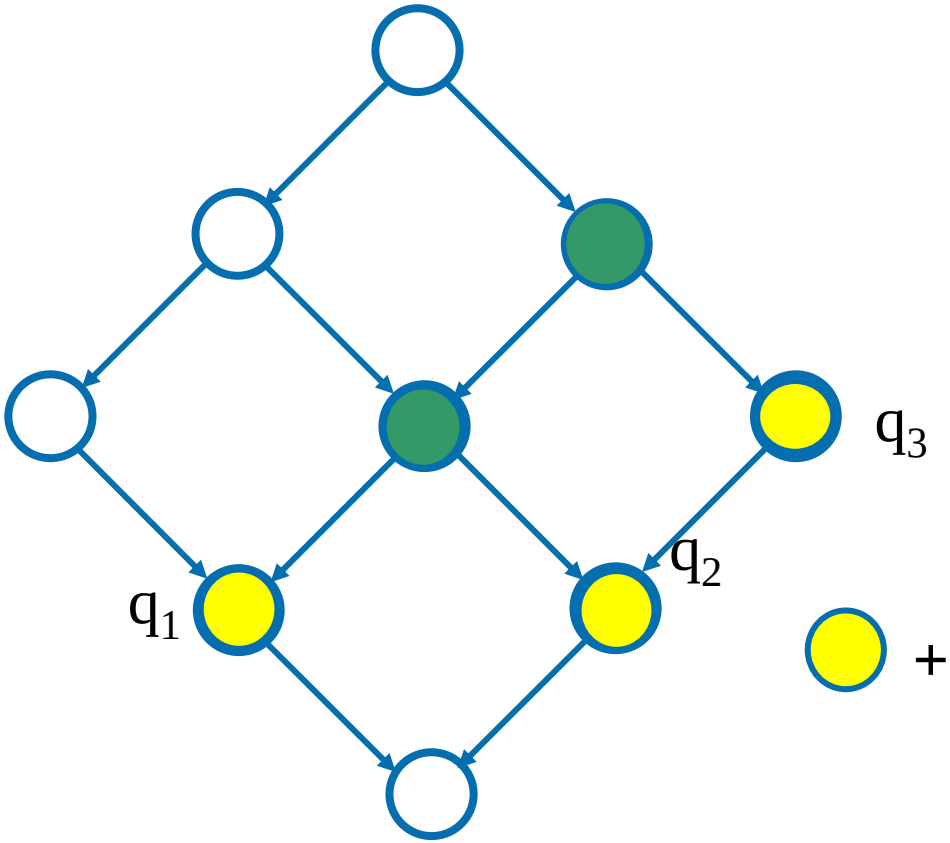
Elena Baralis
Politecnico di Torino

Scelta delle viste

- Numero di possibili combinazioni di aggregazioni molto elevato
  - quasi tutte le combinazioni di attributi sono eleggibili
- Scelta dell'insieme "ottimo" di viste materializzate
- Minimizzazione di funzioni di costo
  - esecuzione delle interrogazioni
  - aggiornamento delle viste materializzate
- Vincoli
  - spazio disponibile
  - tempo a disposizione per l'aggiornamento
  - tempo di risposta
  - freschezza dei dati

Elena Baralis
Politecnico di Torino

Scelta delle viste

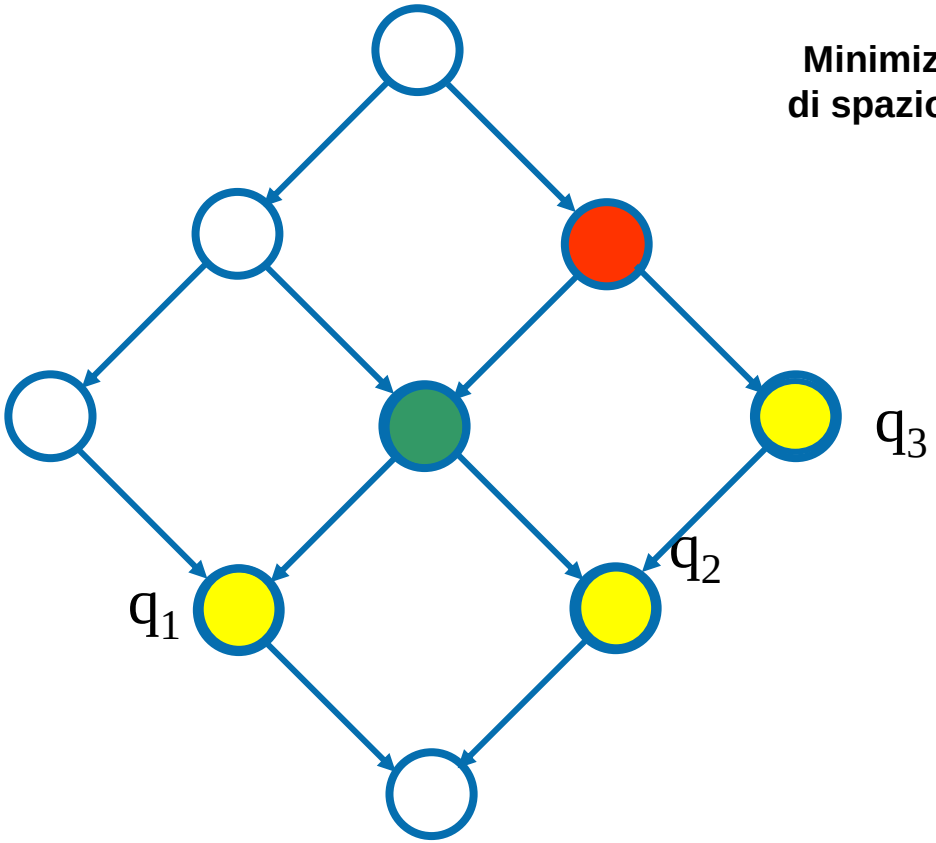


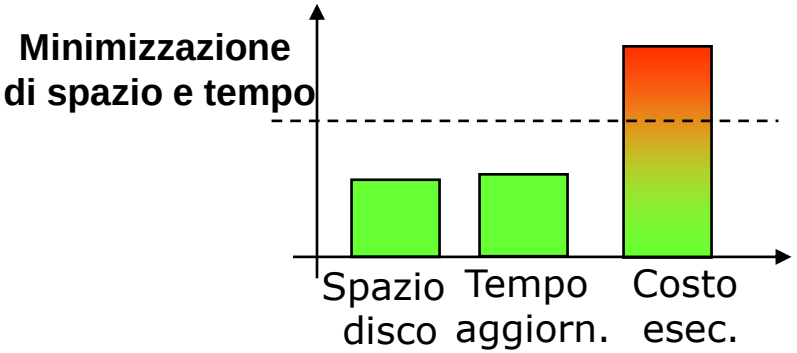
+ = *viste candidate*, **ossia potenzialmente utili a ridurre il costo di esecuzione del carico di lavoro**

Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Scelta delle viste

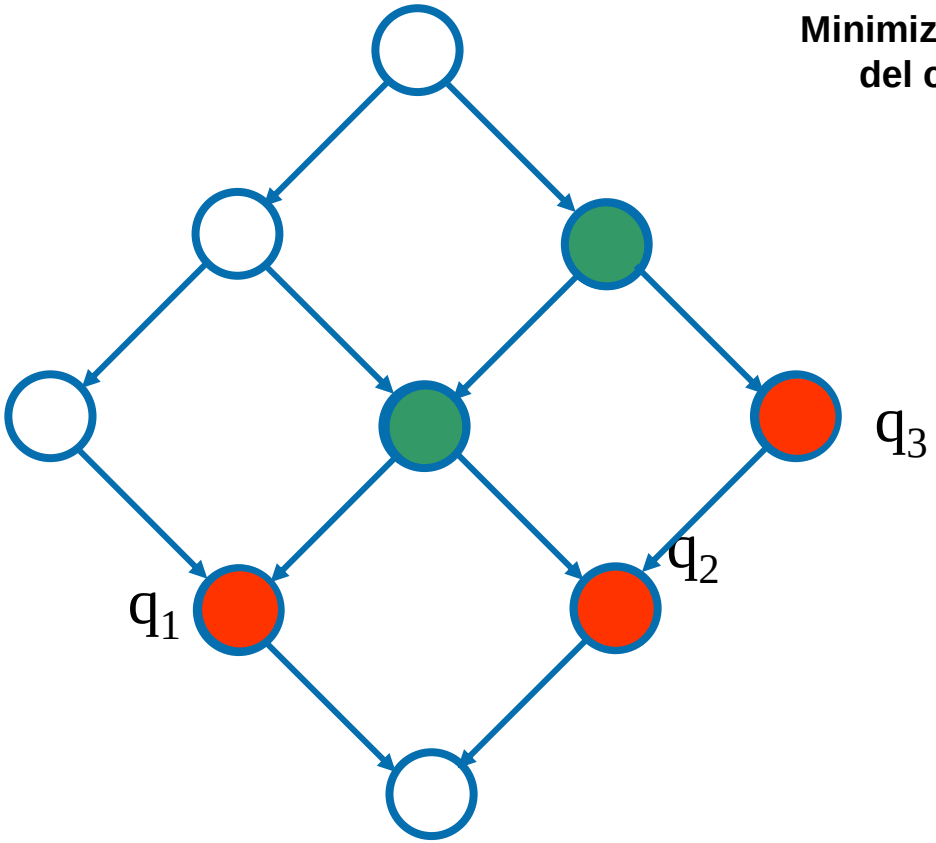


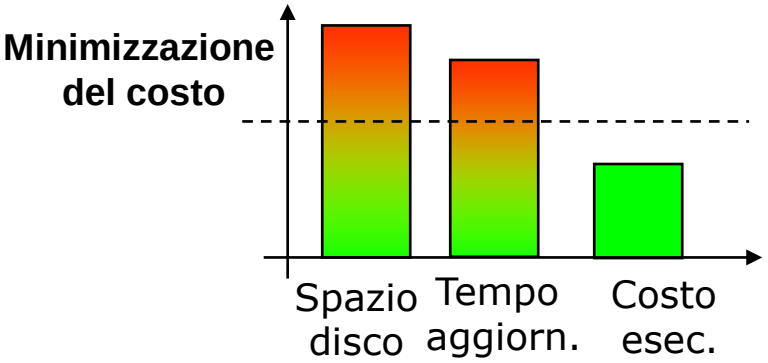


Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Scelta delle viste

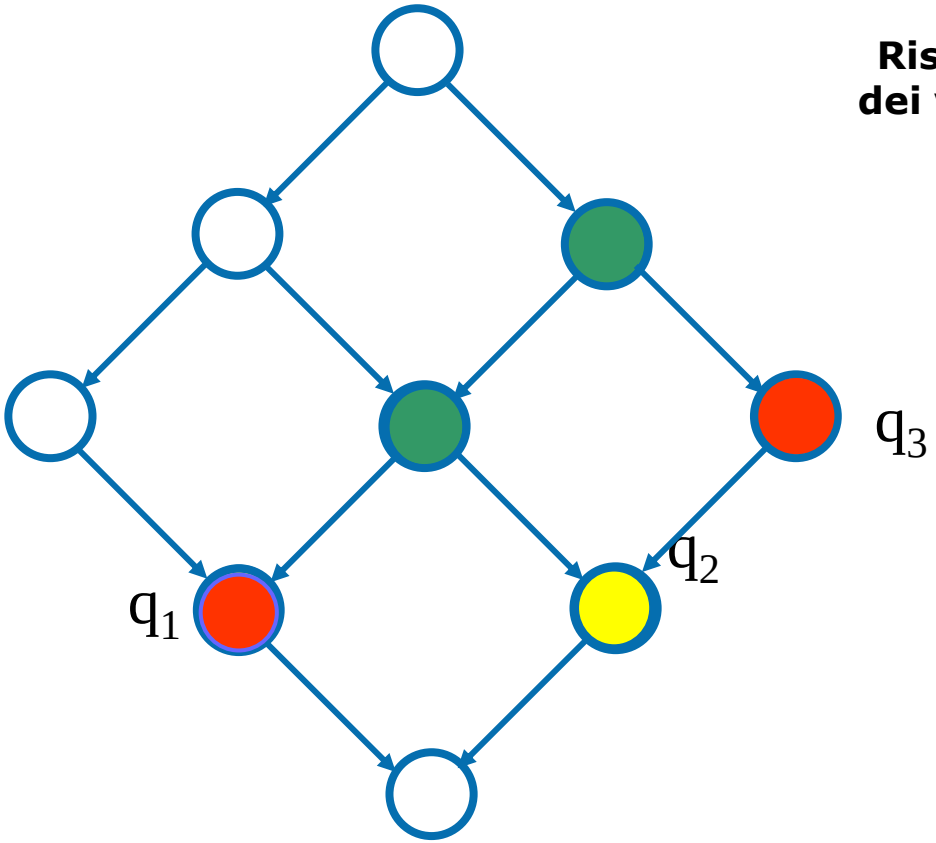


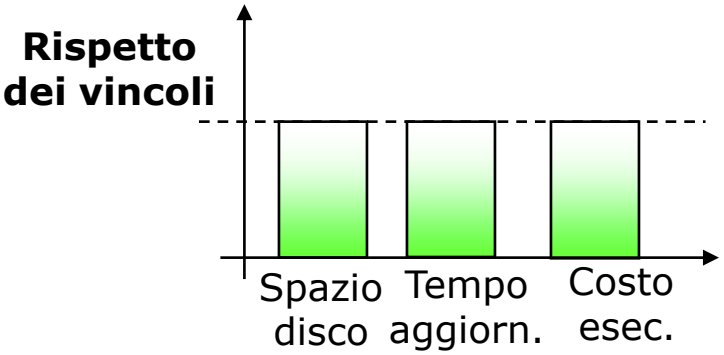


Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Scelta delle viste





Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Progettazione fisica

Tania Cerquitelli
Politecnico di Torino

Elena Baralis
Politecnico di Torino

Progettazione fisica

- Caratteristiche del carico di lavoro
  - interrogazioni con aggregati che richiedono l'accesso a una frazione significativa di ogni tabella
  - accesso in sola lettura
  - aggiornamento periodico dei dati con eventuale ricostruzione delle strutture fisiche di accesso (indici, viste)
- Strutture fisiche
  - tipologie di indici diverse da quelle tradizionali
    - indici bitmap, indici di join, bitmapped join index, ...
    - l'indice B^+-tree non è adatto per
      - attributi con dominio a cardinalità bassa
      - interrogazioni poco selettive
  - viste materializzate
    - richiedono la presenza di un ottimizzatore che le sappia sfruttare

Elena Baralis
Politecnico di Torino

Progettazione fisica

- Caratteristiche dell'ottimizzatore
  - deve considerare le statistiche nella definizione del piano di accesso ai dati (cost based)
  - funzionalità di aggregate navigation
- Procedimento di progettazione fisica
  - selezione delle strutture adatte per supportare le interrogazioni più frequenti (o più rilevanti)
  - scelta di strutture in grado di contribuire al miglioramento di più interrogazioni contemporaneamente
  - vincoli
    - spazio su disco
    - tempo disponibile per l'aggiornamento dei dati

Elena Baralis
Politecnico di Torino

Progettazione fisica

- Tuning
  - variazione a posteriori delle strutture fisiche di supporto
  - richiede strumenti di monitoraggio del carico di lavoro
  - spesso necessario per applicazioni OLAP
- Parallelismo
  - frammentazione dei dati
  - parallelizzazione delle interrogazioni
    - inter-query
    - intra-query
  - le operazioni di join e group by si prestano bene all'esecuzione parallela

Elena Baralis
Politecnico di Torino

Indice bitmap

- Composto da una matrice di bit
  - una colonna per ogni valore diverso del dominio dell'attributo indicizzato
  - una riga per ogni tupla (RID della tabella)
  - la posizione (*i,j*) è 1 se la tupla *i* assume il valore *j*, 0 altrimenti

Esempio: Indice sul campo Posizione della tabella impiegati
Ingegnere – Consulente – Manager – Programmatore
Assistente – Ragioniere

RID	Ing.	Cons.	Man.	Prog.	Assis.	Rag.
1	0	0	1	0	0	0
2	0	0	0	1	0	0
3	0	0	0	0	1	0
4	0	0	0	1	0	0
5	0	0	0	0	0	1

Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

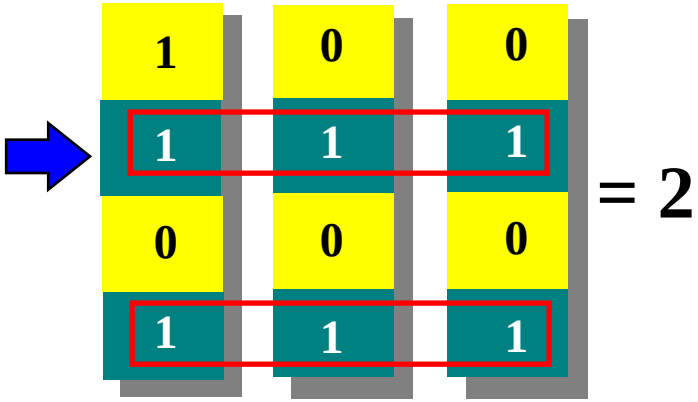
Elena Baralis
Politecnico di Torino

Indice bitmap

- Efficiente per la verifica di espressioni booleane di predicati
  - and/or bit a bit sulle bitmap

Esempio: "*Quanti maschi in Romagna sono assicurati ?*"

RID	Sesso	Assic.	Regione
1	M	No	LO
2	M	Sì	E/R
3	F	No	LA
4	M	Sì	E/R



Esempio tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Indice di join

- Precalcola il join tra due tabelle
  - memorizzazione delle coppie di RID delle tuple che soddisfano il predicato di join

Vendite

	IDNeg	IDSett	IDProd	Quantità	Incasso
1	123	13	41	100	10,00
2	123	24	17	150	15,00
3	367	36	94	350	35,00
4	367	27	11	120	12,00

Negozi

	IDNeg	Negozio	Città	Stato	RespVendite
1	123	N1	RM	I	R1
2	367	N3	MI	I	R2

RID Vendite	RID Negozi
1	1
2	1
3	2
4	2

Coppie di RID che verificano la condizione di join:
Vendite. IDNegozio= Negozio.IDNegozio

Esempio tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Indice a stella

- Precalcola il join tra due o più tabelle
  - memorizzazione delle n-uple di RID delle tuple che soddisfano i predicati di join

Settimana

IDSett	Sett	Mese	
13	Jan1	Jan.	1
24	Jan2	Jan.	2

Negozi

IDNeg	Negozio	Città	Stato	
123	N1	RM	I	1
367	N3	MI	I	2

Vendite

IDNeg	IDSett	IDProd	Quantità	Incasso	
123	13	41	100	10,00	1
123	24	17	150	15,00	2
367	13	17	350	35,00	3
367	24	41	120	12,00	4

VRID	NRID	SID	PID
1	1	1	1
2	1	2	2
3	2	1	2
4	2	2	1

Prodotti

IDProd	Prodotto	Tipo	Categoria	Fornitore	
41	P1	A	X	F1	1
17	P2	A	X	F1	2

Esempio tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Indice a stella

- Vantaggi
  - efficienza nel calcolo di join che coinvolgono le colonne iniziali dell'indice (o tutte le colonne)
- Svantaggi
  - utile solo per specifiche combinazioni di join
    - è necessario memorizzare un numero elevato di indici per avere generalità
  - lo spazio occupato può essere significativo
    - i join coinvolgono sempre la tabella dei fatti

Elena Baralis
Politecnico di Torino

Bitmapped join index

- Matrice di bit che precalcola il join tra una dimensione e la tabella dei fatti
  - una colonna per ogni RID della dimensione
  - una riga per ogni RID della tabella dei fatti
  - la posizione (*i*,*j*) è 1 se la tupla *i* della dimensione è in join con la tupla *j* della tabella dei fatti, 0 altrimenti
- Può essere utilizzato insieme agli indici bitmap tradizionali per calcolare interrogazioni complesse con condizioni sulle dimensioni e join multipli

RID della tabella NEGOZIO

RID della tabella VENDITE

RID	1	2	3	…
1	1	0	0	…
2	0	1	0	…
3	0	0	1	…
4	0	1	0	…
5	1	0	0	…
…	…	…	…	…

La riga 4 della tabella VENDITE è in join con la riga 2 della tabella NEGOZIO

Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Bitmapped join index

Eseguendo un OR bit a bit si ottiene il valore di RID_i che soddisfa tutte le condizioni relative a una tabella dimensionale

Indice Bitmap sull'attributo $DT_i.b_i$

RID	Val_1	Val_2	…	Val_i	…	Val_n
1	1	0	…	0	…	0
2	0	0	…	0	…	1
3	0	1	…	0	…	0
4	0	0	…	1	…	0
5	0	0	…	1	…	0
…	…	…	…	…	…	…

1

Bitmapped join index
$FT.a_i = DT_i.a_i$

RID	1	2	3	4	5	…
1	0	0	0	1	0	…
2	0	0	0	1	0	…
3	0	1	1	0	0	…
4	1	0	0	0	0	…
5	0	0	0	0	1	…
6	0	1	0	0	0	…
…	…	…	…	…	…	…

2

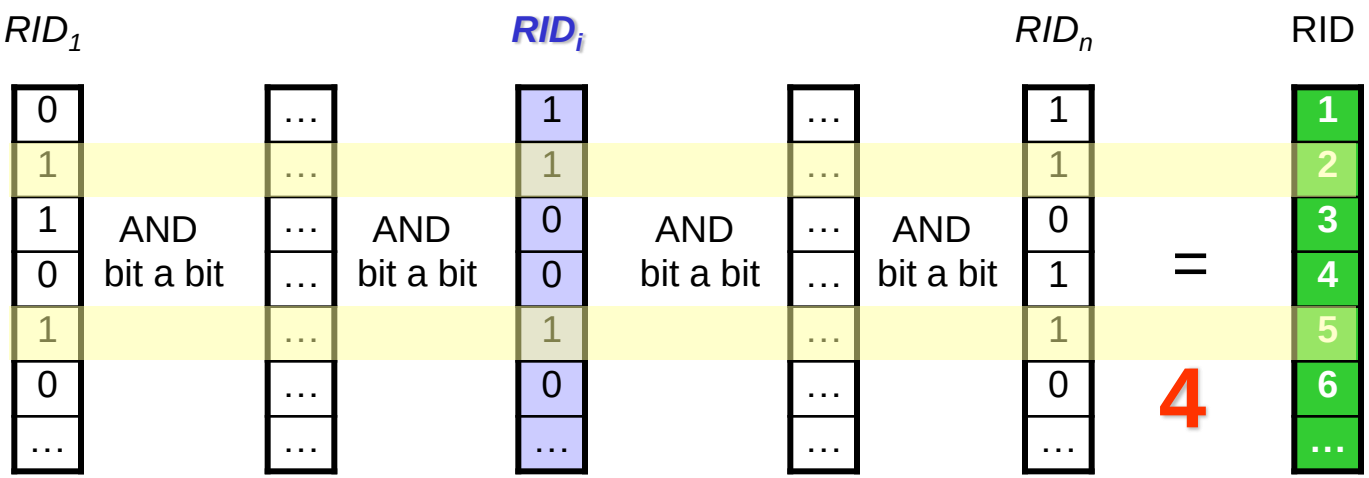
RID 4	RID 5	RID_i
1	0	1
1	0	1
0	0	0
0	0	0
0	1	1
0	0	0
…	…	…

3 OR bit a bit =

Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Bitmapped join index



Le tuple della fact table che soddisfano l'interrogazione vengono determinate eseguendo un AND bit a bit tra gli *n* vettori precedentemente creati

RID che soddisfano tutte le condizioni

Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Scelta degli indici

- Indicizzazione delle dimensioni
  - attributi frequentemente coinvolti in predicati di selezione
  - se il dominio ha cardinalità elevata, indice B-tree
  - se il dominio ha cardinalità ridotta, indice bitmap
- Indici per i join
  - raramente opportuno indicizzare solo le chiavi esterne della tabella dei fatti
  - uso con cautela di star join index (problema dell'ordine delle colonne)
  - consigliati bitmapped join index
- Indici per i group by
  - uso di viste materializzate

Elena Baralis
Politecnico di Torino

Alimentazione del data warehouse

Tania Cerquitelli
Politecnico di Torino

Elena Baralis
Politecnico di Torino

Extraction, Transformation and Loading (ETL)

- Processo di preparazione dei dati da introdurre nel data warehouse
  - estrazione dei dati dalle sorgenti
  - pulitura
  - trasformazione
  - caricamento
- semplificato dalla presenza di una staging area
- eseguito durante
  - il primo popolamento del DW
  - l'aggiornamento periodico dei dati

Elena Baralis
Politecnico di Torino

Estrazione

- Acquisizione dei dati dalle sorgenti
- Modalità di estrazione
  - statica: fotografia dei dati operazionali
    - eseguita durante il primo popolamento del DW
  - incrementale: selezione degli aggiornamenti avvenuti dopo l'ultima estrazione
    - utilizzata per l'aggiornamento periodico del DW
    - immediata o ritardata
- Scelta dei dati da estrarre basata sulla loro qualità

Elena Baralis
Politecnico di Torino

Estrazione

- Dipende dalla natura dei dati operazionali
  - storicizzati: tutte le modifiche sono memorizzate per un periodo definito di tempo nel sistema OLTP
    - transazioni bancarie, dati assicurativi
    - operativamente semplice
  - semi-storicizzati: è conservato nel sistema OLTP solo un numero limitato di stati
    - operativamente complessa
  - transitori: il sistema OLTP mantiene solo l'immagine corrente dei dati
    - scorte di magazzino, dati di inventario
    - operativamente complessa

Elena Baralis
Politecnico di Torino

Estrazione incrementale

- Assistita dall'applicazione
  - le modifiche sono catturate da specifiche funzioni applicative
  - richiede la modifica delle applicazioni OLTP (o delle API di accesso alla base di dati)
  - aumenta il carico applicativo
  - necessaria per sistemi legacy
- Uso del log
  - accesso mediante primitive opportune ai dati del log
  - formato proprietario del log
  - efficiente, non interferisce con il carico applicativo

Elena Baralis
Politecnico di Torino

Estrazione incrementale

- Definizione di trigger
  - i trigger catturano le modifiche di interesse
  - non richiede la modifica dei programmi applicativi
  - aumenta il carico applicativo
- Basata su timestamp
  - i record operazionali modificati sono marcati con il timestamp dell'ultima modifica
  - richiede la modifica dello schema della base di dati OLTP (e delle applicazioni)
  - estrazione differita, può perdere stati intermedi se i dati sono transitori

Elena Baralis
Politecnico di Torino

Confronto tra le tecniche di estrazione

	Statica	*Marche temporali*	*Assistita applicazione*	*Trigger*	*Log*
Gestione dati transitori o semi-storicizzati	NO	Incompleta	Completa	Completa	Completa
Supporto per sistemi basati su file	SI	SI	SI	NO	Raro
Tecnica di realizzazione	Prodotti	Prodotti o sviluppo interno	Sviluppo interno	Prodotti	Prodotti
Costi di sviluppo interno	Nessuno	Medi	Alti	Nessuno	Nessuno
Utilizzo in sistemi legacy	SI	Difficile	Difficile	Difficile	SI
Modifiche ad applicazioni	Nessuna	Probabile	Probabile	Nessuna	Nessuna
Dipendenza delle procedure dal DBMS	Limitata	Limitata	Variabile	Alta	Limitata
Impatto sulle prestazioni del sistema operaz.	Nessuna	Nessuna	Medio	Medio	Nessuna
Complessità delle procedure di estrazione	Bassa	Bassa	Alta	Media	Bassa

Tratto da Devlin, Data warehouse: from architecture to implementation, Addisono-Wesley, 1997

Elena Baralis
Politecnico di Torino

Estrazione incrementale

4/4/2010

Cod	Prodotto	Cliente	Qtà
1	Greco di tufo	Malavasi	50
2	Barolo	Maio	150
3	Barbera	Lumini	75
4	Sangiovese	Cappelli	45

6/4/2010

Cod	Prodotto	Cliente	Qtà
1	Greco di tufo	Malavasi	50
2	Barolo	Maio	150
4	Sangiovese	Cappelli	145
5	Vermentino	Maltoni	25
6	Trebbiano	Maltoni	150

Differenza incrementale

Cod	Prodotto	Cliente	Qtà	Azione
3	Barbera	Lumini	75	D
4	Sangiovese	Cappelli	145	U
5	Vermentino	Maltoni	25	I
6	Trebbiano	Maltoni	150	I

Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Pulitura

- Operazioni volte al miglioramento della qualità dei dati (correttezza e consistenza)
  - dati duplicati
  - dati mancanti
  - uso non previsto di un campo
  - valori impossibili o errati
  - inconsistenza tra valori logicamente associati
- Problemi dovuti a
  - errori di battitura
  - differenze di formato dei campi
  - evoluzione del modo di operare dell'azienda

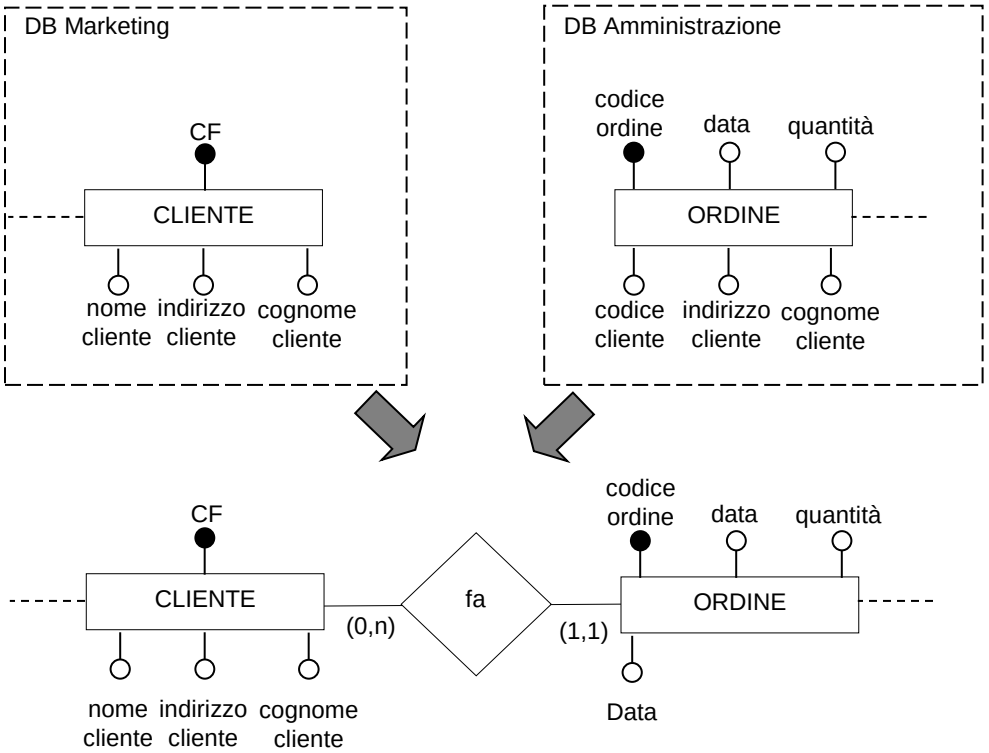
Elena Baralis
Politecnico di Torino

Pulitura

- Ogni problema richiede una tecnica specifica di soluzione
  - tecniche basate su dizionari
    - adatte per errori di battitura o formato
    - utilizzabili per attributi con dominio ristretto
  - tecniche di fusione approssimata
    - adatte per riconoscimento di duplicati/correlazioni tra dati simili
      - join approssimato
      - problema purge/merge
  - identificazione di outliers o deviazioni da business rules
- La strategia migliore è la prevenzione, rendendo più affidabili e rigorose le procedure di data entry OLTP

Elena Baralis
Politecnico di Torino

Join approssimato

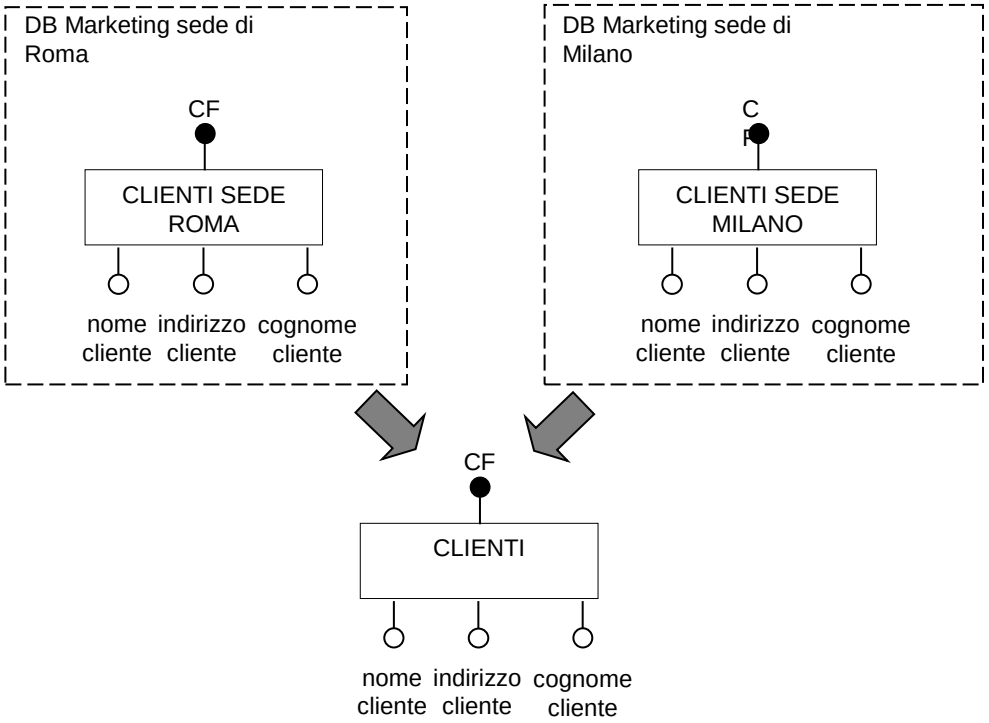


Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

- Il join deve essere eseguito sulla base dei campi comuni, che non rappresentano un identificatore per il cliente

Elena Baralis
Politecnico di Torino

Problema purge/merge



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

- I record duplicati devono essere identificati ed eliminati
- E` necessario un criterio per valutare la somiglianza tra due record

Elena Baralis
Politecnico di Torino

Trasformazione

- Conversione dei dati dal formato operazionale a quello del data warehouse (integrazione)
- Richiede una rappresentazione uniforme dei dati operazionali (schema riconciliato)
- Può avvenire in due passi
  - dalle sorgenti operazionali ai dati riconciliati nella staging area
    - conversioni e normalizzazioni
    - matching
    - (eventuale) filtraggio dei dati significativi
  - dai dati riconciliati al data warehouse
    - generazione di chiavi surrogate
    - generazione di valori aggregati

Elena Baralis
Politecnico di Torino

Esempio di pulitura e trasformazione

Elena Baralis
C.so Duca degli Abruzzi 24
20129 Torino (I)

Normalizzazione

nome:	Elena
cognome:	Baralis
indirizzo:	C.so Duca degli Abruzzi 24
CAP:	20129
città:	Torino
nazione:	I

Standardizzazione

nome:	Elena
cognome:	Baralis
indirizzo:	Corso Duca degli Abruzzi 24
CAP:	20129
città:	Torino
nazione:	Italia

Correzione

nome:	Elena
cognome:	Baralis
indirizzo:	Corso Duca degli Abruzzi 24
CAP:	10129
città:	Torino
nazione:	Italia

Adattato da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

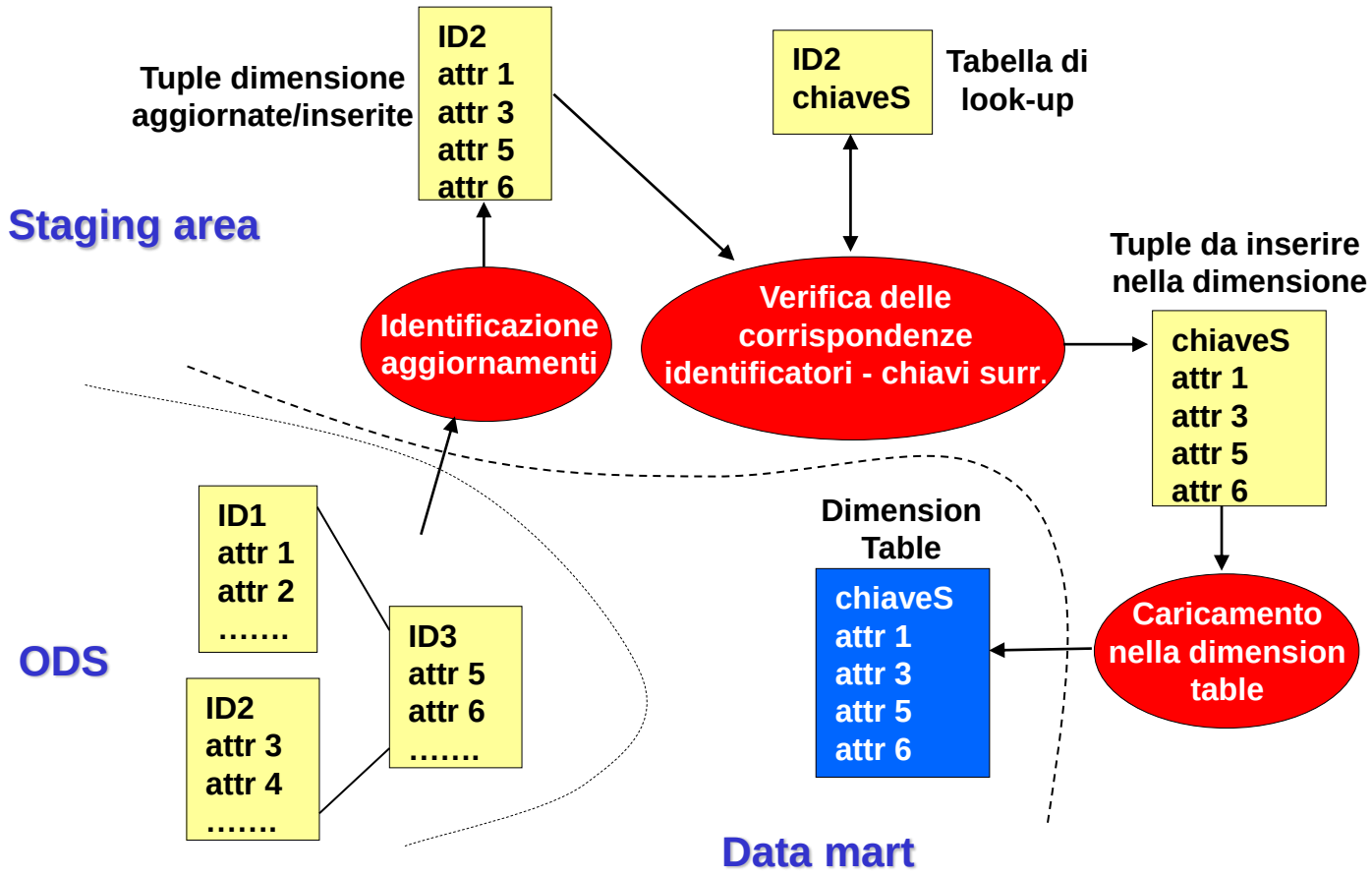
Elena Baralis
Politecnico di Torino

Caricamento

- Propagazione degli aggiornamenti al data warehouse
- Per mantenere l'integrità dei dati, si aggiornano in ordine
  1. dimensioni
  2. tabelle dei fatti
  3. viste materializzate e indici
- Finestra temporale limitata per eseguire gli aggiornamenti
- Richiede proprietà transazionali (affidabilità, atomicità)

Elena Baralis
Politecnico di Torino

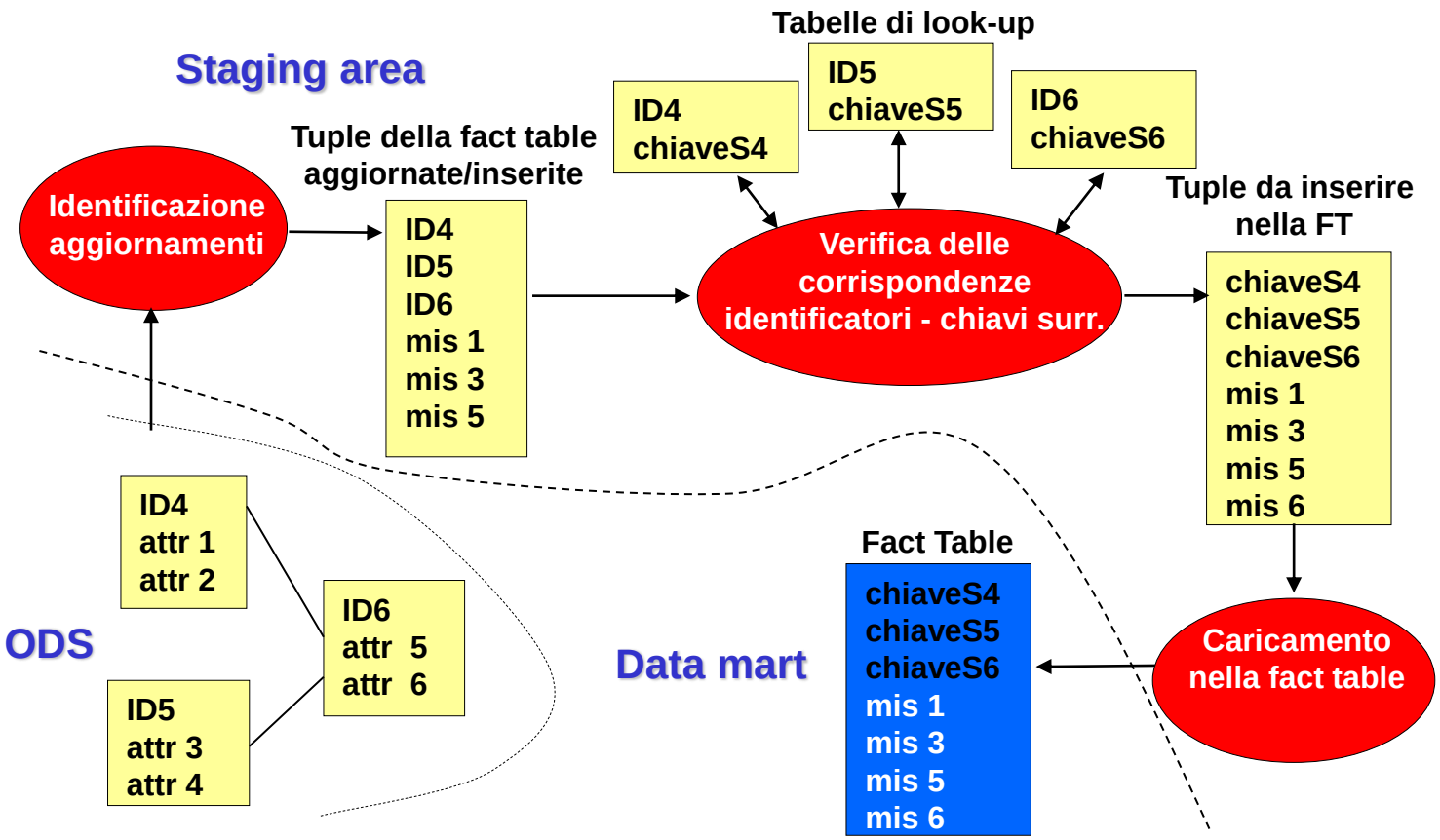
Alimentazione delle dimensioni



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

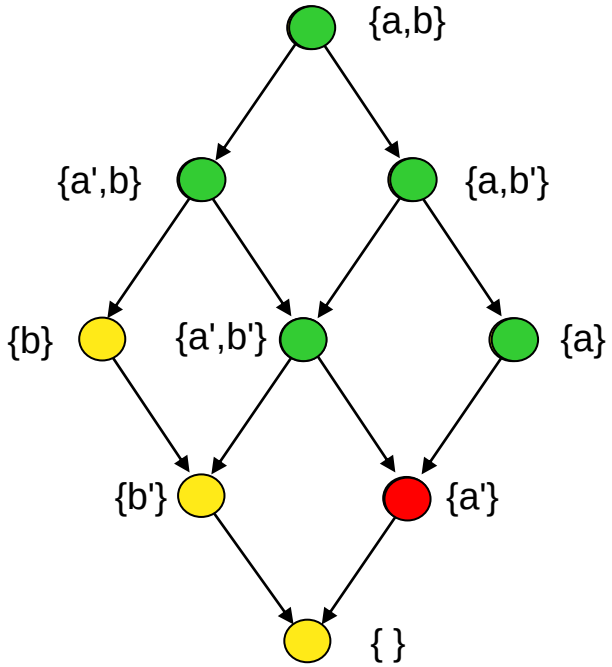
Alimentazione delle fact table



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Alimentazione delle viste materializzate



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino