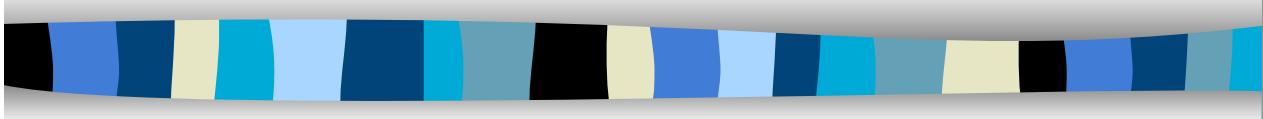


Il ciclo di vita del Data Warehouse



Perché?

- Molte organizzazioni mancano della necessaria esperienza e capacità per affrontare con successo le sfide implicite nei progetti di data warehousing
- Uno dei fattori che maggiormente minaccia la riuscita dei progetti è la mancata adozione di una **approccio metodologico**, che minimizza i rischi di insuccesso essendo basato su un'analisi costruttiva degli errori commessi

Fattori di rischio

- ✓ Rischi legati alla gestione del progetto
- ✓ Rischi legati alle tecnologie
- ✓ Rischi legati ai dati e alla progettazione
- ✓ Rischi legati all' organizzazione
- Il rischio di ottenere un risultato insoddisfacente nei progetti di data warehousing è particolarmente alto a causa delle elevatissime aspettative degli utenti
- Nella cultura aziendale contemporanea è infatti diffusissima la credenza che attribuisce al data warehousing il ruolo di panacea
- In realtà una larga parte della responsabilità della riuscita del progetto ricade sulla qualità dei dati sorgente e sulla lungimiranza, disponibilità e dinamismo del personale dell' azienda

3

Approccio top-down

- Analizza i bisogni globali dell' intera azienda e pianifica lo sviluppo del DW per poi progettarlo e realizzarlo nella sua interezza
 - ➔ Promette ottimi risultati poiché si basa su una visione globale dell' obiettivo e garantisce in linea di principio di produrre un DW consistente e ben integrato
 - ➔ Il preventivo di costi onerosi a fronte di lunghi tempi di realizzazione scoraggia la direzione dall' intraprendere il progetto
 - ➔ Affrontare contemporaneamente l' analisi e la riconciliazione di tutte le sorgenti di interesse è estremamente complesso
 - ➔ Riuscire a prevedere a priori nel dettaglio le esigenze delle diverse aree aziendali impegnate è pressoché impossibile, e il processo di analisi rischia di subire una paralisi
 - ➔ Il fatto di non prevedere la consegna a breve termine di un prototipo non permette agli utenti di verificare l' utilità del progetto e ne fa scemare l' interesse e la fiducia

4

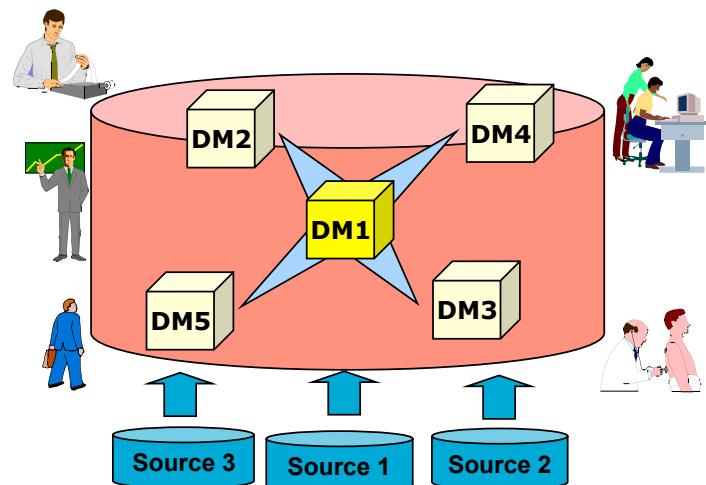
Approccio bottom-up

- Il DW viene costruito in modo incrementale, assemblando iterativamente più data mart, ciascuno dei quali incentrato su un insieme di fatti collegati a uno specifico settore aziendale e di interesse per una certa categoria di utenti
 - Determina risultati concreti in tempi brevi
 - Non richiede elevati investimenti finanziari
 - Permette di studiare solo le problematiche relative al data mart in oggetto
 - Fornisce alla dirigenza aziendale un riscontro immediato sull'effettiva utilità del sistema in via di realizzazione
 - Mantiene costantemente elevata l'attenzione sul progetto
 - Determina una visione parziale del dominio di interesse

5

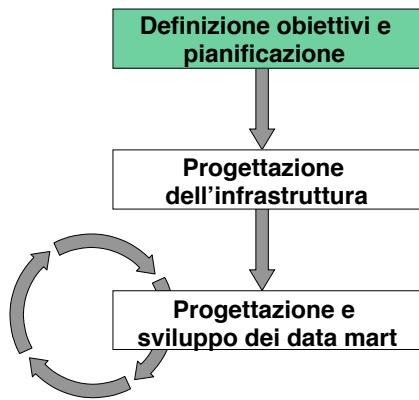
Il primo data mart da prototipare...

- ✓ deve essere quello che gioca il ruolo più strategico per l'azienda
- ✓ deve ricoprire un ruolo centrale e di riferimento per l'intero DW
- ✓ si deve appoggiare su fonti dati già disponibili e consistenti



6

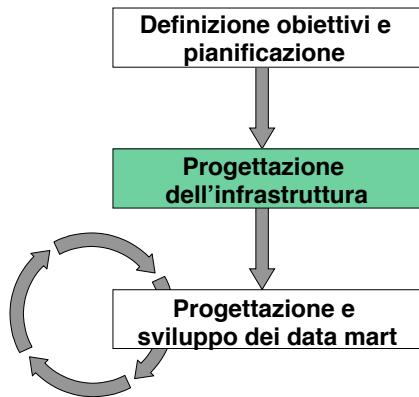
Il ciclo di sviluppo



- individuazione degli obiettivi e dei confini del sistema
- stima delle dimensioni
- scelta dell' approccio per la costruzione
- valutazione dei costi e del valore aggiunto
- analisi dei rischi e delle aspettative
- studio delle competenze del gruppo di lavoro

7

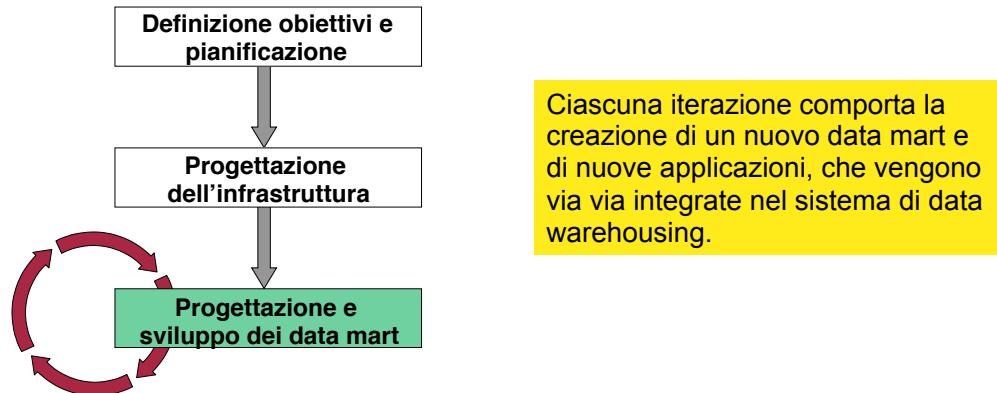
Il ciclo di sviluppo



Si analizzano e si comparano le possibili soluzioni architettoniche valutando le tecnologie e gli strumenti disponibili, al fine di realizzare un progetto di massima dell' intero sistema.

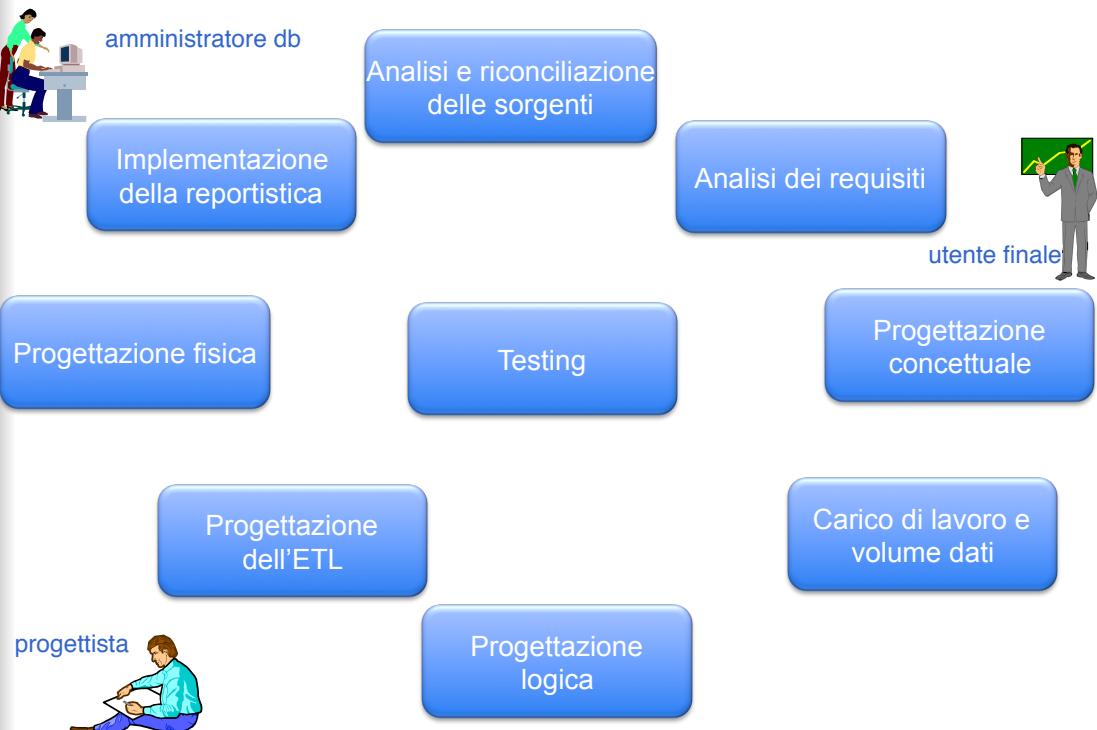
8

Il ciclo di sviluppo

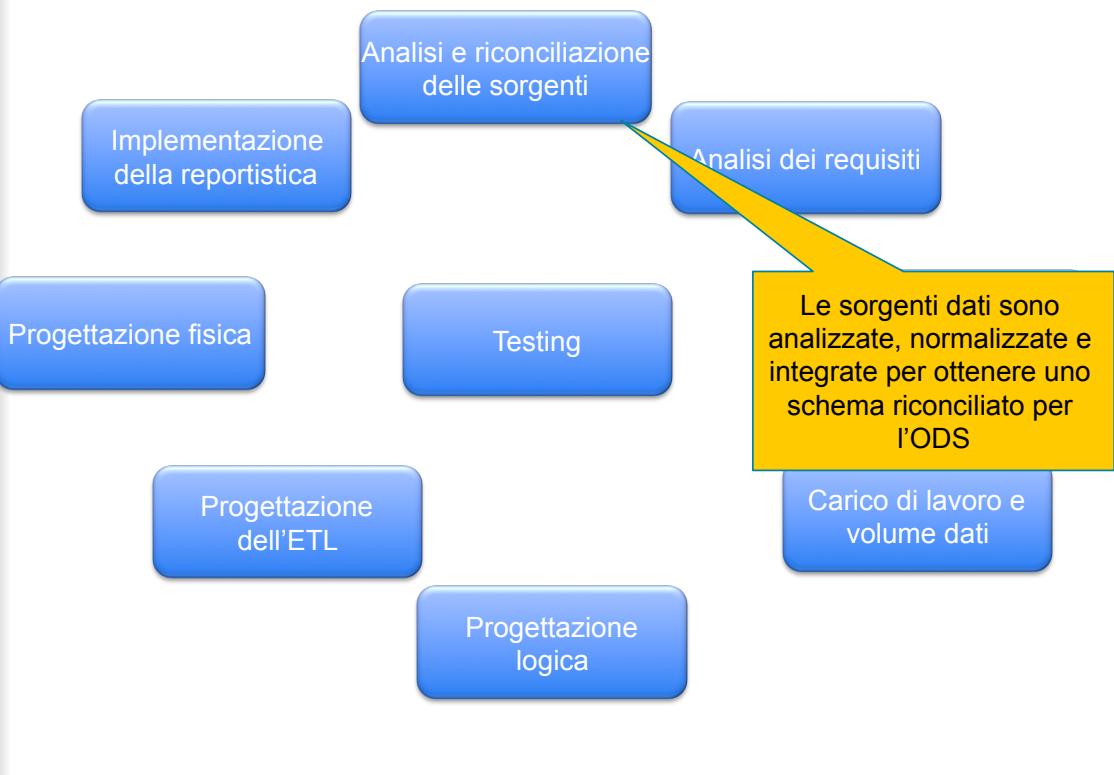


9

La progettazione di data mart



La progettazione di data mart



La progettazione di data mart



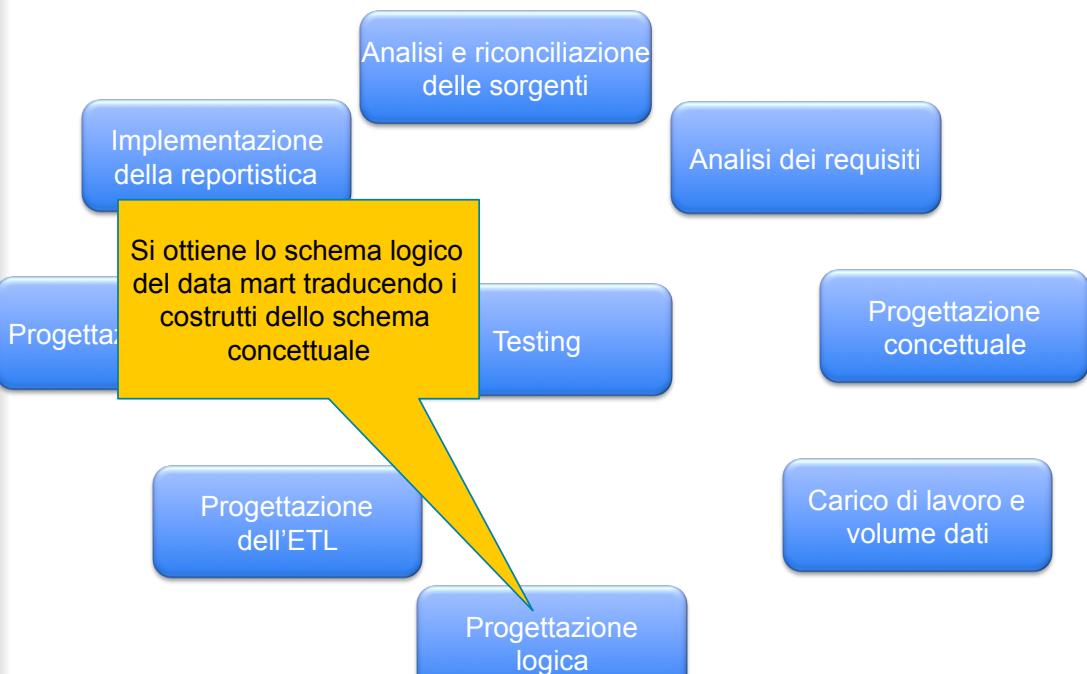
La progettazione di data mart



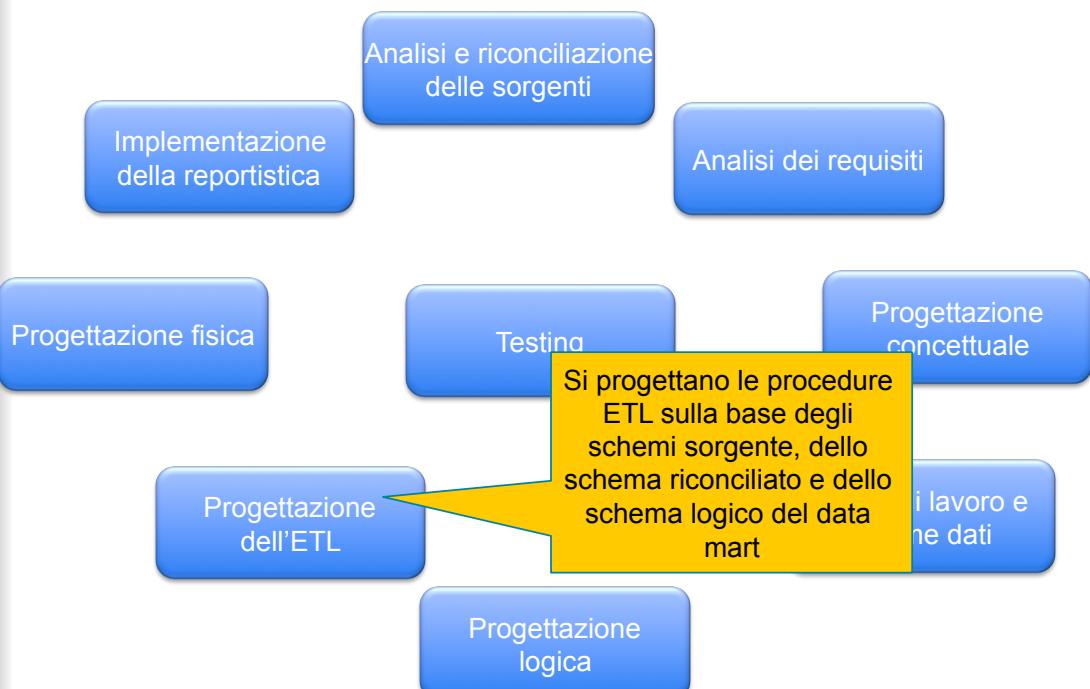
La progettazione di data mart



La progettazione di data mart



La progettazione di data mart



La progettazione di data mart



La progettazione di data mart



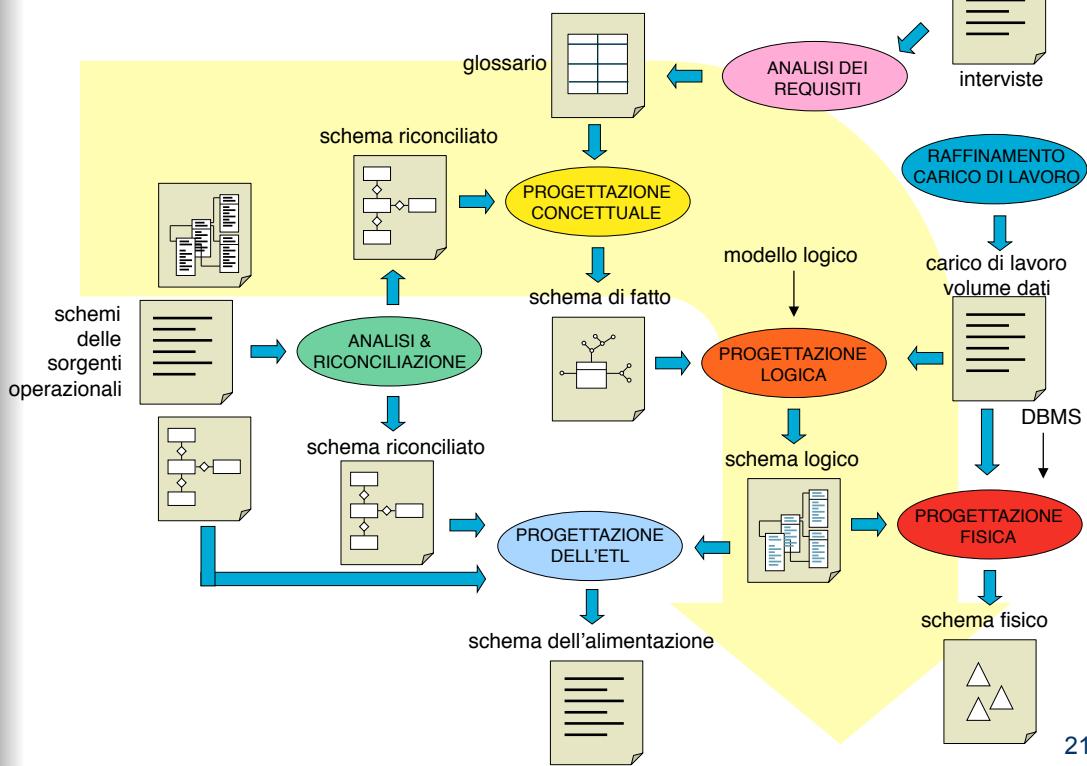
La progettazione di data mart



Quadro metodologico

- **Approcci guidati dai dati (*supply-driven*)**
 - ✓ progettano il data mart a partire da una dettagliata analisi delle sorgenti operazionali
 - ✓ i requisiti utente impattano sul progettista nella selezione delle porzioni di dati considerate rilevanti per il processo decisionale, e determinando la loro strutturazione secondo il modello multidimensionale
- **Approcci guidati dai requisiti (*demand-driven*)**
 - ✓ iniziano determinando i requisiti informativi degli utenti del data mart
 - ✓ il problema di come creare una mappatura tra questi requisiti e le sorgenti dati disponibili viene affrontato solo in seguito, attraverso l'implementazione di procedure ETL adatte

Approccio guidato dai dati



21

Approccio guidato dai dati

■ Vantaggi

- ✓ uno schema concettuale di massima per il data mart può essere derivato algoritmamente a partire dal livello dei dati riconciliati, ossia in funzione della struttura delle sorgenti
- ✓ la progettazione dell' ETL risulta notevolmente semplificata, poiché ciascuna informazione nel data mart è direttamente associata a uno o più attributi delle sorgenti

■ Svantaggi

- ✓ ai requisiti utente viene assegnato un ruolo secondario nel determinare i contenuti informativi per l' analisi
- ✓ al progettista viene dato un supporto limitato per l' identificazione di fatti, dimensioni e misure

22

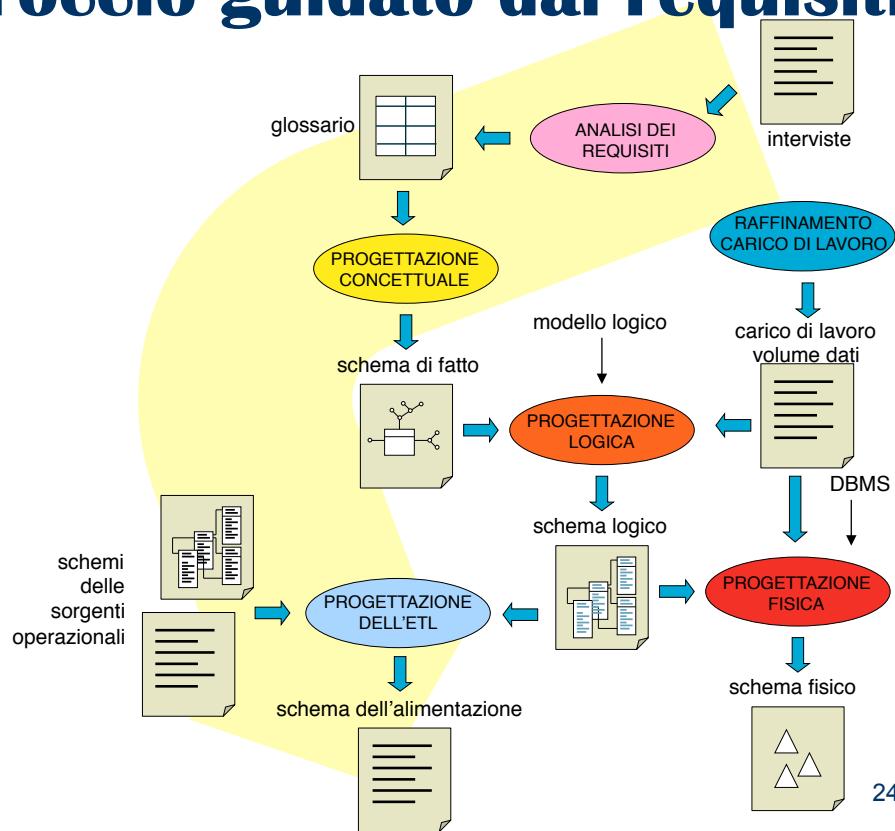
Approccio guidato dai dati

■ Applicabilità

- ✓ E' applicabile quando:
 1. è disponibile preliminarmente, oppure ottenibile con costi e tempi contenuti, una conoscenza approfondita delle sorgenti da cui il data mart si alimenterà;
 2. gli schemi delle sorgenti mostrano un buon grado di normalizzazione;
 3. la complessità degli schemi delle sorgenti non è eccessiva
- ✓ Quando l' architettura prescelta prevede l' adozione di un livello riconciliato questi requisiti sono soddisfatti: la normalizzazione e la conoscenza approfondita sono garantite dalla riconciliazione. Lo stesso vale nel caso in cui la sorgente si riduca a un singolo database, ben progettato e di dimensioni limitate
- ✓ L' esperienza di progettazione mostra che, qualora applicabile, l' approccio guidato dai dati risulta preferibile agli altri poiché permette di raggiungere i risultati prefissati in tempi estremamente contenuti

23

Approccio guidato dai requisiti



24

Approccio guidato dai requisiti

■ Vantaggi

- ✓ i desiderata degli utenti vengono portati in primo piano

■ Svantaggi

- ✓ è richiesto al progettista uno sforzo consistente durante il disegno dell' alimentazione
- ✓ fatti, misure e gerarchie vengono desunte direttamente dalle specifiche dettate dagli utenti, e solo a posteriori si verifica che le informazioni richieste siano effettivamente disponibili nei database operazionali
- ✓ la fiducia del cliente verso il progettista e verso l' utilità del data mart può venir meno

25

Approccio guidato dai requisiti

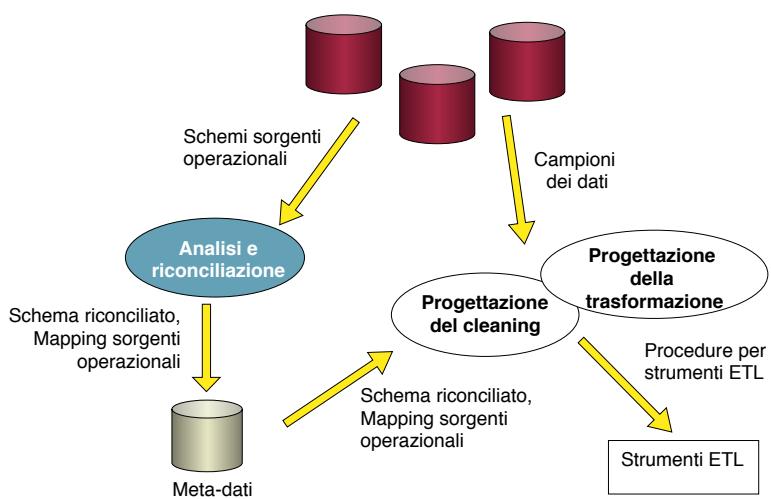
■ Applicabilità

- ✓ Questo approccio costituisce l' unica alternativa nei casi in cui non sia fattibile a priori un' analisi approfondita delle sorgenti (per esempio quando il data mart viene alimentato da un sistema ERP), oppure qualora le sorgenti siano rappresentate da sistemi legacy di tale complessità da sconsigliarne la ricognizione e la normalizzazione
- ✓ E' più difficilmente perseguitabile dell' approccio guidato dai dati

26

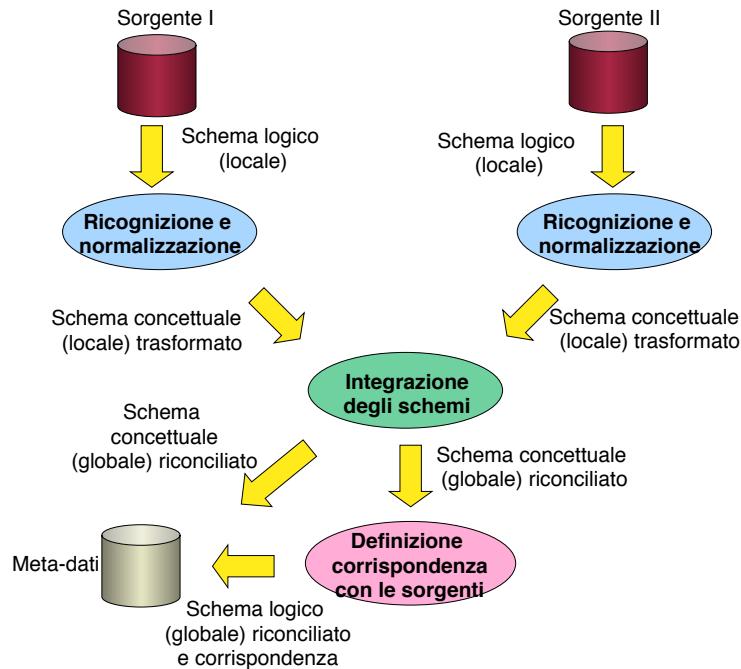
Analisi e riconciliazione delle sorgenti operazionali

Progettazione del livello riconciliato



- ✓ La fase di integrazione è incentrata sulla componente intensionale delle sorgenti operazionali, ossia riguarda la consistenza degli schemi che le descrivono
- ✓ Pulizia e trasformazione dei dati operano a livello estensionale, ossia coinvolgono direttamente i dati veri e propri

Analisi e riconciliazione delle sorgenti operazionali



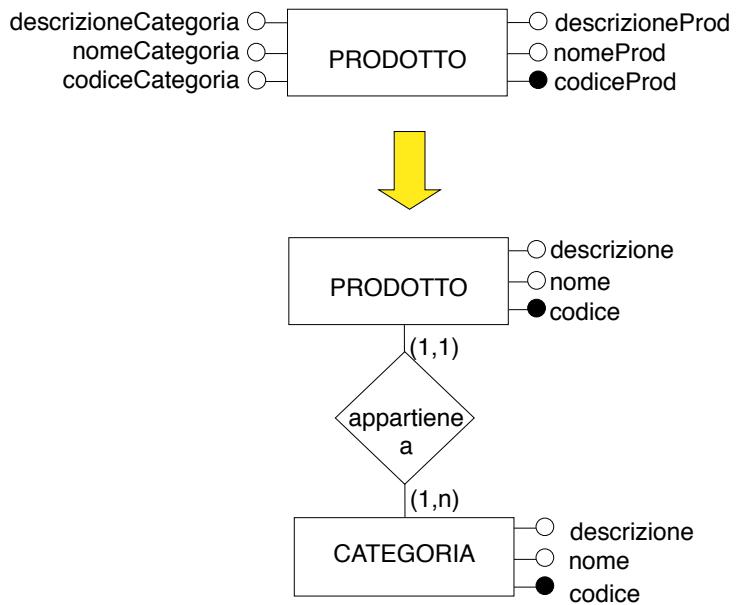
29

Ricognizione e normalizzazione

- Il progettista, confrontandosi con gli esperti del dominio applicativo, acquisisce un' approfondita conoscenza delle sorgenti operazionali attraverso:
 - ✓ *ricognizione*, che consiste in un esame approfondito degli schemi locali mirato alla piena comprensione del dominio applicativo;
 - ✓ *normalizzazione*, il cui obiettivo è correggere gli schemi locali al fine di modellare in modo più accurato il dominio applicativo
- Ricognizione e normalizzazione devono essere svolte anche qualora sia presente una sola sorgente dati; qualora esistano più sorgenti, l' operazione dovrà essere ripetuta per ogni singolo schema

30

Ricognizione e normalizzazione



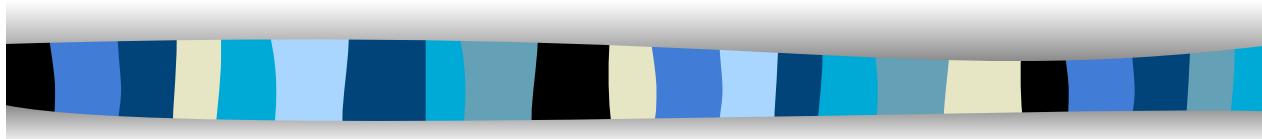
31

Integrazione

- L'integrazione di un insieme di sorgenti dati eterogenee (basi di dati relazionali, file dati, sorgenti legacy) consiste nell' individuazione delle corrispondenze tra i concetti rappresentati negli schemi locali e nella risoluzione dei conflitti evidenziati, finalizzate alla creazione di un unico schema globale i cui elementi possano essere correlati con i corrispondenti elementi degli schemi locali (*mapping*)
- La fase di integrazione non si deve limitare a evidenziare le differenze di rappresentazione dei concetti comuni a più schemi locali, ma deve anche identificare l'insieme di concetti distinti e memorizzati in schemi differenti che sono correlati attraverso proprietà semantiche (*proprietà interschema*)
- Per poter ragionare sui concetti espressi negli schemi delle diverse sorgenti dati è necessario utilizzare **un unico formalismo** in modo da fissare i costrutti utilizzabili e la potenza espressiva

32

Analisi dei requisiti



Obiettivi

- La fase di analisi dei requisiti ha l' obiettivo di raccogliere le esigenze di utilizzo del data mart espresse dai suoi utenti finali
- Essa ha un' importanza strategica poiché influenza le decisioni da prendere riguardo:
 - ✓ lo schema concettuale dei dati
 - ✓ il progetto dell' alimentazione
 - ✓ le specifiche delle applicazioni per l' analisi dei dati
 - ✓ il piano di avviamento e formazione
 - ✓ le linee guida per la manutenzione e l' evoluzione del sistema

Fonti

- La “fonte” principale da cui attingere i requisiti sono i futuri utenti del data mart (*business users*)
 - ✓ La differenza nel linguaggio usato da progettisti e utenti, e la percezione spesso distorta che questi ultimi hanno del processo di warehousing, rendono il dialogo difficile e a volte infruttuoso
- Per gli aspetti più tecnici, saranno gli amministratori del sistema informativo e/o i responsabili del CED a fungere da riferimento per il progettista
 - ✓ In questo caso, i requisiti che dovranno essere catturati riguardano principalmente vincoli di varia natura imposti sul sistema di data warehousing



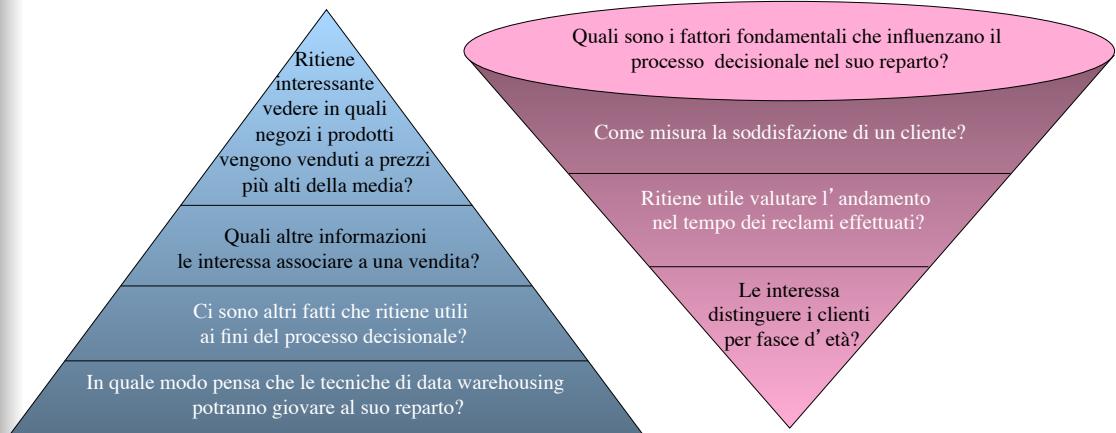
35

Le interviste

- **A piramide.** Approccio induttivo: l’ intervistatore parte da domande molto dettagliate per poi ampliare l’ argomento dell’ intervista mediante domande aperte che richiedono risposte più generali.
 - ✓ Questo tipo di intervista permette di superare la riluttanza di un intervistato scettico poiché inizialmente non richiede un forte coinvolgimento da parte dell’ intervistato.
- **A imbuto.** Approccio deduttivo: l’ intervistatore parte da domande molto generali per poi restringere l’ argomento dell’ intervista a temi specifici
 - ✓ Questo approccio è utile nel caso in cui l’ intervistato sia emozionato o eccessivamente deferente, poiché il fatto che le domande di carattere generale (normalmente in forma aperta) non prevedano una risposta “sbagliata” allevia la tensione dell’ intervistato.

36

Le interviste



37

Le domande

Ruolo	Domande chiave
Dirigente	Quali sono gli obiettivi aziendali? Come misuri il successo della tua azienda? Quali sono oggi i principali problemi dell'azienda? In che modo ti aspetti che una maggiore disponibilità di informazioni possa migliorare la situazione aziendale?
Direttore di reparto	Quali sono gli obiettivi del tuo reparto? Come misuri il successo del tuo reparto? Descrivi i soggetti coinvolti nel tuo settore di interesse. Ci sono colli di bottiglia nell'accesso ai dati? Che analisi di routine esegui? Che tipi di analisi ti piacerebbe poter eseguire? A che livello di dettaglio occorre vedere le informazioni? Quanta informazione storica è necessaria?
Amministratore del sistema informativo	Illustra le caratteristiche delle principali fonti dati disponibili. Che strumenti vengono usati per analizzare i dati? Come vengono gestite le richieste di analisi ad hoc? Quali sono i principali problemi di qualità dei dati?

38

I fatti

- I **fatti** sono i concetti su cui gli utenti finali del data mart baseranno il processo decisionale; ogni fatto descrive una categoria di eventi che si verificano in azienda
 - ✓ Fissare le dimensioni di un fatto è importante poiché significa determinarne la **granularità**, ovvero il più fine livello di dettaglio a cui i dati saranno rappresentati. La scelta della granularità di un fatto nasce da un delicato compromesso tra due esigenze contrapposte: quella di raggiungere un' elevata flessibilità d' utilizzo e quella di conseguire buone prestazioni
 - ✓ Per ogni fatto occorre definire l' **intervallo di storicizzazione**, ovvero l'arco temporale che gli eventi memorizzati dovranno coprire

39

I fatti

	Data mart	Fatti
commerciale/ manifatturiero	approvvigionamenti	acquisti, inventario di magazzino, distribuzione
	produzione	confezionamento, inventario, consegna, manifattura
	gestione domanda	vendite, fatturazione, ordini, spedizioni, reclami
	marketing	promozioni, fidelizzazione, campagne pubblicitarie
finanziario	bancario	conti correnti, bonifici, prestiti ipotecari, mutui
	investimenti	acquisto titoli, transazioni di borsa
	servizi	carte di credito, domiciliazioni bollette
sanitario	scheda di ricovero	ricoveri, dimissioni, interventi chirurgici, diagnosi
	pronto soccorso	accessi, esami, dimissioni
	medicina di base	scelte, revoche, prescrizioni
trasporti	merci	domanda, offerta, trasporti
	passeggeri	domanda, offerta, trasporti
	manutenzione	interventi
telecomunicazioni	traffico	traffico in rete, chiamate
	CRM	fidelizzazione, reclami, servizi
turismo	gestione domanda	biglietteria, noleggi auto, soggiorni
	CRM	frequent-flyers, reclami
gestionale	logistica	trasporti, scorte, movimentazione
	risorse umane	assunzioni, dimissioni, promozioni, incentivi
	budgeting	budget commerciale, budget di marketing
	infrastrutture	acquisti, opere

40

Glossario dei requisiti

Fatto	Possibili dimensioni	Possibili misure	Storicità
inventario di magazzino	prodotto, data, magazzino	quantità in magazzino	1 anno
vendite	prodotto, data, negozio	quantità venduta, importo, sconto	5 anni
linee d'ordine	prodotto, data, fornitore	quantità ordinata, importo, sconto	3 anni

41

Il carico di lavoro preliminare

- Il riconoscimento di fatti, dimensioni e misure è strettamente collegato all' identificazione di un *carico di lavoro preliminare*.
 - ✓ Oltre che dall' interazione diretta con l' utente, indicazioni al riguardo potranno essere ricavate da un esame della reportistica correntemente in uso in azienda.
 - ✓ In questa fase il carico di lavoro può essere espresso in linguaggio naturale; esso sarà comunque utile per valutare la granularità dei fatti e le misure di interesse, nonché per iniziare ad affrontare il problema dell' aggregazione

42

Il carico di lavoro preliminare

Fatto	Interrogazione
inventario di magazzino	Quantità media di ciascun prodotto presente mensilmente in tutti i magazzini. Prodotti per i quali è stata esaurita la scorta contemporaneamente in tutti i magazzini in almeno un'occasione durante la settimana passata. Andamento giornaliero delle scorte complessive per ciascun tipo di prodotto.
vendite	Quantità totali di ciascun tipo di prodotto vendute durante l'ultimo mese. Incasso totale giornaliero di ciascun negozio. Per un dato negozio, incassi relativi alle diverse categorie di prodotti durante un certo giorno. Riepilogo annuale degli incassi per regione relativamente a un dato prodotto.
linee d'ordine	Quantità totale ordinata annualmente presso un certo fornitore. Importo giornaliero ordinato nell'ultimo mese per un certo tipo di prodotto. Sconto massimo applicato da ciascun fornitore durante l'ultimo anno per ciascuna categoria di prodotto.

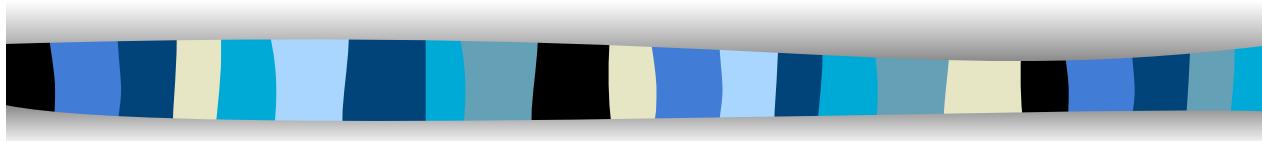
43

Altri requisiti

- **Vincoli di progettazione logica e fisica (spazio disponibile)**
- **Progetto dell' alimentazione (periodicità dell' alimentazione)**
- **Architettura del sistema di data warehousing** (tipo di architettura da implementare, numero dei livelli, presenza di data mart dipendenti o indipendenti, materializzazione del livello riconciliato)
- **Applicazioni per l' analisi dei dati** (disamina delle tipologie di interrogazioni e dei rapporti analitici normalmente richiesti)
- **Piano di avviamento**
- **Piano di formazione**

44

Progettazione concettuale



Quale formalismo?

- Mentre è universalmente riconosciuto che un DW si appoggia sul modello multidimensionale, non c' è accordo sulla metodologia di progetto concettuale
- Il modello Entity/Relationship è molto diffuso nelle imprese come formalismo per la documentazione dei sistemi informativi relazionali, ma *non può essere usato per modellare il DW*
- Alcuni progettisti di DW disegnano direttamente gli schemi a stella: ma uno schema a stella non è altro che uno schema relazionale, e racchiude pertanto solo la definizione di un insieme di relazioni e di vincoli di integrità!



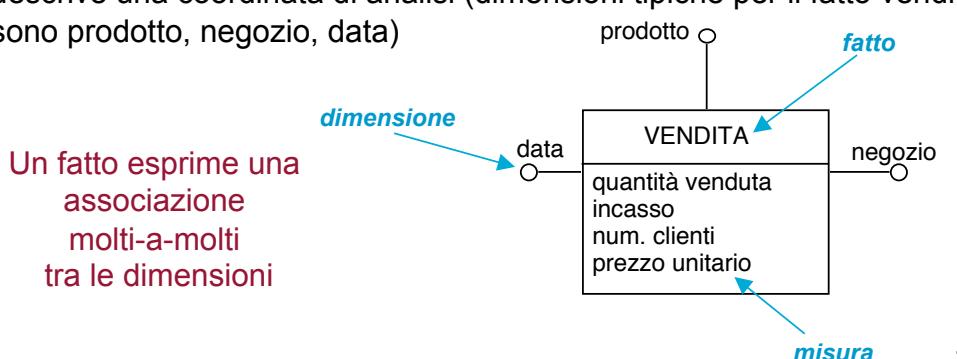
Il Dimensional Fact Model

- Il DFM è un modello concettuale grafico per data mart, pensato per:
 - ✓ supportare efficacemente il progetto concettuale;
 - ✓ creare un ambiente su cui formulare in modo intuitivo le interrogazioni dell'utente;
 - ✓ permettere il dialogo tra progettista e utente finale per raffinare le specifiche dei requisiti;
 - ✓ creare una piattaforma stabile da cui partire per il progetto logico (*indipendentemente dal modello logico target*);
 - ✓ restituire una documentazione a posteriori espressiva e non ambigua.
- La rappresentazione concettuale generata dal DFM consiste in un insieme di **schemi di fatto**. Gli elementi di base modellati dagli schemi di fatto sono i fatti, le misure, le dimensioni e le gerarchie

47

Il DFM: costrutti di base

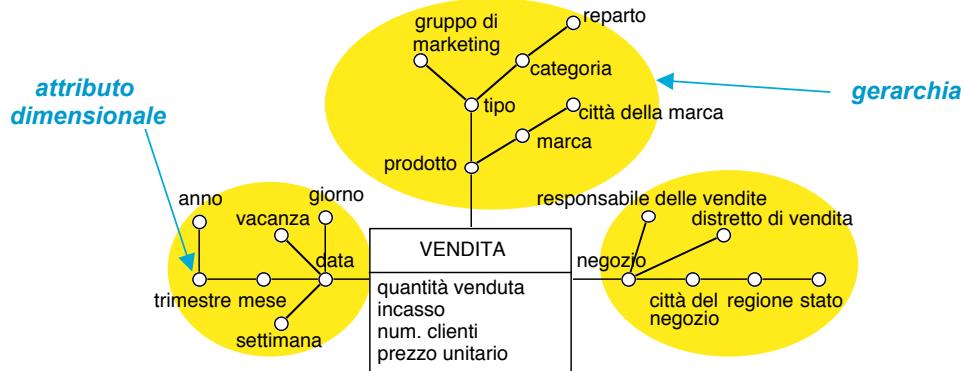
- Un **fatto** è un concetto di interesse per il processo decisionale; tipicamente modella un insieme di eventi che accadono nell'impresa (ad esempio: vendite, spedizioni, acquisti, ...). È essenziale che un fatto abbia aspetti dinamici, ovvero evolva nel tempo
- Una **misura** è una proprietà numerica di un fatto e ne descrive un aspetto quantitativo di interesse per l'analisi (ad esempio, ogni vendita è misurata dal suo incasso)
- Una **dimensione** è una proprietà con dominio finito di un fatto e ne descrive una coordinata di analisi (dimensioni tipiche per il fatto vendite sono prodotto, negozio, data)



48

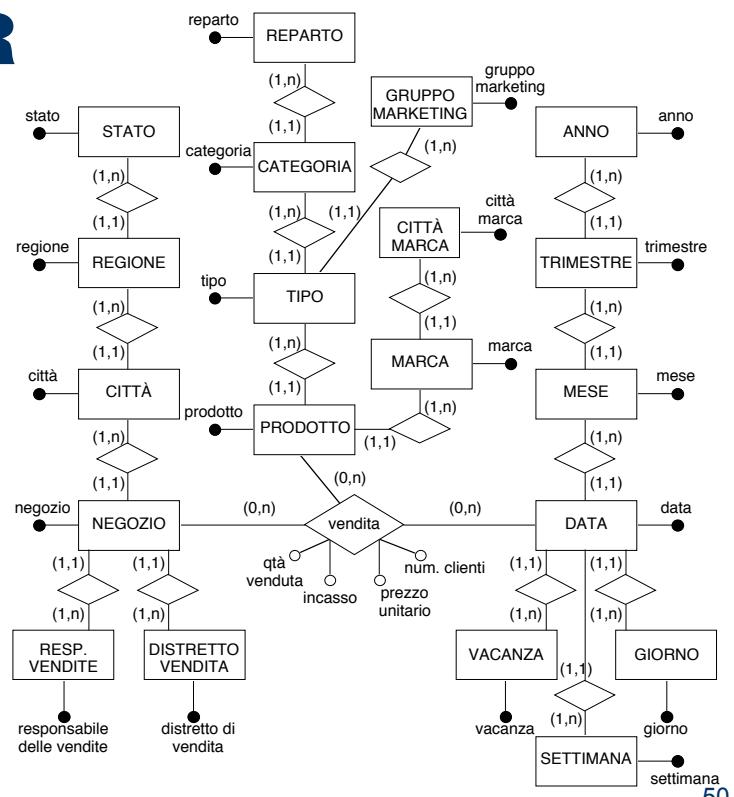
Il DFM: costrutti di base

- Con il termine generale **attributi dimensionali** si intendono le dimensioni e gli eventuali altri attributi, sempre a valori discreti, che le descrivono (per esempio, un prodotto è descritto dal suo tipo, dalla categoria cui appartiene, dalla sua marca, dal reparto in cui è venduto)
- Una **gerarchia** è un albero direzionale i cui nodi sono attributi dimensionali e i cui archi modellano associazioni multi-a-uno tra coppie di attributi dimensionali. Essa racchiude una dimensione, posta alla radice dell' albero, e tutti gli attributi dimensionali che la descrivono



49

Il DFM: corrispondenza con l'E/R



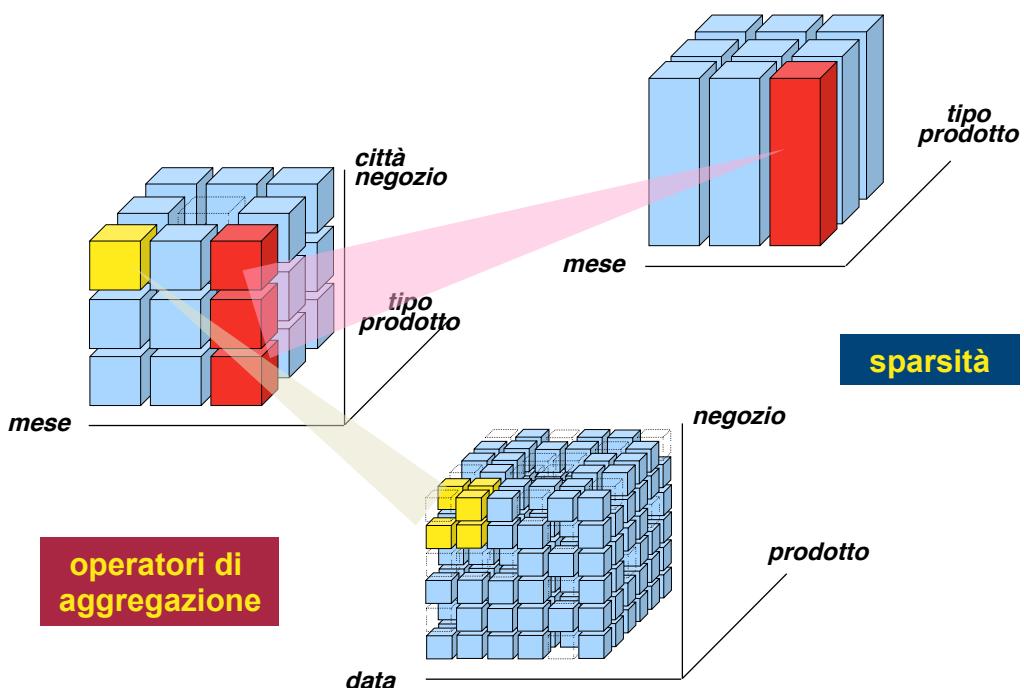
50

Eventi e aggregazione

- Un **evento primario** è una particolare occorrenza di un fatto, individuata da una ennupla costituita da un valore per ciascuna dimensione. A ciascun evento primario è associato un valore per ciascuna misura
 - ✓ Con riferimento alle vendite, un possibile evento primario registra per esempio che, il 10/10/2001, nel negozio NonSoloPappa sono state vendute 10 confezioni di detersivo Brillo per un incasso complessivo pari a 25 euro
- Dato un insieme di attributi dimensionali (**group-by set**), ciascuna ennupla di loro valori individua un **evento secondario** che aggrega tutti gli eventi primari corrispondenti. A ciascun evento secondario è associato un valore per ciascuna misura, che riassume in sé tutti i valori della stessa misura negli eventi primari corrispondenti
 - ✓ Pertanto, le gerarchie definiscono il modo in cui gli eventi primari possono essere aggregati e selezionati significativamente per il processo decisionale; mentre la dimensione in cui una gerarchia ha radice ne definisce la granularità più fine di aggregazione, agli altri attributi dimensionali corrispondono granularità via via crescenti

51

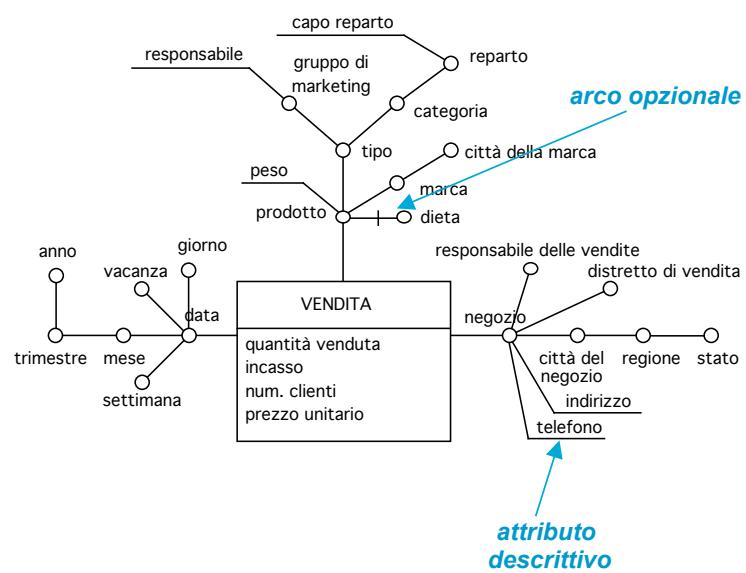
Eventi e aggregazione



52

II DFM: costrutti avanzati

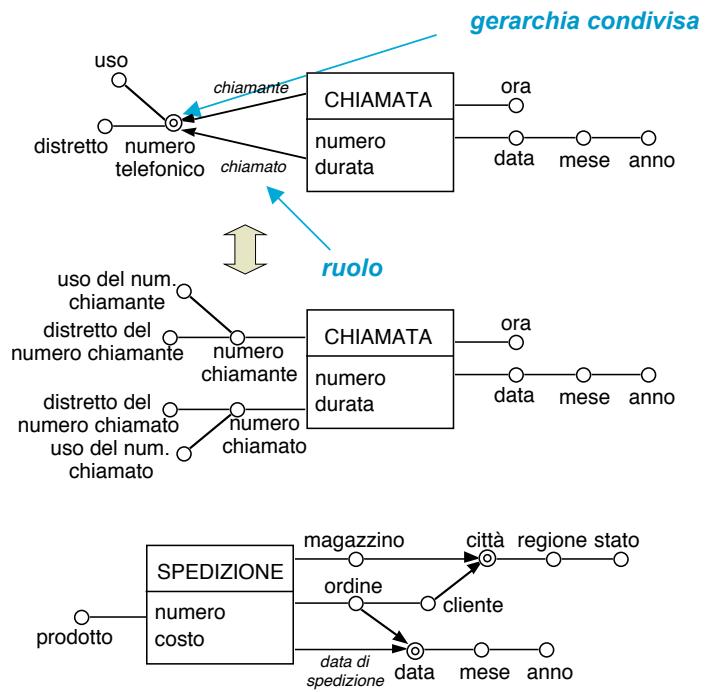
- Un *attributo descrittivo* contiene informazioni aggiuntive su un attributo dimensionale di una gerarchia, a cui è connesso da una associazione -a-uno. Non viene usato per l' aggregazione poiché ha valori continui e/o poiché deriva da un' associazione uno-a-uno
- Alcuni archi dello schema di fatto possono essere *opzionali*



53

II DFM: costrutti avanzati

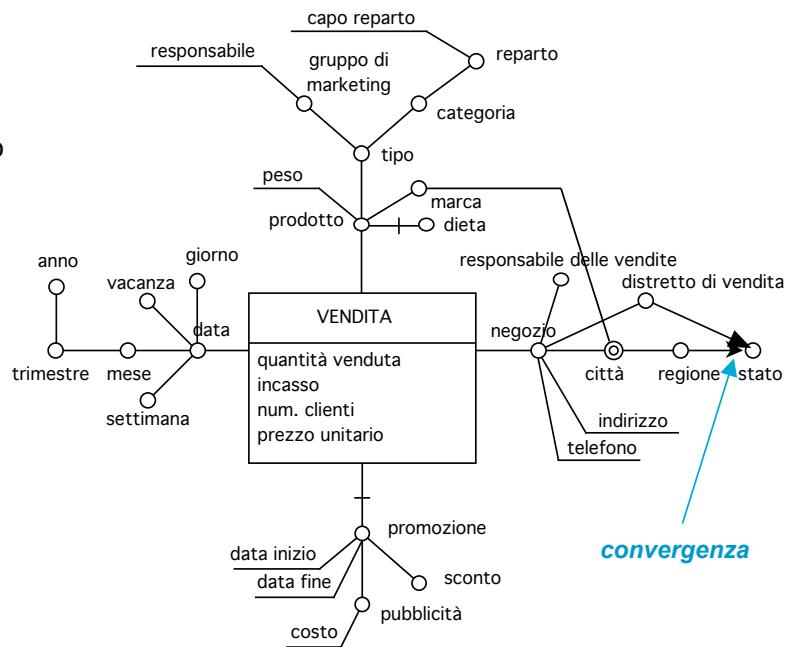
- La *gerarchia condivisa* è un' abbreviazione usata per denotare il fatto che una porzione di gerarchia è replicata più volte nello schema



54

Il DFM: costrutti avanzati

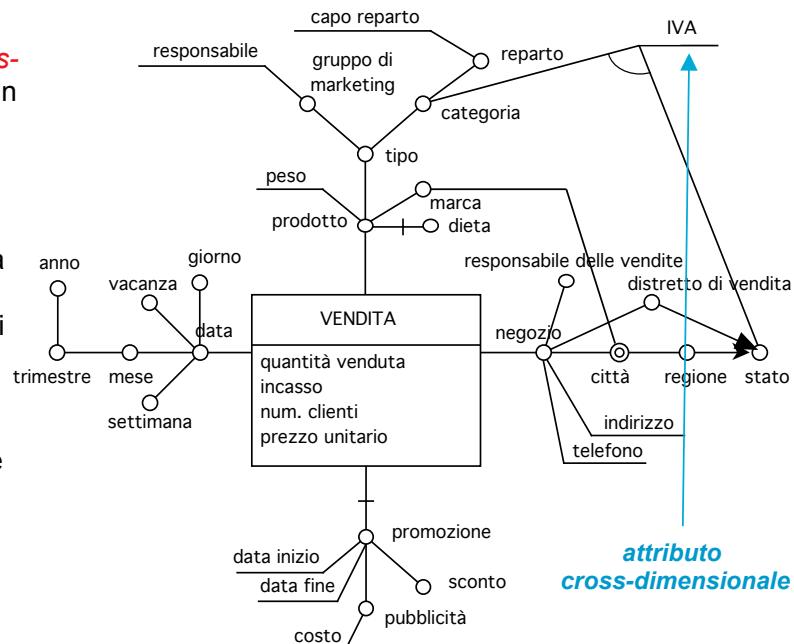
- Due attributi dimensionali possono essere connessi da due o più cammini direzionali distinti, a patto che ciascuno di essi rappresenti ancora una dipendenza funzionale (*convergenza*)



55

Il DFM: costrutti avanzati

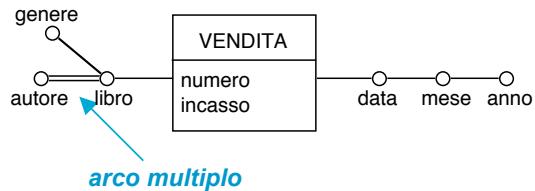
- Un *attributo cross-dimensionale* è un attributo, dimensionale o descrittivo, il cui valore è determinato dalla combinazione di due o più attributi dimensionali, eventualmente appartenenti a gerarchie distinte



56

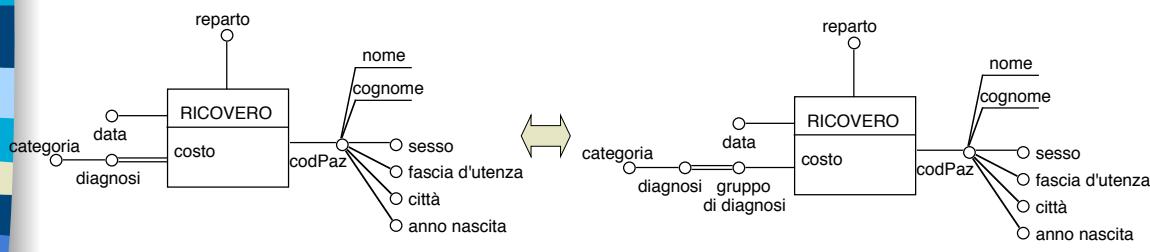
II DFM: costrutti avanzati

- Un *arco multiplo* modella un' associazione molti-a-molti tra due attributi dimensionali



Il DFM	Golfarelli, Rizzi	3
Mi Sembra Logico	Golfarelli	5
La Giusta Misura	Rizzi	10
Un Fatto Come e Perchè	Golfarelli, Rizzi	4
La Quarta Dimensione	Golfarelli	8

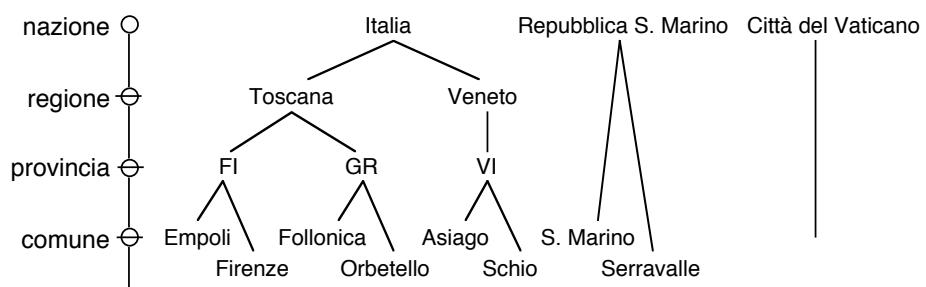
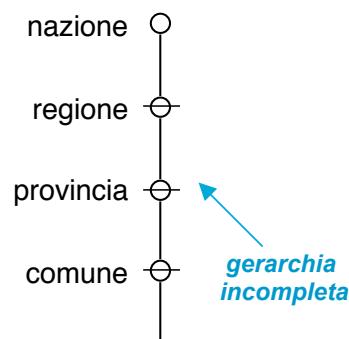
Quanto ha venduto Rizzi?



57

II DFM: costrutti avanzati

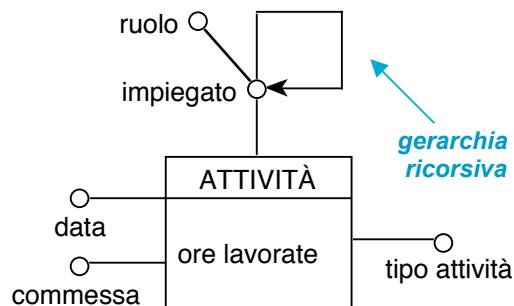
- Una *gerarchia incompleta* è una gerarchia in cui, per alcune istanze, risultano assenti (in quanto non noti oppure non definiti) uno o più livelli di aggregazione



58

II DFM: costrutti avanzati

- Nelle *gerarchie ricorsive* le relazioni padre-figlio tra i livelli sono consistenti, ma le istanze possono avere lunghezze differenti

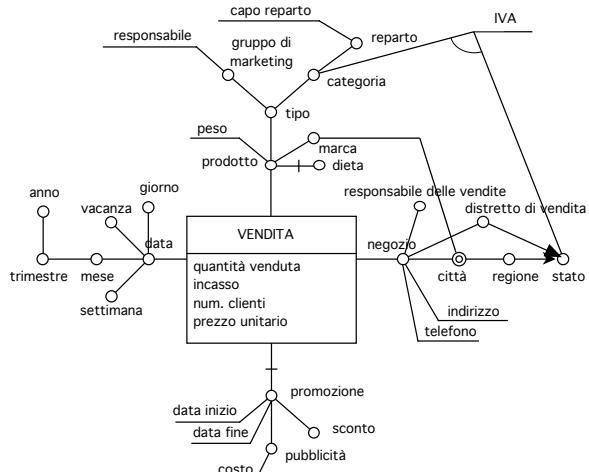


59

II DFM: costrutti avanzati

- L' *additività* esprime in che modo le misure possono essere aggregate

additività



	quantità	incasso	num. clienti	prezzo unit.
data	SUM	SUM	SUM	AVG
prodotto	SUM	SUM	---	AVG
negozio	SUM	SUM	SUM	AVG
promozione	SUM	SUM	SUM	AVG

60

Additività

- L' aggregazione richiede di definire un operatore adatto per comporre i valori delle misure che caratterizzano gli eventi primari in valori da abbinare a ciascun evento secondario
- Da questo punto di vista è possibile distinguere tre categorie di misure:
 - ✓ **Misure di flusso:** si riferiscono a un periodo, al cui termine vengono valutate in modo cumulativo (il numero di prodotti venduti in un giorno, l'incasso mensile, il numero di nati in un anno)
 - ✓ **Misure di livello:** vengono valutate in particolari istanti di tempo (il numero di prodotti in inventario, il numero di abitanti di una città)
 - ✓ **Misure unitarie:** vengono valutate in particolari istanti di tempo, ma sono espresse in termini relativi (il prezzo unitario di un prodotto, la percentuale di sconto, il cambio di una valuta)

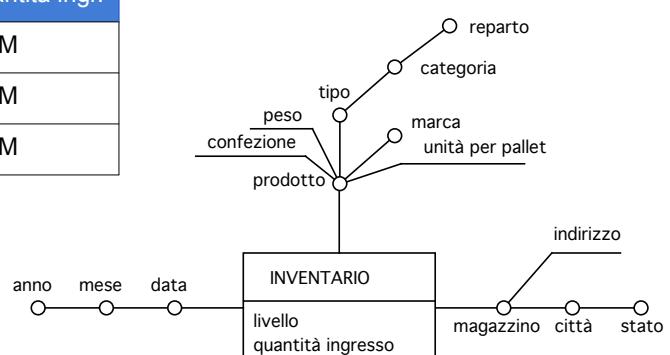
	Gerarchie temporali	Gerarchie non temporali
Misure di flusso	SUM, AVG, MIN, MAX	SUM, AVG, MIN, MAX
Misure di livello	Avg, MIN, MAX	SUM, AVG, MIN, MAX
Misure unitarie	Avg, MIN, MAX	Avg, MIN, MAX

61

Additività

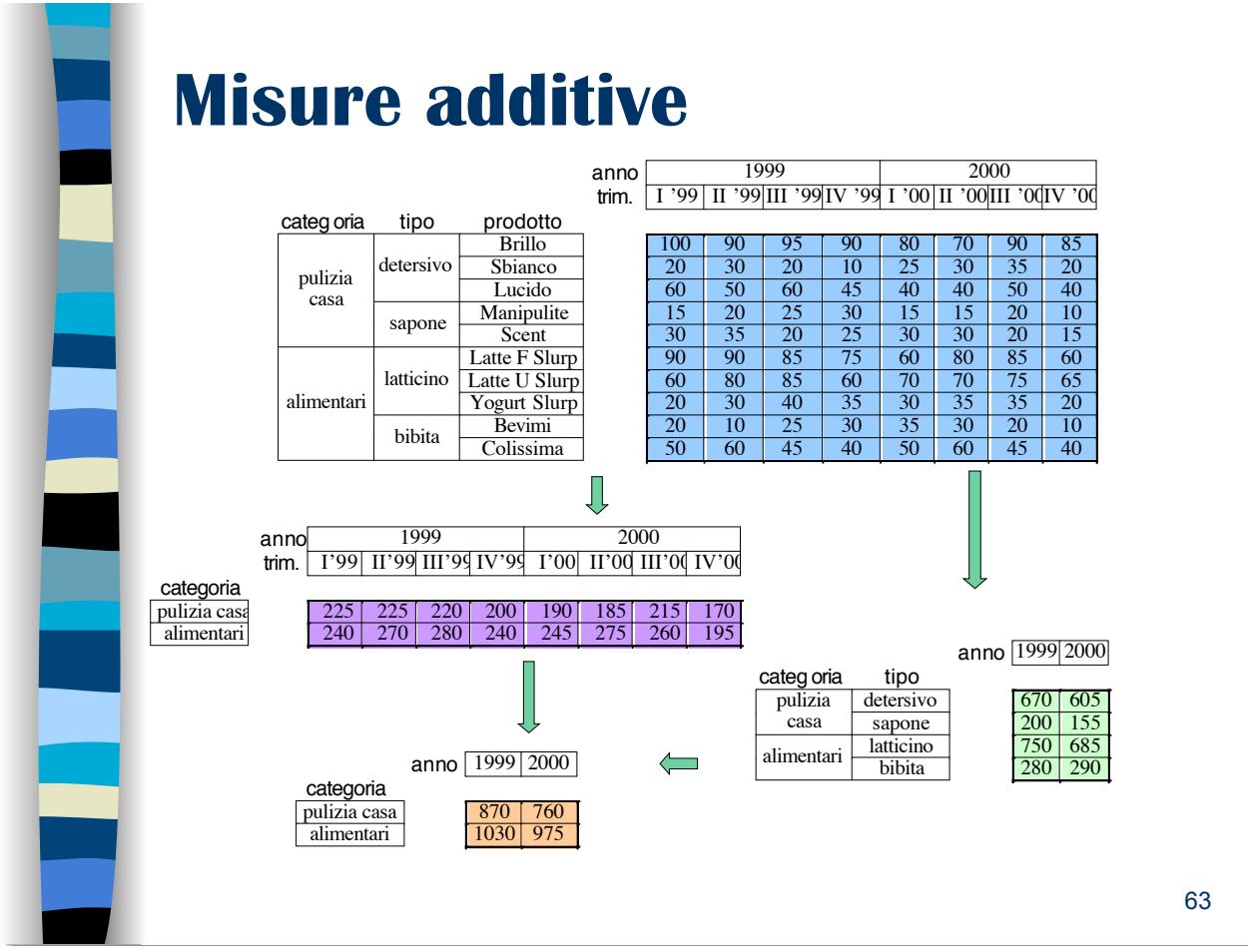
- Una misura è detta **additiva** su una dimensione se i suoi valori possono essere aggregati lungo la corrispondente gerarchia tramite l'operatore di somma, altrimenti è detta **non-additiva**. Una misura non-additiva è **non-aggregabile** se nessun operatore di aggregazione può essere usato su di essa

	livello	quantità ingr.
data	AVG,MIN	SUM
prodotto	SUM	SUM
magazzino	SUM	SUM



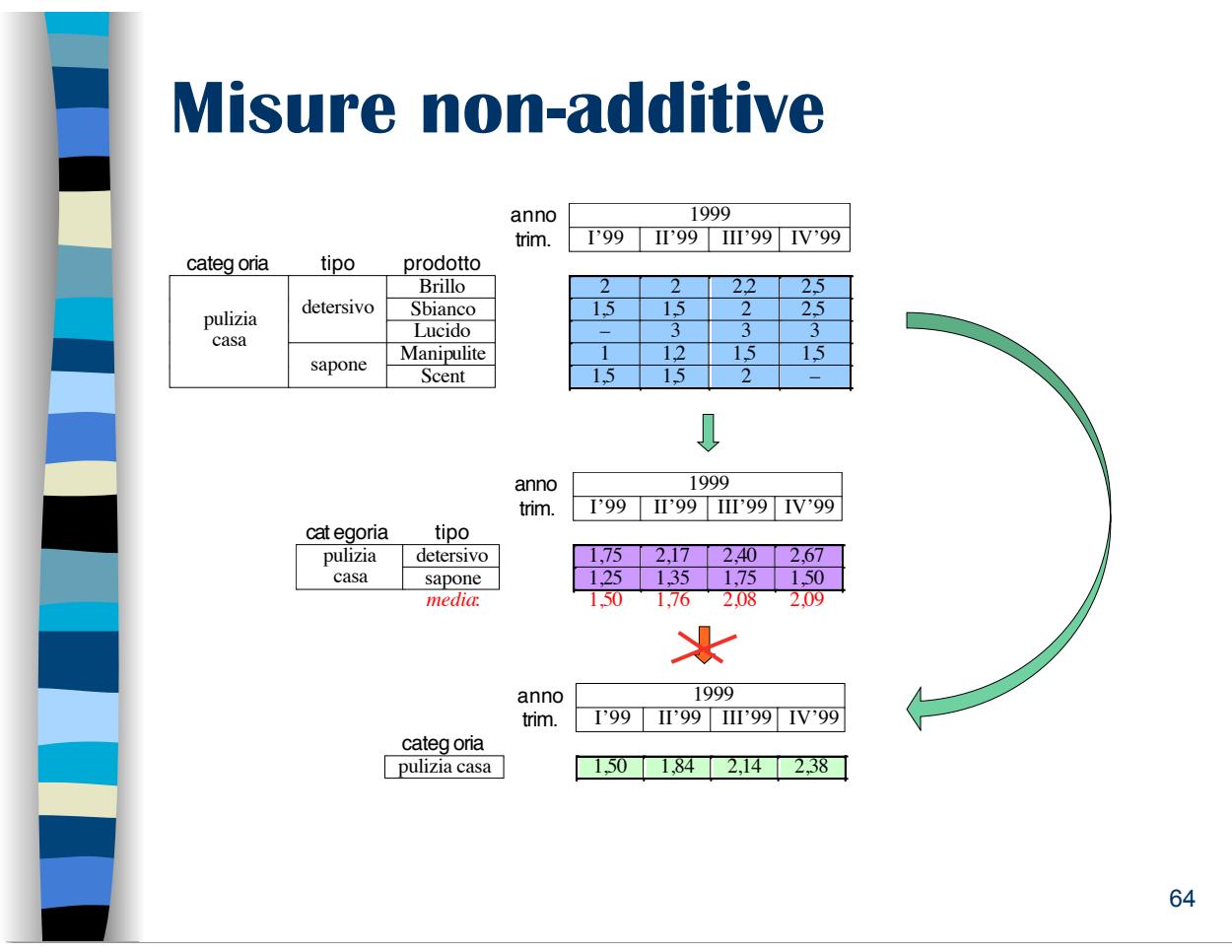
62

Misure additive



63

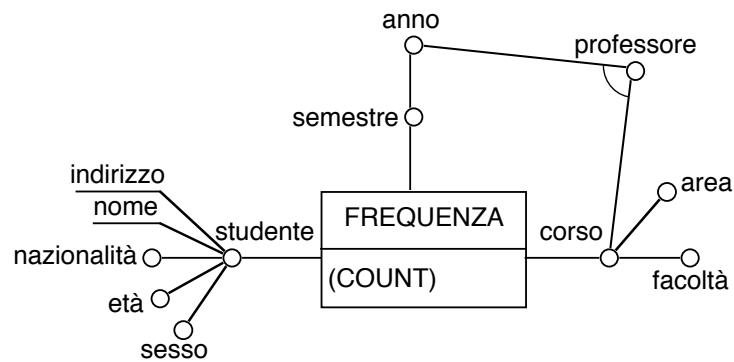
Misure non-additive



64

Schemi di fatto vuoti

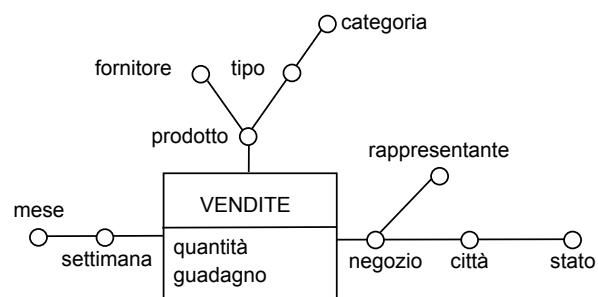
- Uno schema di fatto si dice **vuoto** se non ha misure
 - ✓ In questo caso, il fatto registra solo il verificarsi di un evento



65

Schemi di fatto transazionali

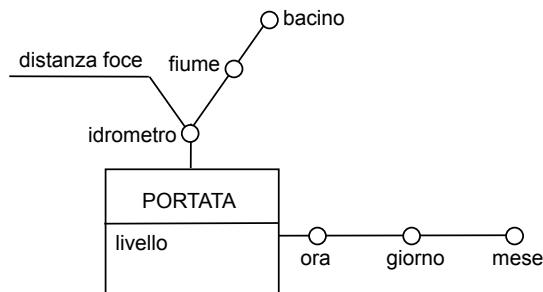
- Ciascun evento registra una singola transazione o riassume un insieme di transazioni che avvengono durante lo stesso intervallo di tempo
 - ✓ La maggior parte delle misure sono di flusso



66

Schemi di fatto istantanei

- Ciascun evento corrisponde a una fotografia periodica del fatto
 - ✓ La maggior parte delle misure sono di livello



67

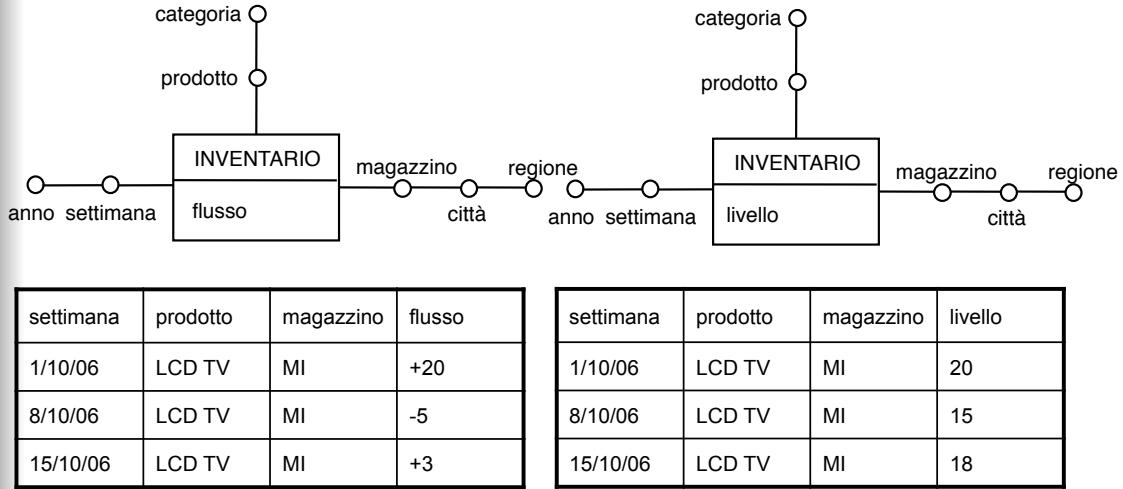
Transazionale vs. istantaneo

- Uno schema transazionale...
 - ✓ ...è la soluzione migliore se, nel dominio applicativo, gli eventi sono misurati come "flussi" entranti e uscenti (delta)
 - ✓ non può essere adottato se gli eventi sono misurati come livelli, a meno che non sia possibile decomporli univocamente in flussi
- Uno schema istantaneo...
 - ✓ ...è la soluzione migliore se, nel dominio applicativo, gli eventi sono misurati come "livelli"
 - ✓ può essere adottato anche quando gli eventi sono misurati come flussi, se è nota la funzione che compone i flussi per determinare i livelli; in questo caso, può comportare perdita di informazione
- In generale, la scelta dipende comunque anche dal carico di lavoro

68

Transazionale vs. istantaneo

- Esempio:



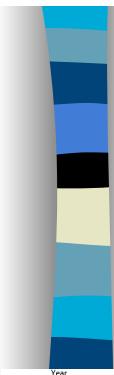
69

Il DFM in azione

Lauree universitarie

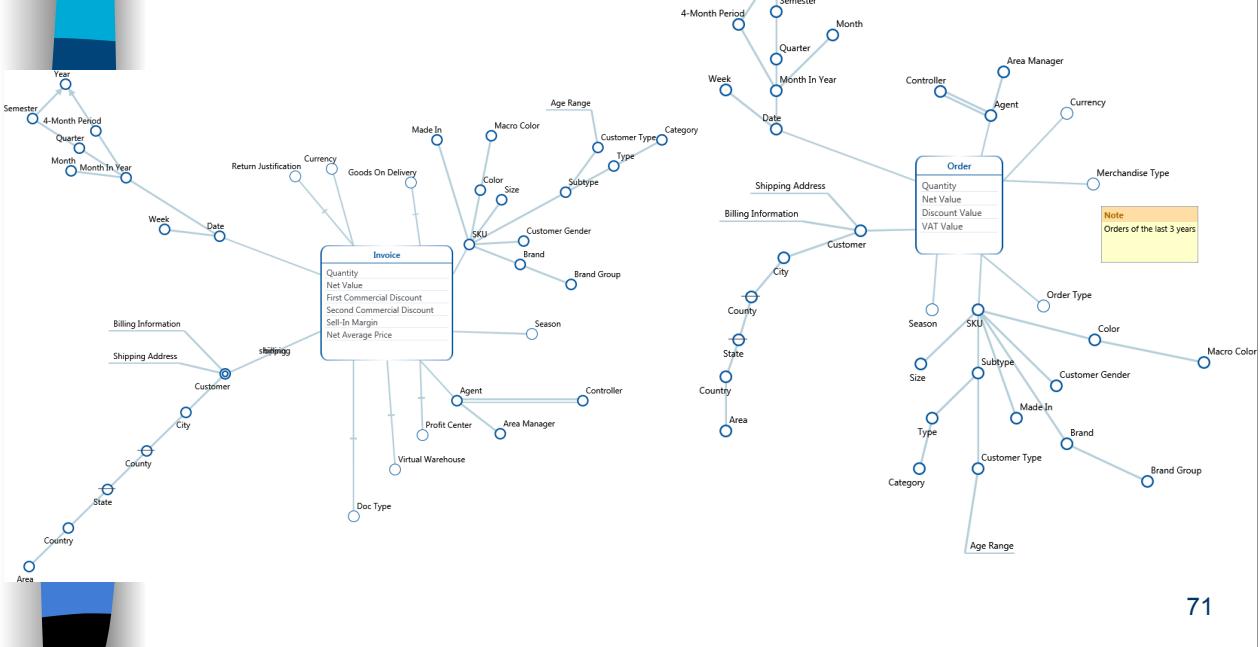


70

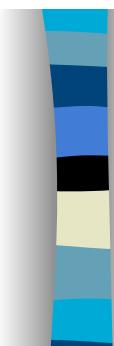


Il DFM in azione

Ordini e fatture

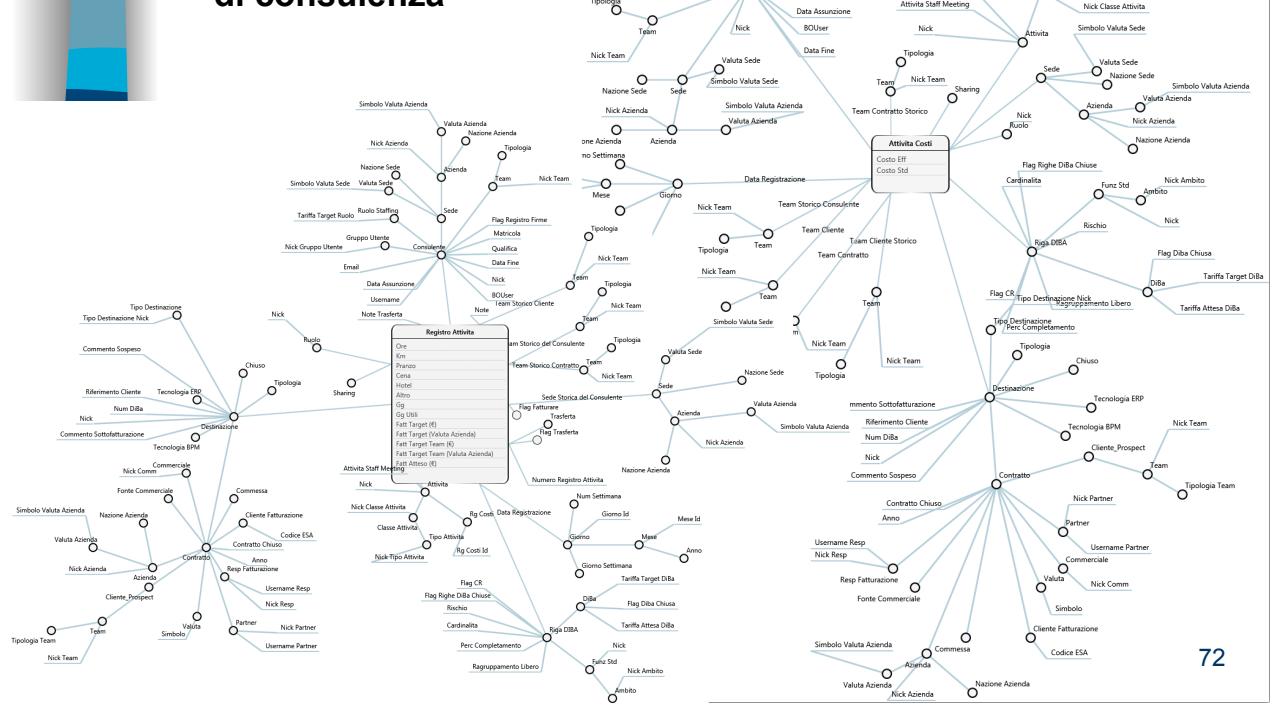


71

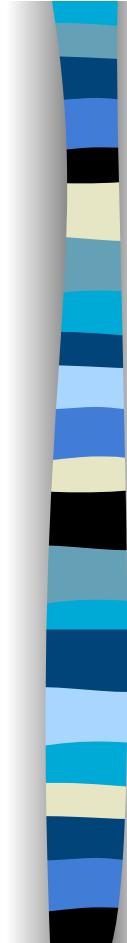


Il DFM in azione

Attività di un'azienda di consulenza



72



Progettazione concettuale: approcci

- Approccio demand-driven
 - ✓ Il progettista deve essere in grado di enucleare, dalle interviste condotte presso l'utente, un'indicazione precisa circa i fatti da rappresentare, le misure che li descrivono e le gerarchie attraverso cui aggregarli utilmente. Il problema del collegamento tra lo schema concettuale così determinato e le sorgenti operazionali viene affrontato in un secondo tempo
- Approccio supply-driven 
 - ✓ È possibile definire lo schema concettuale in funzione della struttura delle sorgenti, evitando il complesso compito di stabilire il legame con esse a posteriori. Inoltre, è possibile derivare uno schema concettuale prototipale dagli schemi operazionali in modo pressoché automatico

73

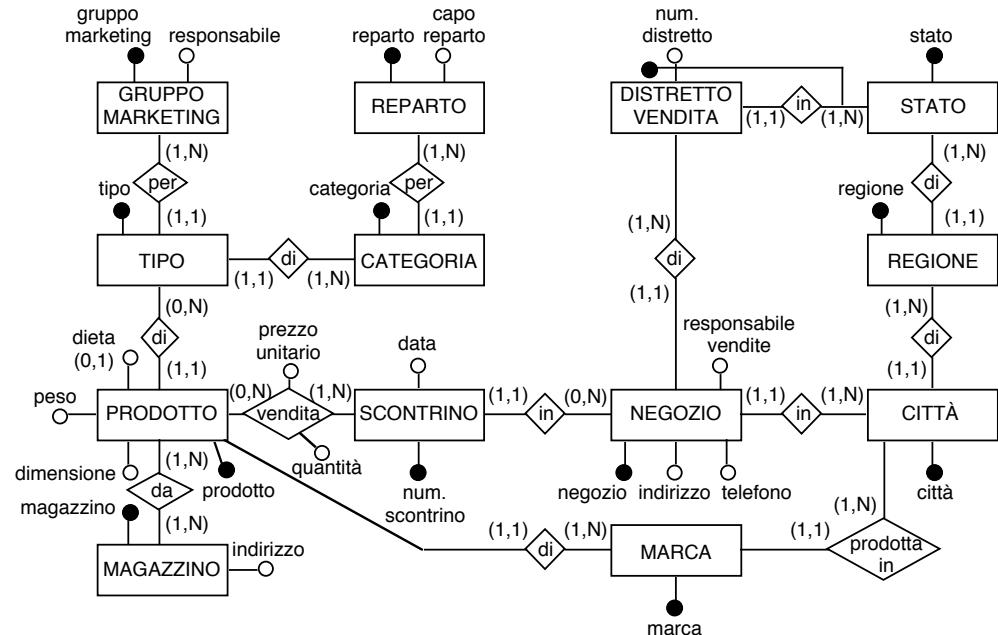


Progettazione concettuale: come

- La progettazione concettuale viene effettuata a partire dalla documentazione relativa al database riconciliato:
 - ✓ Schemi E/R
 - ✓ Schemi Relazionali
 - ✓ Schemi XML
 - ✓
- Passi di progettazione:
 - ① Scelta dei fatti
 - ② Per ogni fatto:
 1. Costruzione di un *albero degli attributi*
 2. Editing dell' albero degli attributi
 3. Scelta delle dimensioni
 4. Scelta delle misure
 5. Creazione dello schema di fatto

74

L'esempio delle vendite (da E/R)



75

L'esempio delle vendite (da schema logico)

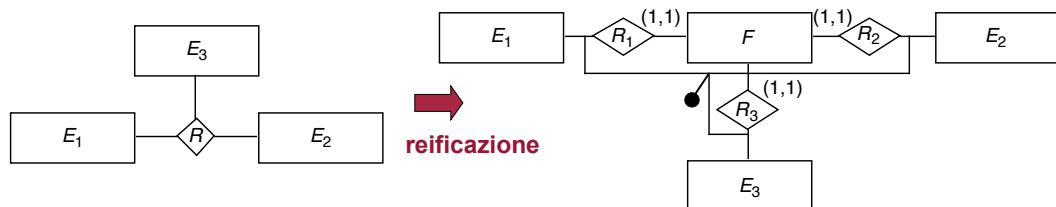
PRODOTTI (prodotto, peso, dimensione, dieta,
diMarca:**MARCHE**, diTipo:**TIPI**)
NEGOZI (negozi, indirizzo, telefono, respVendite,
numDistr, stato):**DISTRETTI**, inCittà:**CITTÀ**)
SCONTRINI (numScontrino, data, negozio:**NEGOZI**)
VENDITE (prodotto:**PRODOTTI**, numScontrino:**SCONTRINI**,
quantità, prezzoUnitario)
MAGAZZINI (magazzino, indirizzo)
CITTÀ (città, regione:**REGIONI**)
REGIONI (regione, stato:**STATI**)
STATI (stato)
DISTRETTI (numDistr, stato:**STATI**)
PROD_IN_MAGAZZ (prodotto:**PRODOTTI**, magazzino:**MAGAZZINI**)
MARCHE (codMarca, prodottaIn:**CITTÀ**)
TIPI (tipo, gruppoMarketing:**GRUPPIMARK**,
categoria:**CATEGORIE**)
GRUPPIMARK (gruppoMarketing, responsabile)
CATEGORIE (categoria, reparto:**REPARTI**)
REPARTI (reparto, capoReparto)

76

Scelta dei fatti

I fatti sono concetti di interesse primario per il processo decisionale; tipicamente, corrispondono a eventi che accadono dinamicamente nel mondo aziendale

- Sullo schema E/R un fatto può corrispondere o a un' entità F o a un' associazione n-aria R tra le entità E1, E2..., En



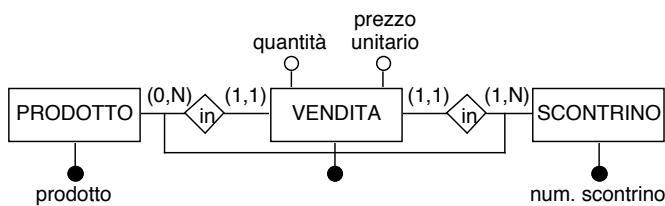
- Sullo schema relazionale un fatto corrisponde a una relazione F

77

Scelta dei fatti

Le entità o relazioni che rappresentano archivi frequentemente modificati (come VENDITA) sono buoni candidati per definire fatti; quelli che rappresentano archivi quasi-statici (come NEGOZIO e CITTÀ) no

- Nell' esempio delle vendite si sceglie come fatto l' associazione VENDITA, corrispondente alla relazione VENDITE.



- Ogni fatto identificato diviene la radice di un nuovo schema

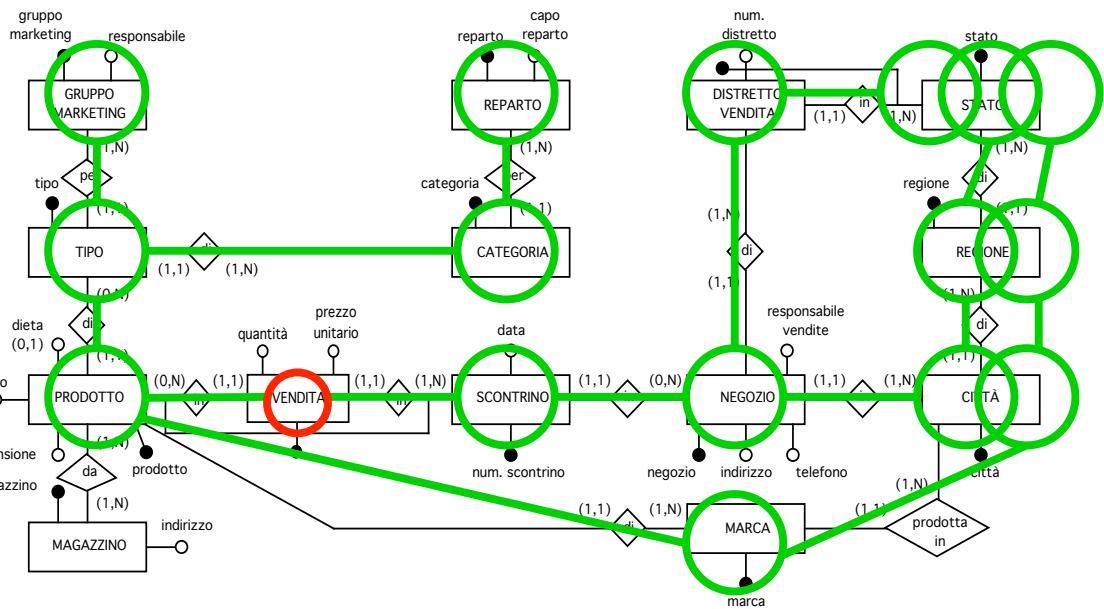
78

Costruzione dell' albero degli attributi

- L' albero degli attributi è un albero in cui:
 - ✓ ogni vertice corrisponde a un attributo - semplice o composto - dello schema sorgente;
 - ✓ la radice corrisponde all' identificatore (chiave primaria) di F;
 - ✓ per ogni vertice v, l' attributo corrispondente determina funzionalmente tutti gli attributi corrispondenti ai discendenti di v
- L' albero degli attributi corrispondente a F può essere costruito in modo automatico applicando una procedura che naviga ricorsivamente le dipendenze funzionali espresse, nello schema sorgente, dagli identificatori e dalle associazioni a-uno

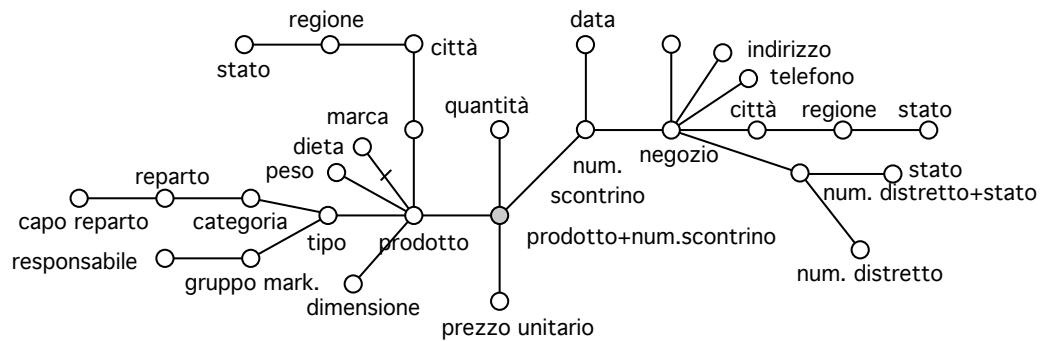
79

L' esempio delle vendite



80

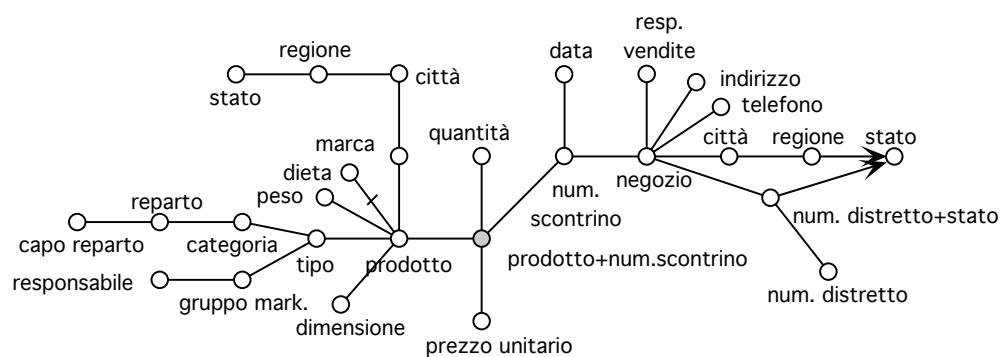
L' esempio delle vendite



81

Problemi

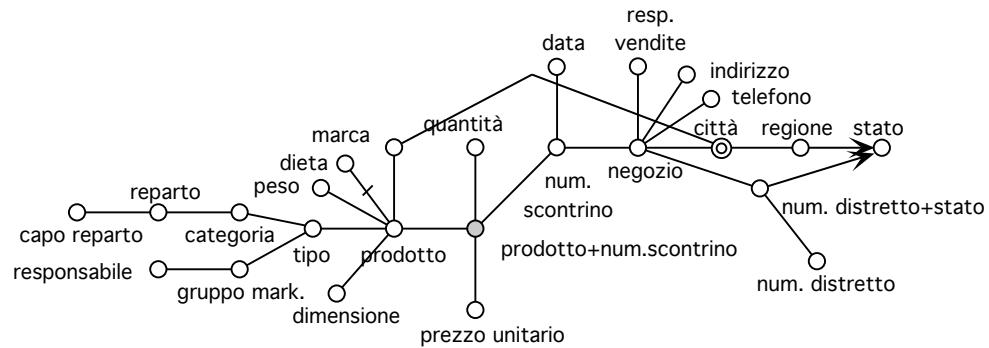
- Stessa entità raggiunta due volte
✓ convergenza



82

Problemi

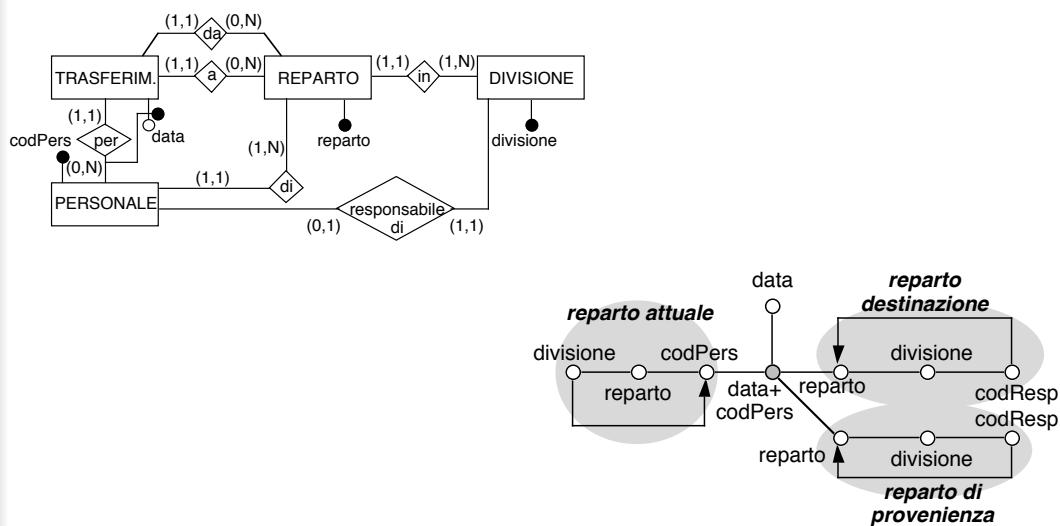
- Stessa entità raggiunta due volte
 - ✓ gerarchia condivisa



83

Problemi

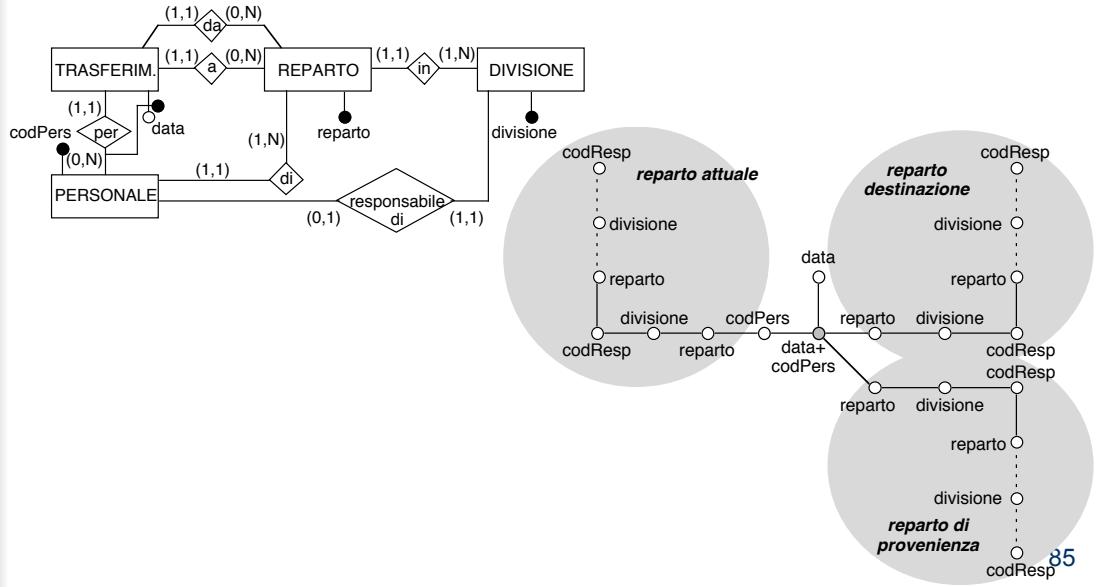
- Cicli di associazioni multi-a-uno
 - ✓ uso di gerarchie ricorsive



84

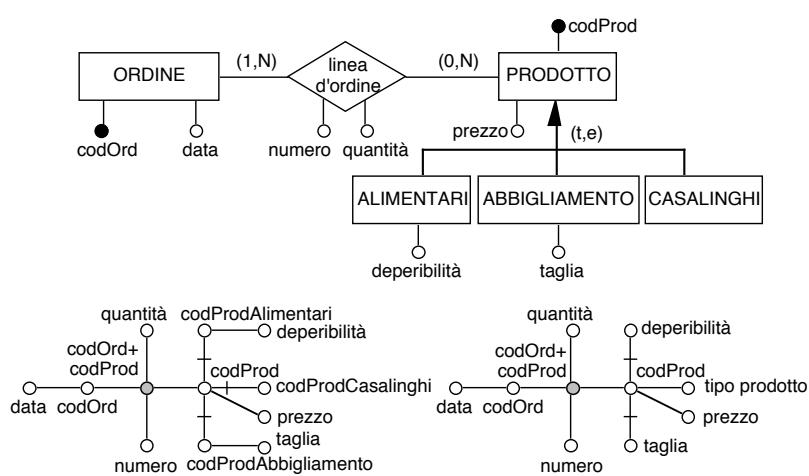
Problemi

- Cicli di associazioni multi-a-uno
 - ✓ “taglio” della gerarchia



Problemi

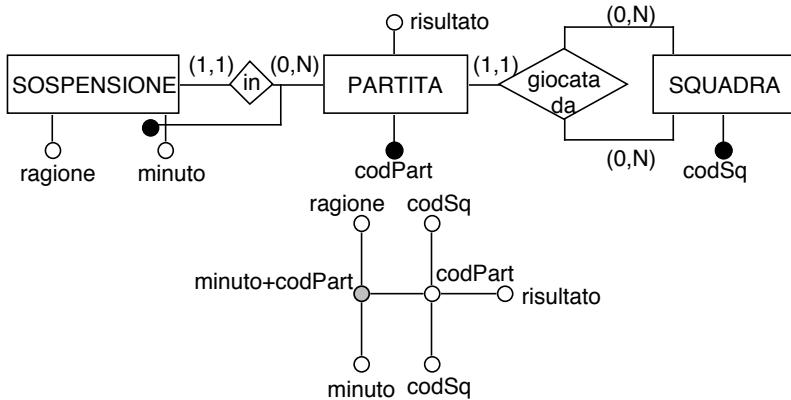
- Gerarchie di specializzazione
 - ✓ equivalenza con associazioni uno-a-uno opzionali



Problemi

■ Associazioni n-arie

✓ percorribili solo le “false n-arie”

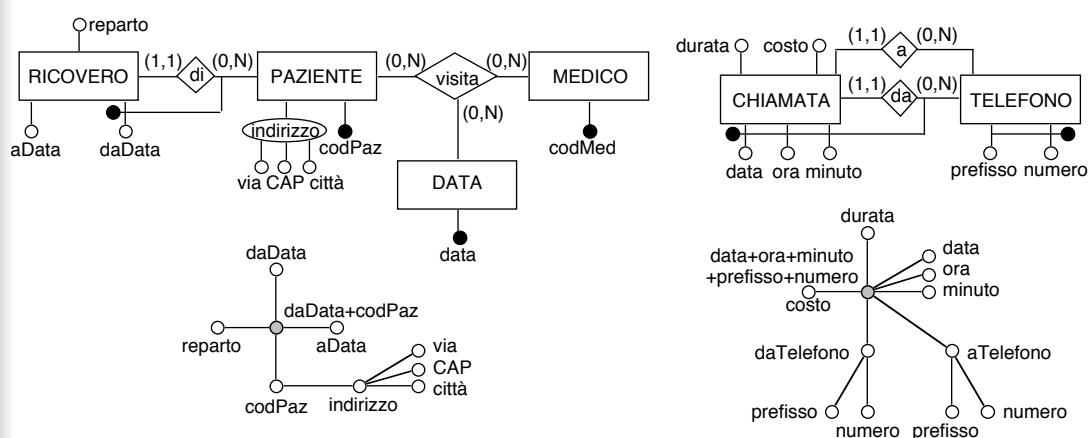


87

Problemi

■ Attributi composti

✓ generano due livelli nell’ albero



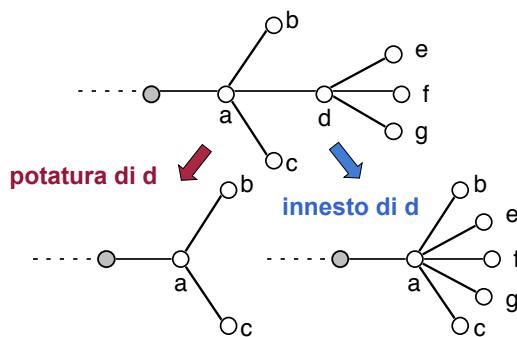
88

Editing dell' albero

- In genere non tutti gli attributi dell' albero sono d' interesse per il data mart; quindi, l' albero può essere manipolato per eliminare i livelli di dettaglio non necessari
 - ✓ La **potatura** di un vertice v si effettua eliminando l' intero sottoalbero con radice in v
 - Gli attributi eliminati non verranno inclusi nello schema di fatto, quindi non potranno essere usati per aggregare i dati
 - ✓ L' **innesto** viene utilizzato quando, sebbene un vertice esprima un' informazione non interessante, è necessario mantenere nell' albero i suoi discendenti
 - L' innesto del vertice v , con padre v' , viene effettuato collegando tutti i figli di v direttamente a v' ed eliminando v ; come risultato verrà perduto il livello di aggregazione corrispondente all' attributo v ma non i livelli corrispondenti ai suoi discendenti

89

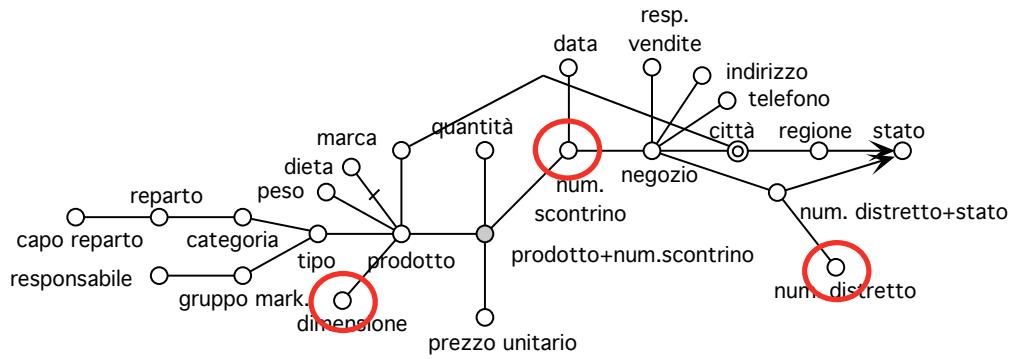
Editing dell' albero



- Quando un vertice opzionale viene innestato, tutti i suoi figli ereditano il trattino di opzionalità
 - ✓ Nel caso di potatura o innesto di un vertice opzionale v con padre v' è possibile aggiungere a v' un nuovo figlio b corrispondente a un attributo booleano che esprima l' opzionalità
- Potare o innestare un figlio della radice che corrisponde, sullo schema sorgente, a un attributo incluso nell' identificatore dell' entità scelta come fatto significa rendere più grossolana la granularità del fatto
 - ✓ Se il vertice innestato ha più di un figlio, si può avere un aumento del numero di dimensioni nello schema di fatto

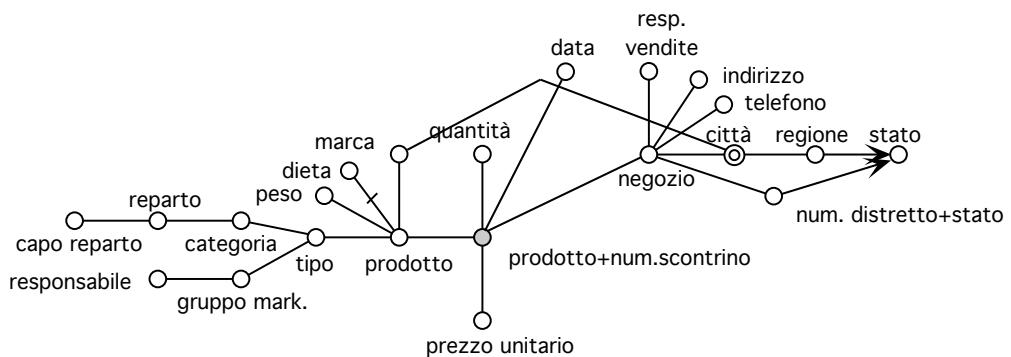
90

L' esempio delle vendite



91

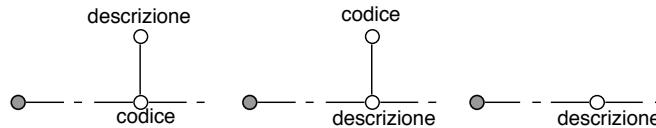
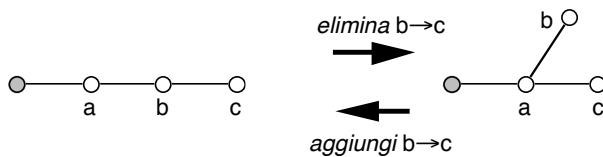
L' esempio delle vendite



92

Editing dell' albero

- Nella pratica possono rendersi necessarie ulteriori manipolazioni sull' albero degli attributi
 - ✓ Può essere necessario modificarne radicalmente la struttura sostituendo il padre di un certo nodo: ciò corrisponde ad aggiungere o eliminare una dipendenza funzionale
 - ✓ In presenza di un' associazione uno-a-uno sono consigliabili due soluzioni:
 - quando il vertice v determinato dall' associazione uno-a-uno ha dei discendenti di interesse lo si può eliminare dall' albero tramite innesto;
 - quando v non ha discendenti di interesse lo si può rappresentare come attributo descrittivo.
 - in alcuni casi può convenire *invertire* i due nodi coinvolti



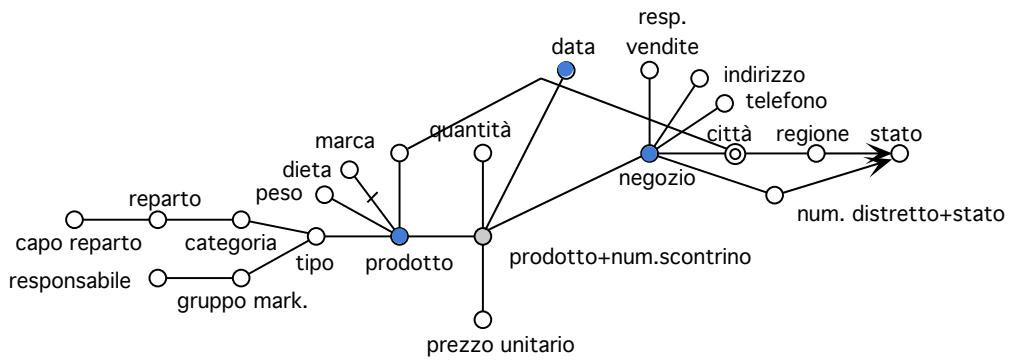
93

Scelta delle dimensioni

- Le dimensioni devono essere scelte nell' albero degli attributi tra i vertici figli della radice; possono corrispondere ad attributi discreti o a intervalli di valori di attributi discreti o continui
- La loro scelta è cruciale per il progetto poiché definisce la *granularità* degli eventi primari

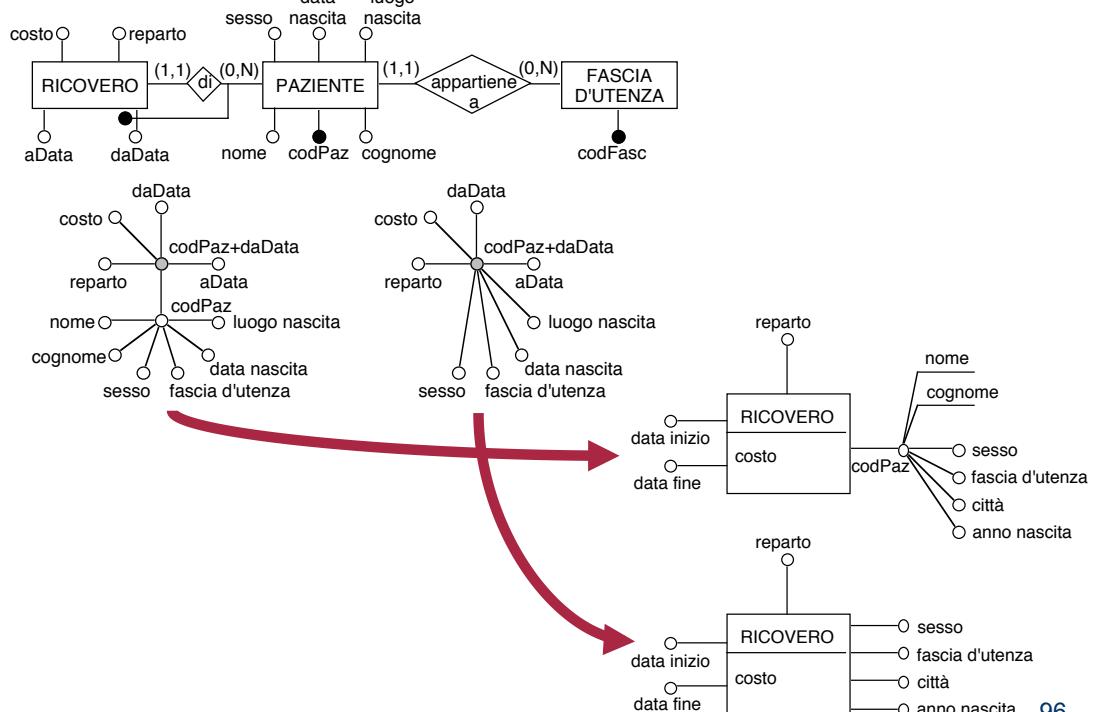
94

L' esempio delle vendite



95

L' esempio dei ricoveri



96

Il tempo

- Il tempo dovrebbe sempre essere una dimensione:
 - ✓ Se la sorgente è uno schema storico, il tempo è rappresentato esplicitamente come un attributo; se appare nell' albero degli attributi come figlio di un vertice diverso dalla radice, si può effettuare un innesto o eliminare una dipendenza funzionale al fine di farlo diventare un figlio diretto della radice e quindi una dimensione
 - ✓ Nelle sorgenti snapshot il tempo non viene rappresentato esplicitamente; in questo caso il tempo viene aggiunto "manualmente" allo schema di fatto
- In entrambi i casi, il significato che si dà alla dimensione tempo è quello di *tempo di validità*, inteso come l' istante in cui l' evento si è verificato nel mondo aziendale. Al tempo di transazione, ossia l' istante in cui l' evento è stato memorizzato nel database, non viene data tipicamente importanza nei DW, non essendo considerato rilevante per il supporto decisionale

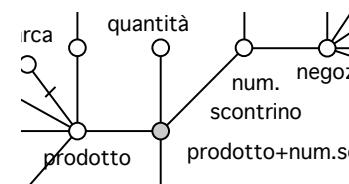
97

Scelta delle misure

- Se tra le dimensioni compaiono tutti gli attributi che costituiscono un identificatore dell'entità fatto, allora le misure corrispondono ad attributi numerici figli della radice dell'albero

SCONTRINI (numScontrino, data, negozio:NEGOZI)
VENDITE (prodotto:PRODOTTI, numScontrino:SCONTRINI, quantità)

prodotto	numScontrino	quantità	data	negozi
vite	S1	10	2/2/2019	DiTutto
bullone	S1	5	2/2/2019	DiTutto
vite	S2	3	2/2/2019	DiTutto
dado	S2	8	2/2/2019	DiTutto
dado	S3	4	2/2/2019	DiTutto



	vite	bullone	dado
S1	10	5	---
S2	3	---	8
S3	---	---	4

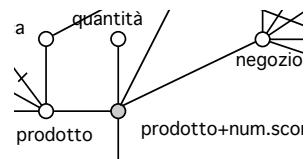
98

Scelta delle misure

- Altrimenti le misure si definiscono applicando, ad attributi numerici dell'albero, funzioni di aggregazione che operano su tutte le istanze di F corrispondenti a ciascun evento primario (somma/media/massimo/minimo di espressioni oppure conteggio del numero di istanze di F)
 - Qualora la granularità del fatto sia differente da quella dello schema sorgente, può essere utile definire più misure che aggregano lo stesso attributo tramite operatori diversi

SCONTRINI (numScontrino, data, negozio:NEGOZI)
VENDITE (prodotto:PRODOTTI, numScontrino:SCTRINI, quantità)

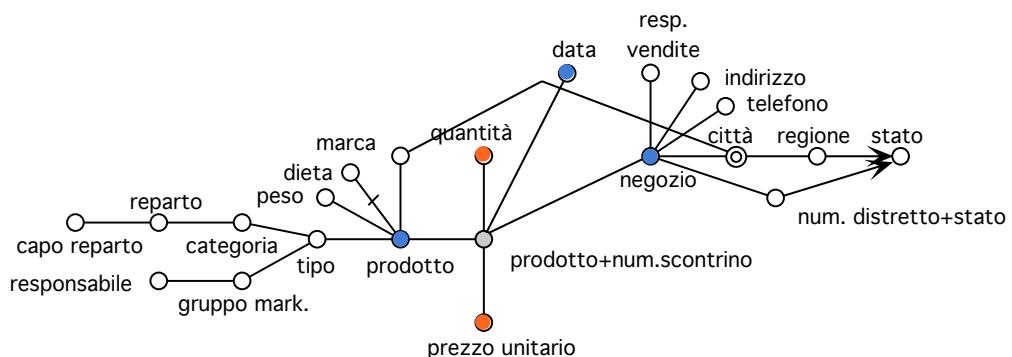
prodotto	numScontrino	quantità	data	negozi
vite	S1	10	2/2/2019	DiTutto
bullone	S1	5	2/2/2019	DiTutto
vite	S2	3	2/2/2019	DiTutto
dado	S2	8	2/2/2019	DiTutto
dado	S3	4	2/2/2019	DiTutto



DiTutto	vite	bullone	dado
2/2/2019	13	5	12

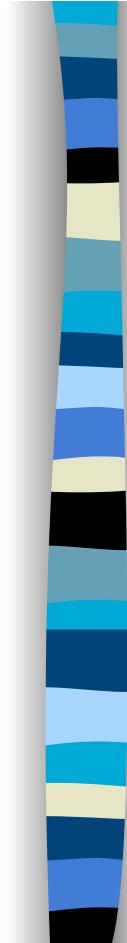
99

L' esempio delle vendite



GLOSSARIO

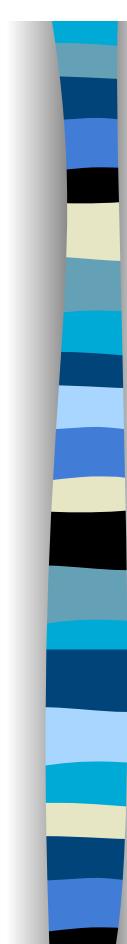
quantità venduta = SUM(VENDITA.quantità)
 incasso = SUM(VENDITA.quantità*VENDITA.prezzoUnitario)
 prezzo unitario = AVG(VENDITA.prezzoUnitario)
 num. clienti = COUNT(*)



Creazione dello schema di fatto

- L' albero degli attributi può ora essere tradotto in uno schema di fatto che include le dimensioni e misure definite
 - ✓ le gerarchie corrispondono ai sottoalberi dell' albero degli attributi con radice nelle diverse dimensioni
 - ✓ il nome del fatto corrisponde al nome dell' entità scelta come fatto
 - ✓ è possibile potare e innestare l' albero per eliminare dettagli inutili
 - ✓ è possibile aggiungere attributi dimensionali definendo opportuni intervalli per attributi numerici (per es. sulla dimensione tempo)
 - ✓ gli attributi che non verranno usati per l' aggregazione possono essere contrassegnati come descrittivi; tra questi compariranno in genere anche gli attributi determinati da associazioni uno-a-uno e privi di discendenti
 - ✓ per quanto riguarda eventuali attributi alfanumerici figli della radice ma non prescelti né come dimensioni né come misure:
 - se la granularità degli eventi primari coincide con quella dell' entità F, essi possono essere rappresentati come attributi descrittivi associati direttamente al fatto, di cui descriveranno ciascuna occorrenza
 - se invece le due granularità sono differenti, essi devono necessariamente essere potati

101

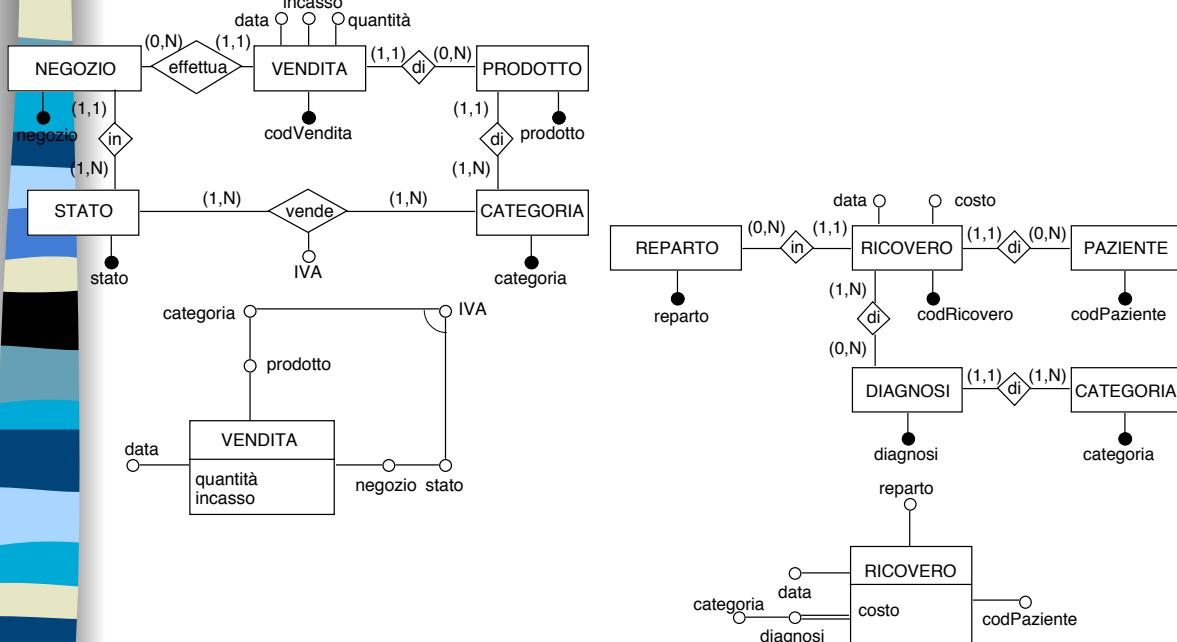


Creazione dello schema di fatto

- Eventuali attributi cross-dimensional e archi multipli possono essere evidenziati in questa fase
 - ✓ Identificare queste tipologie di attributi a partire dallo schema sorgente è complesso, poiché richiede di navigare anche le associazioni a-molti, per cui si preferisce definirli a partire dai requisiti utente per rappresentarli solo successivamente sullo schema di fatto
 - Un attributo cross-dimensional corrisponde in genere a un attributo posto su un' associazione molti-a-molti R dello schema E/R; i suoi padri nello schema di fatto corrisponderanno allora agli identificatori delle entità coinvolte in R
 - Un arco multiplo corrisponde a un' associazione a-molti R da un' entità E a un' entità G; nello schema di fatto, esso potrà allora connettere l' identificatore di E o il fatto con un attributo di R o di G

102

Creazione dello schema di fatto



103

Creazione dello schema di fatto

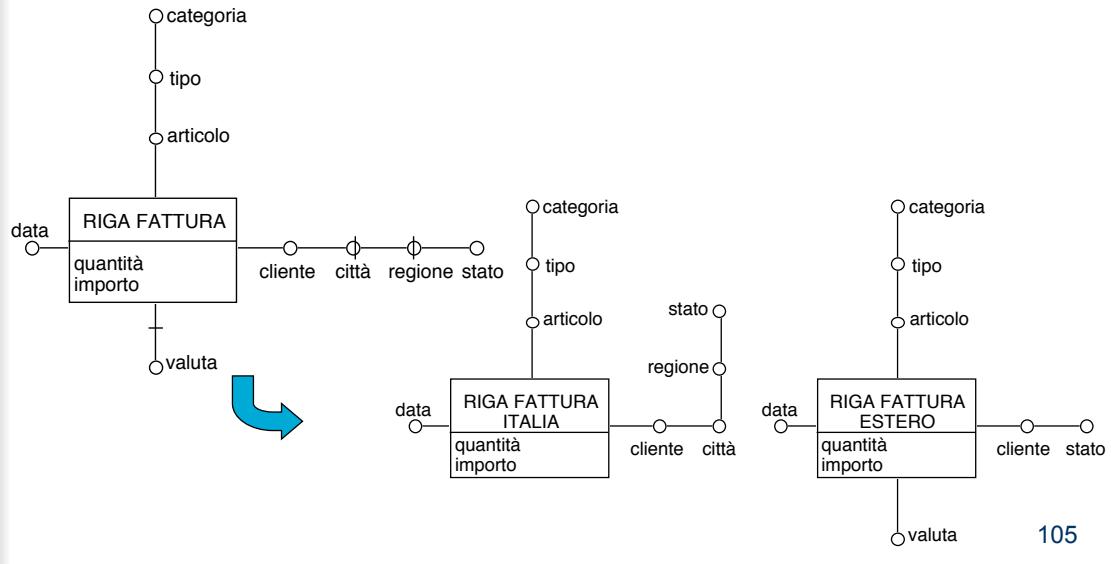
- In questa fase devono anche essere identificate le eventuali non-additività e non-aggregabilità presenti nello schema, considerando tutte le accoppiate dimensione-misura
- Dato uno schema di fatto n-dimensionale, per la dimensione d_i e la misura m_j , la domanda da porsi sarà:

“Siano $\{val_1, \dots, val_k\}$ i valori assunti dalla misura m_j nei k eventi primari corrispondenti a k differenti valori presi dal dominio della dimensione d_i e da un valore prefissato di ciascuna delle altre $n-1$ dimensioni. Volendo caratterizzare complessivamente i k eventi con un unico valore di m_j , quali operatori di aggregazione ha senso utilizzare?”

104

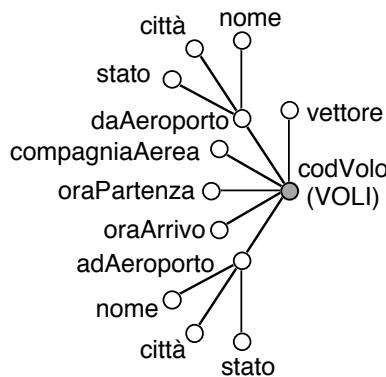
Frammentazione dello schema di fatto

- In alcuni casi, il progettista può valutare la possibilità di frammentare uno schema di fatto in due o più schemi con l'obiettivo di regolarizzare le gerarchie



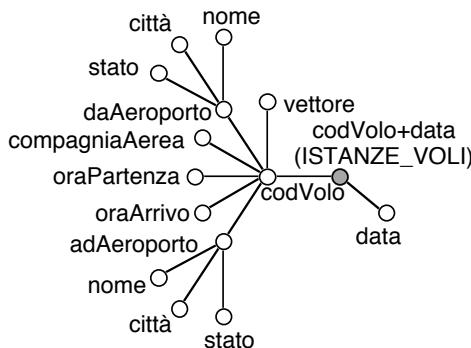
L'esempio dei voli

VOLI(codVolo, compagniaAerea, daAeroporto:AEROPORTI,
 adAeroporto:AEROPORTI, oraPartenza, oraArrivo, vettore)
 ISTANZE_VOLI(codVolo:VOLI, data)
 AEROPORTI(sigla, nome, città, stato)
 BIGLIETTI(numero, (codVolo, data):ISTANZE_VOLI, numPosto, tariffa,
 nomeCliente, cognomeCliente, sessoCliente)
 CHECK-IN(numero:BIGLIETTI, oraCheckIn, numeroColli)



L' esempio dei voli

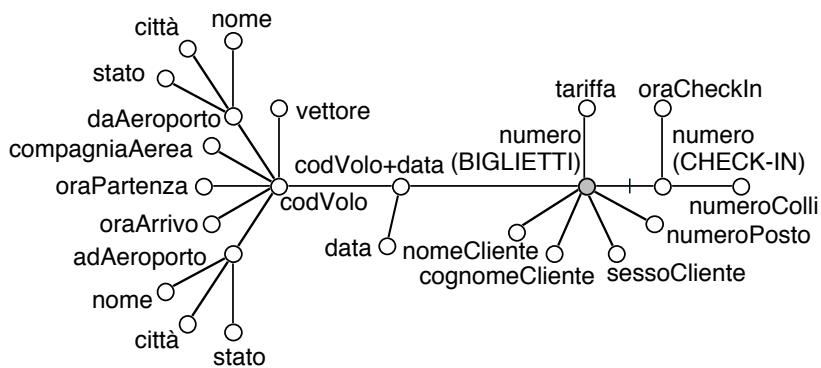
VOLI(codVolo, compagniaAerea, daAeroporto:AEROPORTI,
 adAeroporto:AEROPORTI, oraPartenza, oraArrivo, vettore)
 ISTANZE_VOLI(codVolo:VOLI, data)
 AEROPORTI(sigla, nome, città, stato)
 BIGLIETTI(numero, (codVolo, data):ISTANZE_VOLI, numPosto, tariffa,
 nomeCliente, cognomeCliente, sessoCliente)
 CHECK-IN(numero:BIGLIETTI, oraCheckIn, numeroColli)



107

L' esempio dei voli

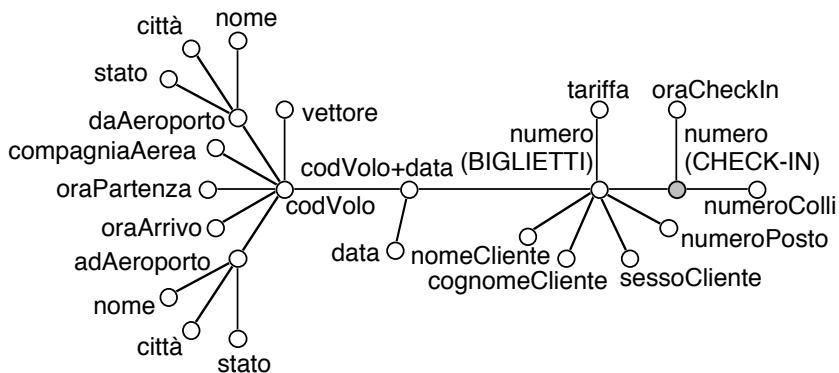
VOLI(codVolo, compagniaAerea, daAeroporto:AEROPORTI,
 adAeroporto:AEROPORTI, oraPartenza, oraArrivo, vettore)
 ISTANZE_VOLI(codVolo:VOLI, data)
 AEROPORTI(sigla, nome, città, stato)
 BIGLIETTI(numero, (codVolo, data):ISTANZE_VOLI, numPosto, tariffa,
 nomeCliente, cognomeCliente, sessoCliente)
 CHECK-IN(numero:BIGLIETTI, oraCheckIn, numeroColli)



108

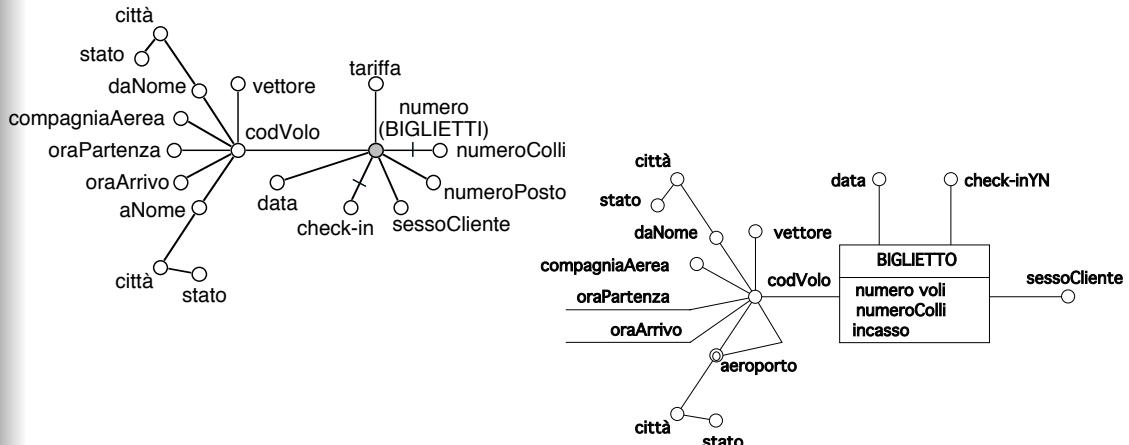
L' esempio dei voli

VOLI(codVolo, compagniaAerea, daAeroporto:AEROPORTI,
 adAeroporto:AEROPORTI, oraPartenza, oraArrivo, vettore)
 ISTANZE_VOLI(codVolo:VOLI, data)
 AEROPORTI(sigla, nome, città, stato)
 BIGLIETTI(numero, (codVolo, data):ISTANZE_VOLI, numPosto, tariffa,
 nomeCliente, cognomeCliente, sessoCliente)
 CHECK-IN(numero:BIGLIETTI, oraCheckIn, numeroColli)



109

L' esempio dei voli

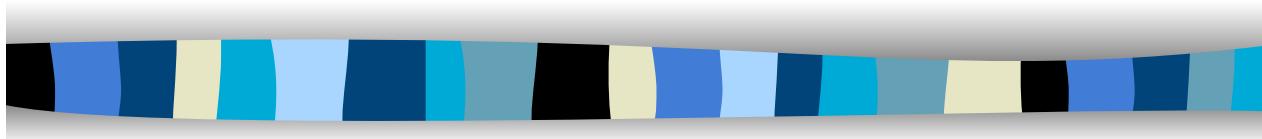


```

numero voli = SELECT COUNT(*)
              FROM BIGLIETTI B, ISTANZE_VOLI I, CHECK-IN C
              WHERE B.codVolo = I.codVolo AND B.data = I.data AND B.numero = C.numero
              GROUP BY B.sessoCliente, I.data, B.codVolo
numero colli = SELECT SUM(C.numeroColli)
              FROM BIGLIETTI B, ISTANZE_VOLI I, CHECK-IN C
              WHERE B.codVolo = I.codVolo AND B.data = I.data AND B.numero = C.numero
              GROUP BY B.sessoCliente, I.data, B.codVolo
incasso = SELECT SUM(B.tariffa)
           FROM BIGLIETTI B, ISTANZE_VOLI I, CHECK-IN C
           WHERE B.codVolo = I.codVolo AND B.data = I.data AND B.numero = C.numero
           GROUP BY B.sessoCliente, I.data, B.codVolo
  
```

110

Carico di lavoro e volume dati

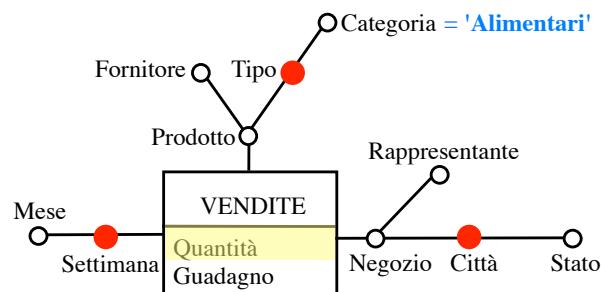


Il carico di lavoro

- Il carico di lavoro di un sistema OLAP è per sua natura estemporaneo
- È necessario identificare in fase di progettazione un carico di lavoro di riferimento
 - ✓ Reportistica standard
 - ✓ Colloqui con gli utenti
- Le interrogazioni OLAP sono facilmente caratterizzabili
 - ✓ Group-by set
 - ✓ Misure richieste
 - ✓ Clausole di selezione



Il carico di lavoro



*Totale della quantità venduta per i diversi tipi di prodotto, in ogni settimana e città
ma solo per i prodotti alimentari*

113

Dinamicità del carico di lavoro

- Il carico di lavoro preliminare non è di per sé sufficiente a ottimizzare le prestazioni del sistema
 - ✓ L'interesse degli utenti cambia nel tempo
 - ✓ Il numero di interrogazioni aumenta al crescere della confidenza degli utenti con il sistema
- Per ottimizzare la struttura logica del data mart è necessaria una fase di tuning attuabile solo dopo che il sistema è stato messo in funzione
- Il carico di lavoro reale può essere desunto dal log delle interrogazioni sottoposte al sistema

114

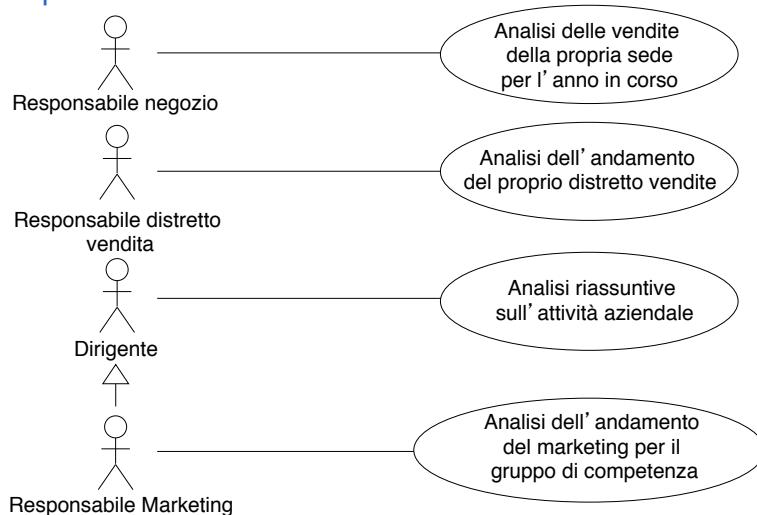
Il carico di lavoro e gli utenti

- Progettare un DW significa anche determinare le modalità di accesso ai dati definendo quali utenti possano accedere a quali dati e con quale modalità
- Per fare ciò è necessario classificare gli utenti finali e le tipologie di interrogazioni che essi prevedono di rivolgere al data mart, al fine di definire una griglia di autorizzazioni che verrà utilizzata dagli implementatori del front-end per configurare opportunamente il sistema

115

Il carico di lavoro e gli utenti

1. Classificare gli utenti in gruppi omogenei (*profilazione*)
 - ✓ il criterio principale da utilizzare è la funzione aziendale svolta, che determina normalmente l'insieme delle informazioni a cui uno specifico utente ha accesso
 - ✓ tra le figure individuate possono essere anche precise relazioni di specializzazione

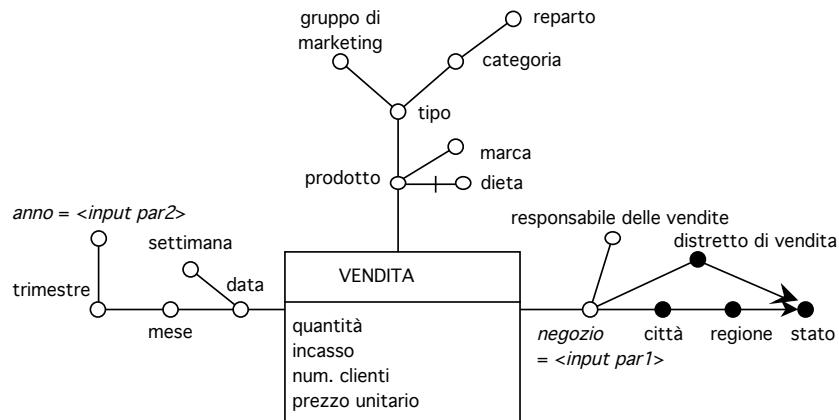


116

Il carico di lavoro e gli utenti

2. Descriverne i permessi di accesso rispetto agli schemi di fatto

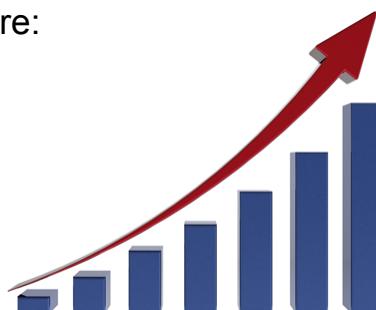
- ✓ Quali misure e quali attributi descrittivi sono visualizzabili
- ✓ Quali gerarchie e quali attributi dimensionali sono navigabili
- ✓ Quali restrizioni a livello di istanze è necessario applicare sui dati



117

Il volume dati

- Consiste nelle informazioni necessarie a determinare/stimare la dimensione del data mart
 - ✓ Numero di valori distinti degli attributi nelle gerarchie
 - ✓ Lunghezza degli attributi
 - ✓ Numero di eventi di ogni fatto
- Deve essere calcolato considerando la quantità di dati necessari a coprire l' intervallo temporale deciso per il data mart
- È utilizzato sia durante la progettazione logica sia durante la progettazione fisica per determinare:
 - ✓ la dimensione delle tabelle
 - ✓ la dimensione degli indici
 - ✓ i costi di accesso



118