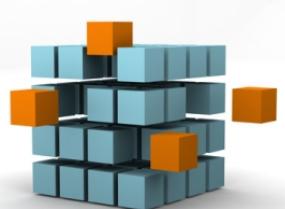


Business Intelligence

Prof. Stefano Rizzi

Obiettivi del corso

- Obiettivo del corso è presentare allo studente una trattazione approfondita dei sistemi di business intelligence, con particolare accento sulle tematiche legate al data warehousing
- Prerequisito per il corso è una sufficiente conoscenza delle basi di dati relazionali
- Parte integrante del corso sono le esercitazioni di laboratorio, basate su strumenti tecnologicamente avanzati e ampiamente diffusi in ambito aziendale
- Al termine del corso, lo studente sarà in grado di comprendere i meccanismi alla base delle piattaforme di business intelligence, nonché di progettare e gestire data warehouse aziendali



Programma

■ Business intelligence:

- ✓ Il ruolo della BI nel sistema informativo aziendale
- ✓ La piramide della BI
- ✓ Data warehousing

■ Data warehousing:

- ⇒ Architetture
- ⇒ Tecniche di analisi dei dati
- ⇒ Il ciclo di sviluppo
 - Analisi delle sorgenti dati
 - Analisi dei requisiti
 - Progettazione concettuale
 - Carico di lavoro e volume dati
 - Progettazione logica
 - Progettazione dell'alimentazione
 - Progettazione fisica
- ⇒ La documentazione di progetto

Orario

■ Lezione

- ✓ 28 ore

■ Esercitazione

- ✓ 12 ore

■ Laboratorio

- ✓ 10 ore



Esercitazioni di laboratorio

- ETL - OLAP - progettazione
(Indyco+TableauPrep+TableauDesktop)



Modalità d' esame

Prova pratica (6/32)

1. Prova al PC con le tecnologie viste in laboratorio

Prova scritta (26/32)

1. Progettazione concettuale/logica
2. Quesiti sui contenuti teorici del corso

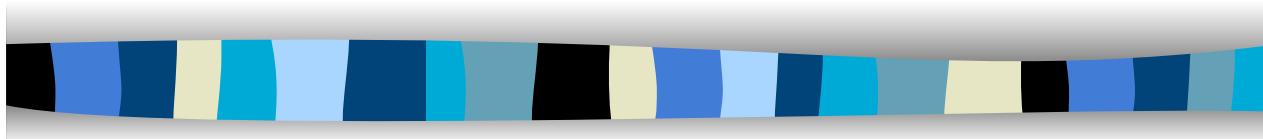




Testi

- 👉 Dispense a cura del docente
- ❑ M. Berry, G. Linoff. *Data mining techniques for marketing, sales, and customer support*. John Wiley & Sons, 1997
- ❑ B. Devlin. *Data warehouse: from architecture to implementation*. Addison-Wesley Longman, 1997
- 👉 M. Golfarelli, S. Rizzi. *Data warehouse: Teoria e pratica della progettazione - Seconda edizione*. McGraw-Hill, 2006
- ❑ W.H. Inmon. *Building the data warehouse*. John Wiley & Sons, 1996
- ❑ M. Jarke, M. Lenzerini, Y. Vassiliou, P. Vassiliadis. *Fundamentals of data warehouse*. Springer, 2000
- ❑ R. Kimball, L. Reeves, M. Ross, W. Thornthwaite. *The data warehouse lifecycle toolkit*. John Wiley & Sons, 1998
- ❑ I. Witten, E. Frank. *Data mining*. Morgan Kaufmann Publishers, 2000

La Business Intelligence



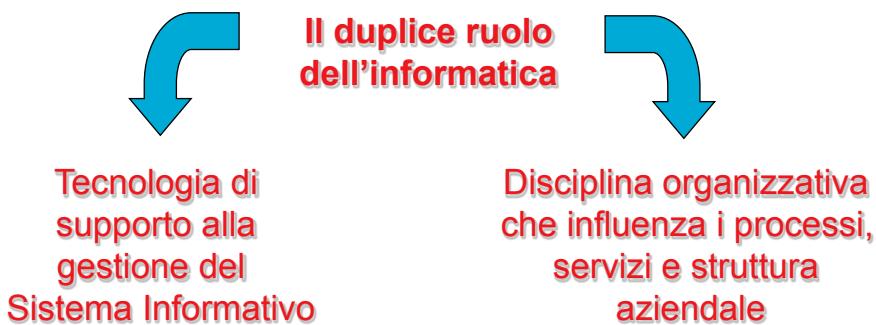
L'informatica in azienda

- La funzione svolta dalle basi di dati in ambito aziendale è stata fino a qualche anno fa solo quella di memorizzare **dati operazionali**, ossia dati generati da operazioni svolte all'interno dei processi gestionali
- L'informatica era vista come una **scienza di supporto** che permetteva di rendere più rapide ed economiche le operazioni di gestione delle informazioni ma che non creava di per sé ricchezza



L'evoluzione dei sistemi informativi

- Il ruolo dei Sistemi Informatici è radicalmente cambiato dai primi anni '70 a oggi. I sistemi informatici si sono trasformati da semplici strumenti per migliorare l'efficienza dei processi a elementi centrali dell'organizzazione aziendale in grado di rivoluzionare la struttura dei processi aziendali



3

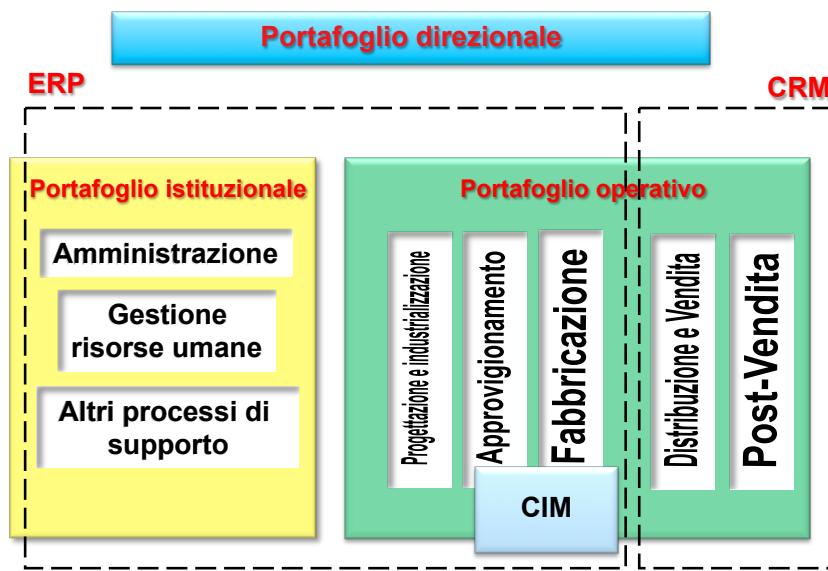
Il ruolo dell'informatica nel processo decisionale

- L'aumento esponenziale del volume dei dati operazionali ha reso il calcolatore l'unico supporto adatto al processo decisionale
- L'utilizzo massiccio di tecniche di analisi dei dati aziendali ha reso il sistema informativo un elemento strategico per la realizzazione del business
- Il ruolo dell'informatica è passato da passivo strumento per la registrazione delle operazioni a fattore decisivo per la individuazione di elementi critici dell'organizzazione e di potenziali aree di business



4

Il portafoglio applicativo



5

Il portafoglio direzionale

- E' l'insieme delle applicazioni utilizzate dai manager aziendali per:
 - ✓ Analizzare lo stato dell'azienda
 - ✓ Prendere decisioni rapide
 - ✓ Prendere le decisioni migliori
- Si parla anche di *piattaforma per la business intelligence*, ossia...



6

Business intelligence



■ Una definizione:

- ✓ La business intelligence è un insieme di strumenti e procedure che consentono a un'azienda di trasformare i propri dati di business in informazioni e conoscenza utili al processo decisionale
- ✓ Le informazioni così ottenute sono utilizzate dai decisori aziendali per definire e supportare le strategie di business, così da **operare decisioni consapevoli e informate** con l'obiettivo di trarre vantaggi competitivi, migliorare le prestazioni operative e la profittabilità e, più in generale, **creare valore** per l'azienda

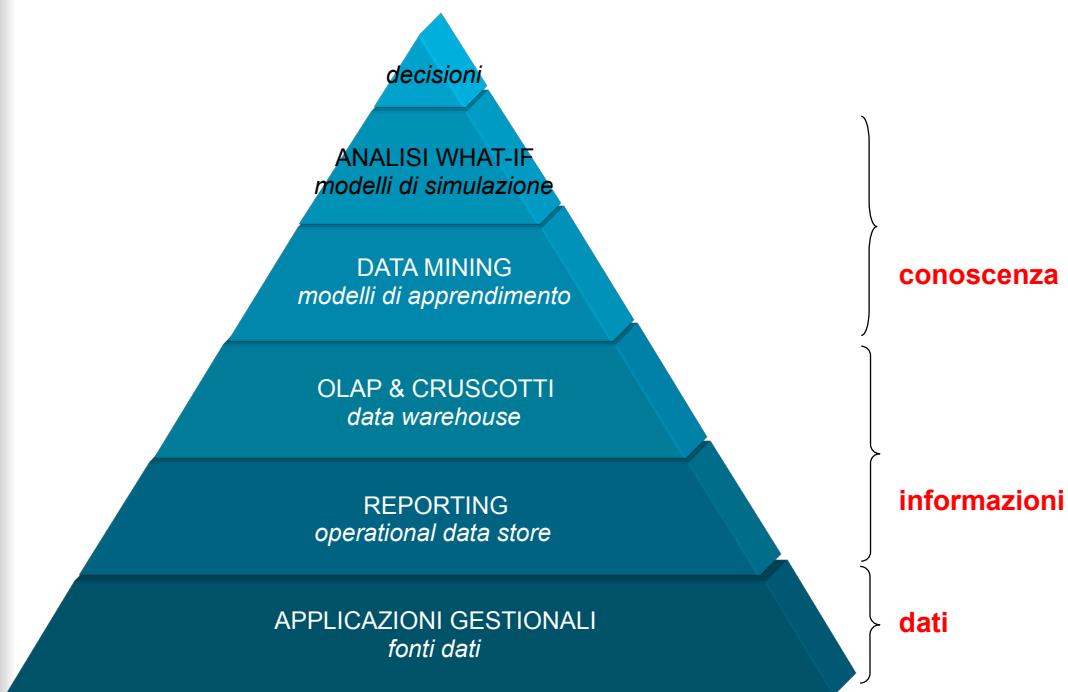
7

Business intelligence

- Si parla di *piattaforma* di BI poiché per consentire ai manager analisi potenti e flessibili è necessario definire un'apposita infrastruttura hardware e software di supporto composta da:
 - ✓ Hardware dedicato
 - ✓ Infrastrutture di rete
 - ✓ DBMS
 - ✓ Software di back-end
 - ✓ Software di front-end
- Il ruolo chiave di una piattaforma di business intelligence è la trasformazione dei *dati* aziendali in *informazioni* fruibili a diversi livelli di dettaglio e, quindi, in *conoscenza*

8

La piramide della BI



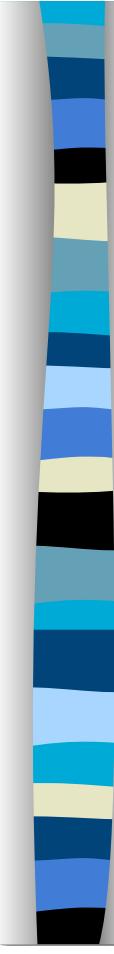
9

Ciclo decisionale in BI

- **Analisi**
 - ✓ identificare e formulare il problema
 - ✓ ottenere dai dati le informazioni rilevanti
- **Comprendione**
 - ✓ comprendere il problema
 - ✓ trasformare le informazioni in conoscenza
- **Decisione**
 - ✓ tradurre la conoscenza in decisioni e quindi in azioni
- **Misura**
 - ✓ misurare le prestazioni conseguenti alle azioni intraprese

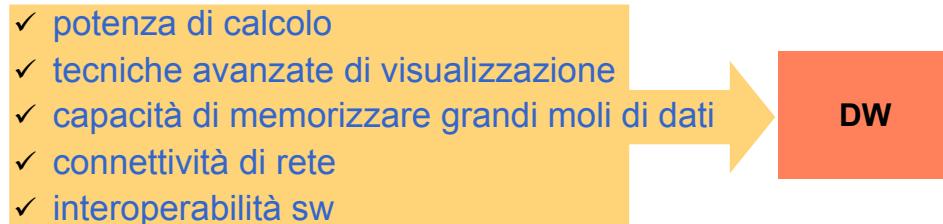


10



Fattori abilitanti per la BI

- **Tecnologie**

- **Tecnologie**
 - ✓ potenza di calcolo
 - ✓ tecniche avanzate di visualizzazione
 - ✓ capacità di memorizzare grandi moli di dati
 - ✓ connettività di rete
 - ✓ interoperabilità sw
- 

DW

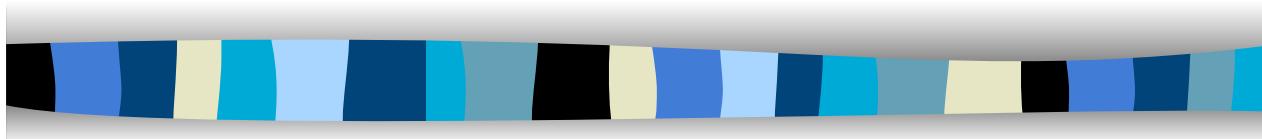
- **Metodologie analitiche**

- **Metodologie analitiche**
- ✓ modelli matematici espressivi, precisi e flessibili
- ✓ tecniche di apprendimento induttivo e di ottimizzazione

- **Risorse umane**

- **Risorse umane**
- ✓ cultura aziendale
- ✓ creatività
- ✓ agilità mentale
- ✓ disponibilità al cambiamento

Il Data Warehousing



Dai dati alle informazioni

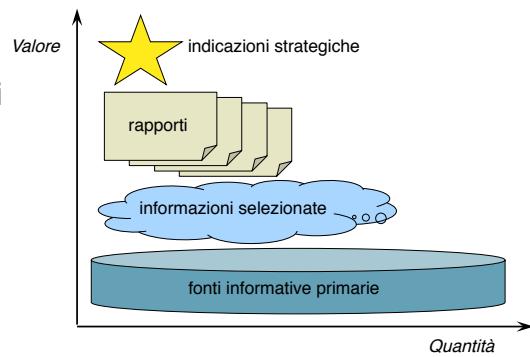
- L'informazione è un bene a valore crescente, necessario per pianificare e controllare le attività aziendali con efficacia
- Essa costituisce la materia prima che viene trasformata dai sistemi informativi, come i semilavorati vengono trasformati dai sistemi di produzione

~~dati = informazione~~

- Spesso la disponibilità di troppi dati rende arduo, se non impossibile, estrarre le informazioni veramente importanti

Dai dati alle informazioni

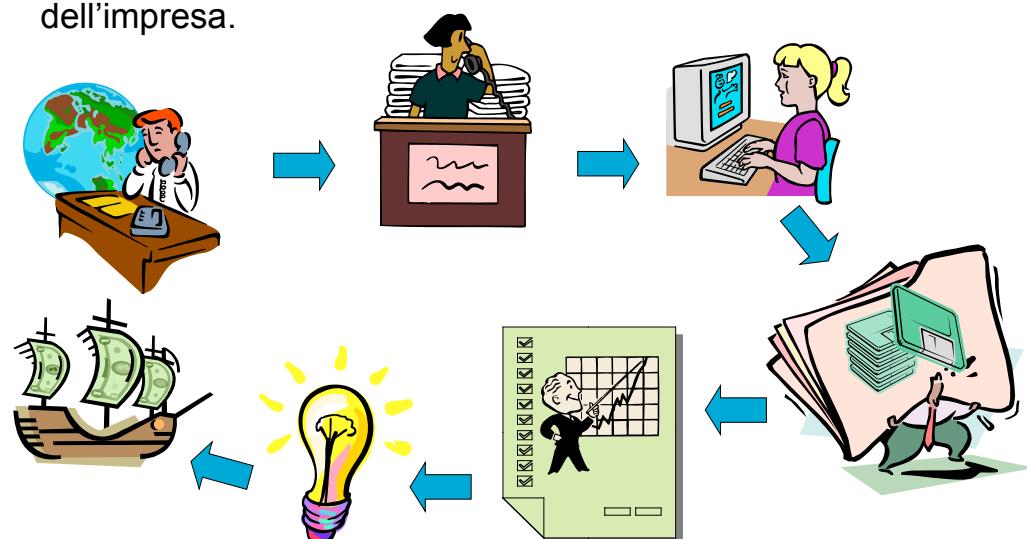
- Per ogni azienda è fondamentale poter disporre in maniera rapida e completa delle informazioni necessarie al processo decisionale: le indicazioni strategiche sono estrapolate principalmente dalla mole dei dati operazionali contenuti nei database aziendali, attraverso un procedimento di selezione e sintesi progressiva



3

Uno scenario tipico...

- .. è quello di una grande azienda, con numerose filiali, i cui dirigenti desiderano quantificare e valutare il contributo dato da ciascuna di esse al rendimento commerciale globale dell'impresa.



4

Uno scenario tipico...

- .. è quello di una grande azienda, con numerose filiali, i cui dirigenti desiderano quantificare e valutare il contributo dato da ciascuna di esse al rendimento commerciale globale dell'impresa.



5

Le interrogazioni

- **OLTP (On-Line Transactional Processing):**
 - ✓ Le interrogazioni eseguono transazioni che leggono e scrivono un ridotto numero di record da diverse tabelle legate da semplici relazioni
 - ✓ Il nucleo sostanziale del carico di lavoro è “congelato” all’interno dei programmi applicativi
- **OLAP (On-Line Analytical Processing):**
 - ✓ Le interrogazioni effettuano un’analisi multidimensionale che richiede la scansione di un’enorme quantità di record per calcolare un insieme di dati numerici di sintesi che quantifichino le prestazioni dell’azienda
 - ✓ L’interattività è una caratteristica irrinunciabile delle sessioni di analisi e fa sì che il carico di lavoro effettivo vari continuamente nel tempo

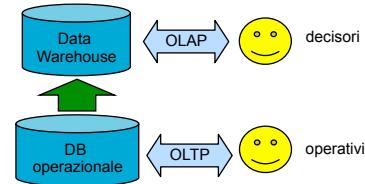
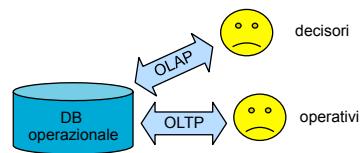
6

OLTP e OLAP

- Mescolare interrogazioni “analitiche” e “transazionali” di routine porta a inevitabili rallentamenti che rendono insoddisfatti gli utenti di entrambe le categorie



separare l’elaborazione di tipo analitico (OLAP) da quella legata alle transazioni (OLTP), costruendo un nuovo raccoglitore di informazioni che integri i dati provenienti da sorgenti di varia natura, li organizzi in una forma appropriata e li renda disponibili per scopi di analisi e valutazione finalizzate alla pianificazione e al processo decisionale



7

Alcune aree di utilità

- Commercio** (analisi delle vendite e dei reclami, controllo di spedizioni e inventari, cura del rapporto con i clienti)
- Manifattura** (controllo dei costi di produzione, supporto fornitori e ordini)
- Servizi finanziari** (analisi del rischio e delle carte di credito, rivelazione di frodi)
- Trasporti** (gestione parco mezzi)
- Telecomunicazioni** (analisi del flusso delle chiamate e del profilo dei clienti)
- Sanità** (analisi di ricoveri e dimissioni, contabilità per centri di costo)
-

8

Data Warehousing:

- Una collezione di metodi, tecnologie e strumenti di ausilio al *knowledge worker* (dirigente, amministratore, gestore, analista) per condurre analisi dei dati finalizzate all'attuazione di processi decisionali e al miglioramento del patrimonio informativo.

9

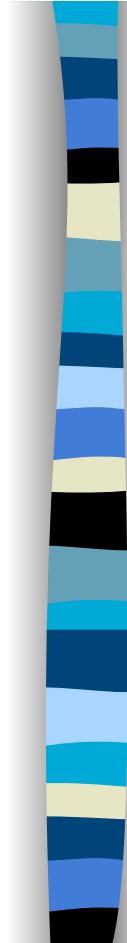
Le lamentele

- ☞ *abbiamo montagne di dati ma non possiamo accedervi!*
- ☞ *come è possibile che persone che svolgono lo stesso ruolo presentino risultati sostanzialmente diversi?*
- ☞ *vogliamo selezionare, raggruppare e manipolare i dati in ogni modo possibile!*
- ☞ *mostratemi solo ciò che è importante!*
- ☞ *tutti sanno che alcuni dati non sono corretti!*

R. Kimball, The Data Warehouse Toolkit



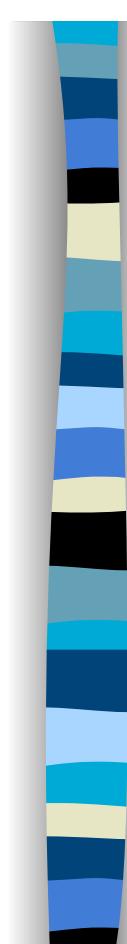
10



Caratteristiche del processo di warehousing

- **accessibilità** a utenti con conoscenze limitate di informatica e strutture dati;
- **integrazione dei dati** sulla base di un modello standard dell'impresa;
- **flessibilità di interrogazione** per trarre il massimo vantaggio dal patrimonio informativo esistente;
- **sintesi** per permettere analisi mirate ed efficaci;
- **rappresentazione multidimensionale** per offrire all'utente una visione intuitiva ed efficacemente manipolabile delle informazioni;
- **correttezza e completezza** dei dati integrati.

11



Il Data Warehouse

- Al centro del processo, il data warehouse è un contenitore di dati che si fa garante dei requisiti esposti.

➤ *Un Data Warehouse è una collezione di dati di supporto per il processo decisionale che presenta le seguenti caratteristiche:*

- ✓ *è orientata ai soggetti di interesse;*
- ✓ *è integrata e consistente;*
- ✓ *è rappresentativa dell'evoluzione temporale;*
- ✓ *non volatile.*

12

...orientato ai soggetti



13

...integrato e consistente

Il DW si appoggia a più fonti di dati eterogenee: dati estratti dall'ambiente di produzione, e quindi originariamente archiviati in basi di dati aziendali, o addirittura provenienti da sistemi informativi esterni all'azienda. Di tutti questi dati il DW restituisce una visione unificata.



14

...rappresentativo dell'evoluzione temporale

DB operazionali



Contenuto storico limitato,
spesso il tempo non è parte
delle chiavi,
i dati sono soggetti
ad aggiornamenti

DW

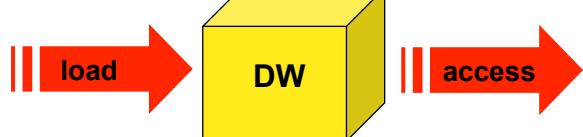
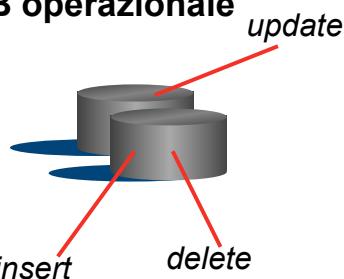


Ricco contenuto storico,
il tempo è parte delle chiavi,
una fotografia del dato
a un certo istante di tempo
non può essere aggiornata

15

...non volatile

DB operazionale



Grandi volumi di dati:
da qualche GB a diversi TB
in pochi anni

- ✓ in un DW non sono necessarie tecniche sofisticate di gestione delle transazioni come invece richiesto da un sistema operazionale
- ✓ problema chiave diventa il query-throughput

16

Riassumendo:

	Database operazionali	Data warehouse
utenti	migliaia	centinaia
carico di lavoro	transazioni predefinite	interrogazioni di analisi <i>ad hoc</i>
accesso	a centinaia di record, in lettura e scrittura	a milioni di record, per lo più in lettura
scopo	dipende dall'applicazione	supporto alle decisioni
dati	elementari, sia numerici sia alfanumerici	di sintesi, prevalentemente numerici
integrazione dei dati	per applicazione	per soggetto
qualità	in termini di integrità	in termini di consistenza
copertura temporale	solo dati correnti	dati correnti e storici
aggiornamenti	continui	periodici
modello	normalizzato	multidimensionale
ottimizzazione	per accessi OLTP su una frazione del database	per accessi OLAP su gran parte del database

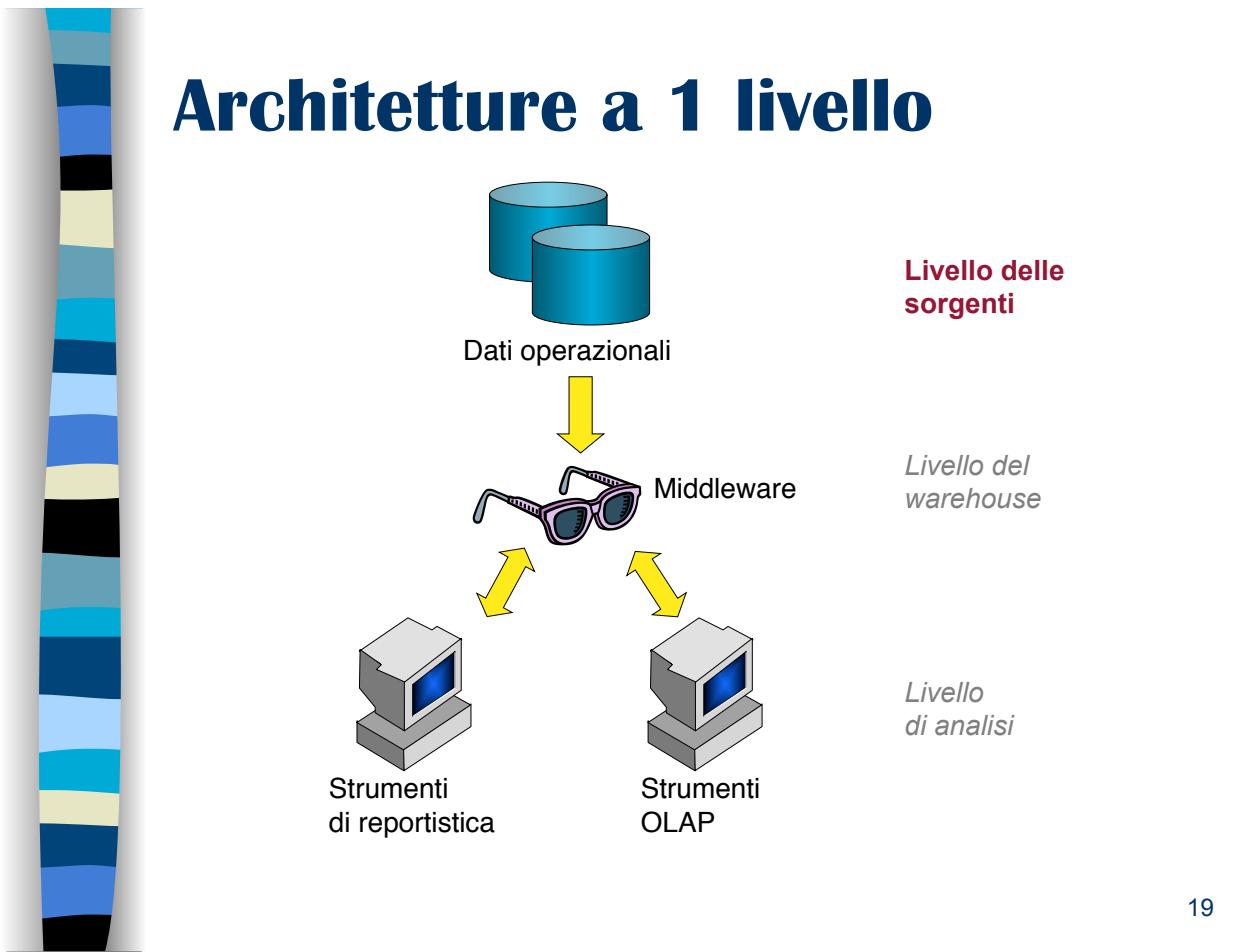
17

Architetture: requisiti

- ✓ **Separazione:** l'elaborazione analitica e quella transazionale devono essere mantenute il più possibile separate.
- ✓ **Scalabilità:** l'architettura hardware e software deve poter essere facilmente ridimensionata a fronte della crescita nel tempo dei volumi di dati da gestire ed elaborare e del numero di utenti da soddisfare.
- ✓ **Estendibilità:** deve essere possibile accogliere nuove applicazioni e tecnologie senza riprogettare integralmente il sistema.
- ✓ **Sicurezza:** il controllo sugli accessi è essenziale a causa della natura strategica dei dati memorizzati.
- ✓ **Amministrabilità:** la complessità dell'attività di amministrazione non deve risultare eccessiva.

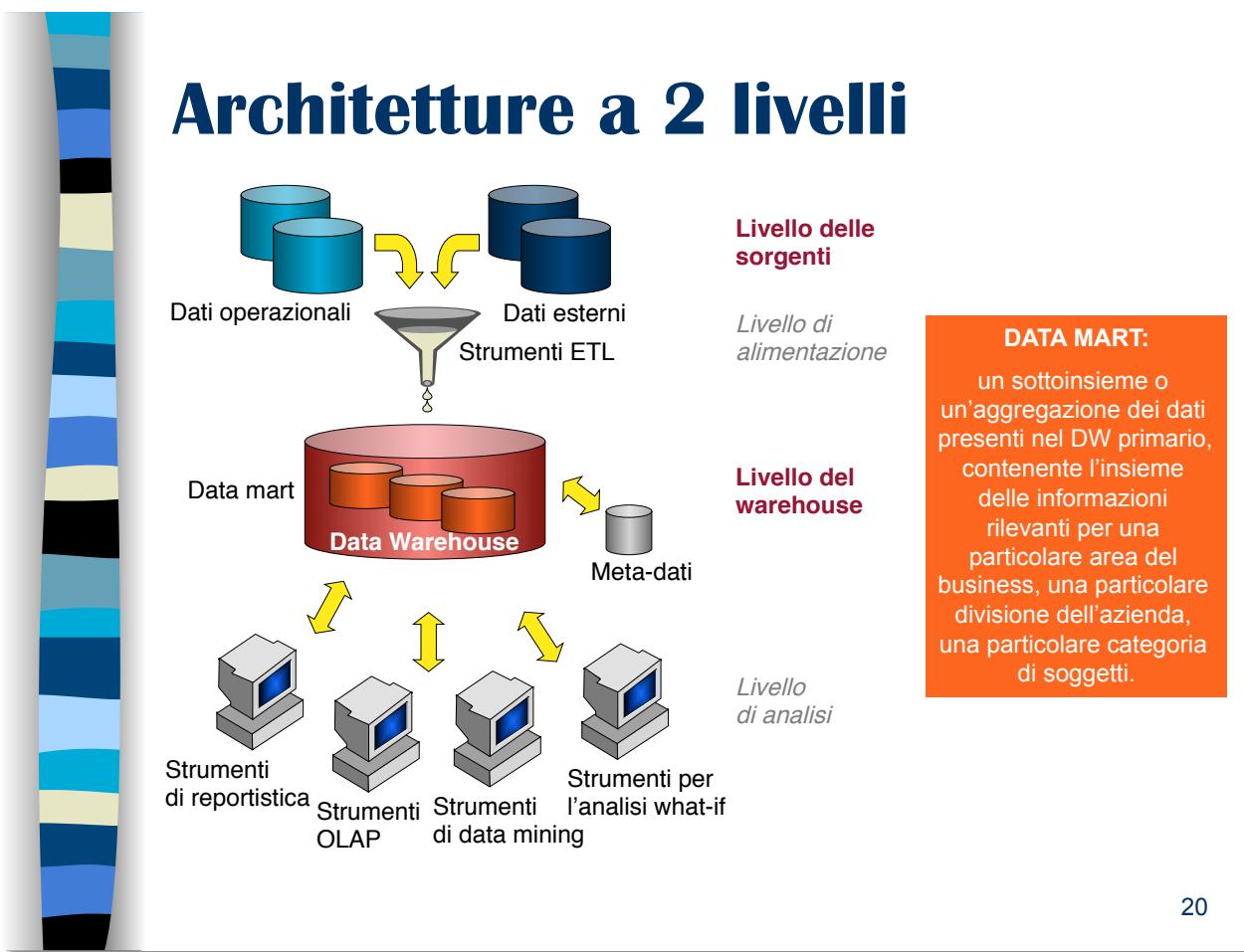
18

Architetture a 1 livello

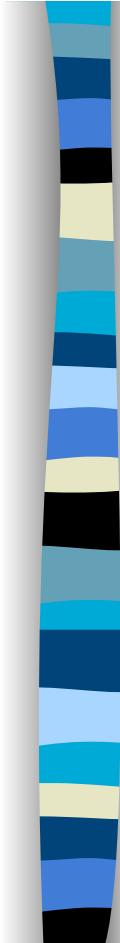


19

Architetture a 2 livelli



20



Architetture a 2 livelli

- I data mart alimentati dal DW primario sono detti *dipendenti*. Per i sistemi collocati all'interno di realtà aziendali medio-grandi essi sono utili:
 - ✓ come blocchi costruttivi durante la realizzazione incrementale del DW;
 - ✓ in quanto delineano i contorni delle informazioni necessarie a un particolare tipo di utenti per le loro interrogazioni;
 - ✓ poiché, essendo di dimensioni inferiori al DW primario, permettono di raggiungere prestazioni migliori
- In alcuni contesti si preferisce adottare data mart alimentati direttamente dalle sorgenti, detti *indipendenti*
 - ✓ L'assenza di un DW primario snellisce le fasi progettuali, ma determina uno schema complesso di accessi ai dati e ingenera il rischio di inconsistenze tra i data mart

21

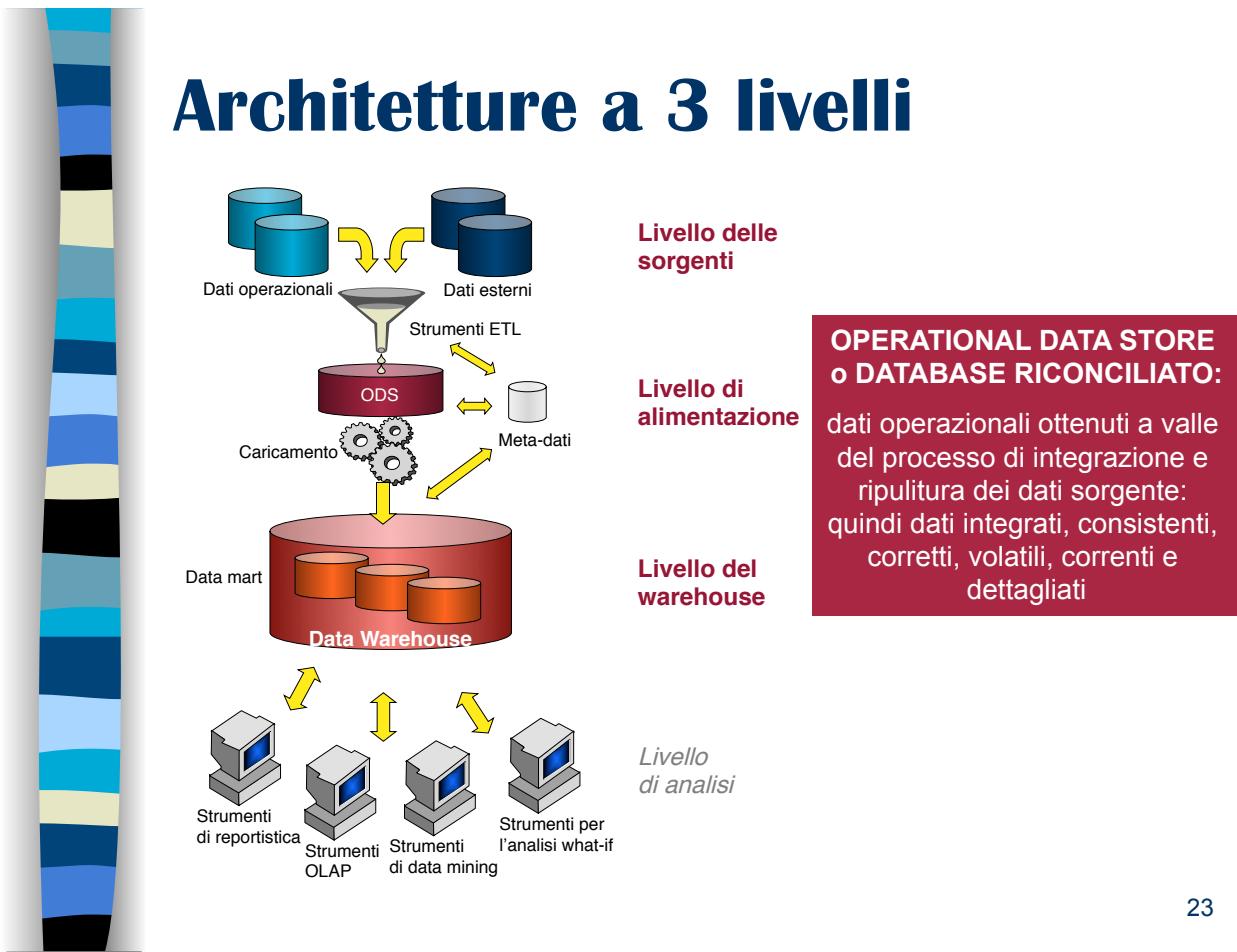


Architetture a 2 livelli

- Vantaggi:
 - ✓ A livello del warehouse è continuamente disponibile informazione di buona qualità anche quando, per motivi tecnici oppure organizzativi, è temporaneamente precluso l'accesso alle sorgenti
 - ✓ L'interrogazione analitica effettuata sul DW non interferisce con la gestione delle transazioni a livello operazionale, la cui affidabilità è essenziale per il funzionamento dell'azienda
 - ✓ L'organizzazione logica del DW è basata sul modello multidimensionale, mentre le sorgenti offrono in genere modelli relazionali o semi-strutturati
 - ✓ C'è una discordanza temporale e di granularità tra sistemi OLTP, che trattano dati correnti e al massimo livello di dettaglio, e sistemi OLAP che operano su dati storici e di sintesi
 - ✓ A livello del warehouse è possibile impiegare tecniche specifiche per ottimizzare le prestazioni per applicazioni di analisi e reportistica

22

Architetture a 3 livelli

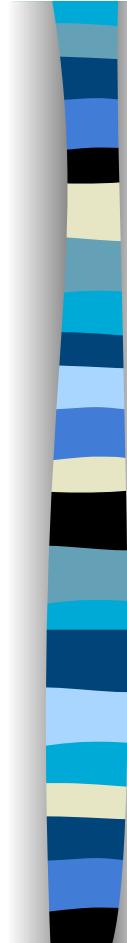


23

Architetture a 3 livelli

- Il vantaggio principale del livello dei dati riconciliati è che esso crea un modello di dati comune e di riferimento per l'intera azienda, introducendo al contempo una separazione netta tra le problematiche legate all'estrazione e integrazione dei dati dalle sorgenti e quelle inerenti l'alimentazione del DW
- D'altro canto, i dati riconciliati introducono un'ulteriore ridondanza rispetto ai dati operazionali sorgente

24

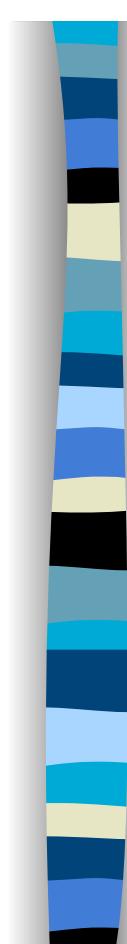


Riconciliazione... terminologica

- Data Warehouse
 - ✓ Contiene dati denormalizzati (tipicamente star e snowflake schema)
- Data Mart
 - ✓ Corrisponde a una porzione del data warehouse
 - ✓ Include più cubi multidimensionali, per cui “data mart” non è sinonimo di “cubo multidimensionale”
- Cubo multidimensionale
 - ✓ Implementa un fatto che lo modella a livello concettuale
- Misura
 - ✓ Chiamata anche “metrica” o “KPI”
- Database riconciliato o Operational Data Store (ODS)
 - ✓ Se presente, è parte integrante dell’architettura
 - ✓ Espone un modello coerente del business, contiene dati normalizzati e può essere usato per la reportistica operativa
 - ✓ Non è sinonimo di “data warehouse” (anche se alcuni lo chiamano erroneamente “data warehouse”)
- Staging area
 - ✓ Area di lavoro dell’ETL, spesso erroneamente confusa con l’ODS
 - ✓ Ha un ruolo di servizio all’interno dell’architettura

© M. Golfarelli e S. Rizzi 2011

25



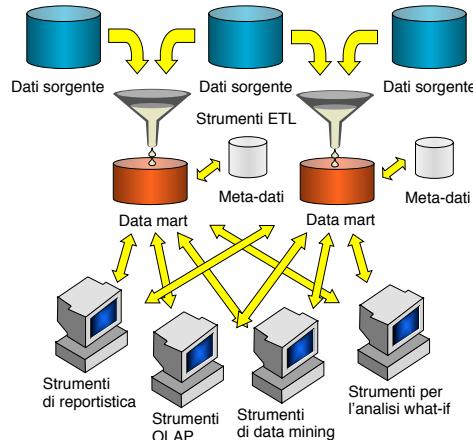
Architetture: un’altra classificazione

- Data mart indipendenti
- Data mart bus
- Hub-and-spoke
- Federazione

26

Data mart indipendenti

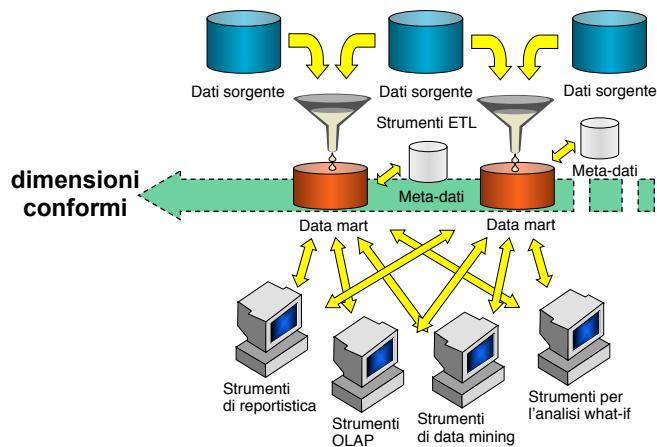
- Primo approccio al data warehousing
- Problema dell'inconsistenza (*data silos*)



27

Data mart bus

- Approccio consigliato da Kimball
- Integrazione a livello logico
- “Enterprise view”

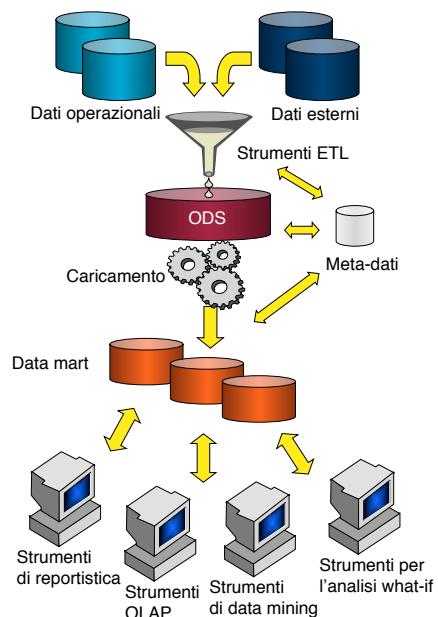


DIMENSIONI CONFORMI:
concetti di primaria
importanza per il
business, condivisi da
larga parte dei data
mart

28

Hub-and-spoke

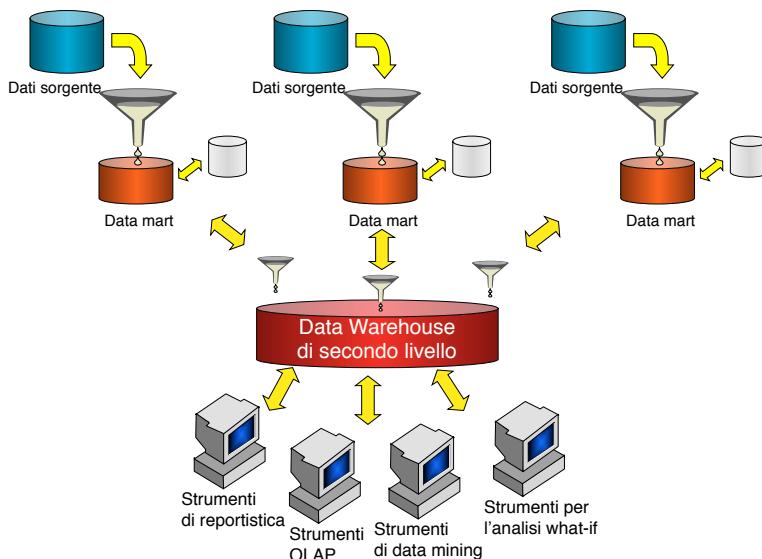
- Una delle architetture più usate in contesti medio-grandi



29

Federazione

- Ideale per contesti molto dinamici (fusioni-acquisizioni)
- Problema dell'integrazione efficace ed efficiente



30



Fattori di scelta dell'architettura

- Interdipendenza informativa tra le unità organizzative aziendali
 - ✓ incoraggia l'adozione di architetture enterprise-wide
- Urgenza del progetto di data warehousing
 - ✓ incoraggia l'adozione di architetture "veloci"
- Vincoli sulle risorse economiche e umane
- Ruolo del progetto di warehousing all'interno della strategia aziendale
 - ✓ data mart indipendenti vs. hub-and-spoke
- Compatibilità con piattaforme esistenti
- Capacità dello staff IT
- Posizione organizzativa dello sponsor di progetto
 - ✓ architetture aziendali vs. architetture dipartimentali

31

ETL

- Il ruolo degli strumenti di *Extraction, Transformation and Loading* è quello di alimentare una sorgente dati singola, dettagliata, esaurente e di alta qualità che possa a sua volta alimentare il DW (*riconciliazione*)
- Durante il processo di alimentazione del DW, la riconciliazione avviene in due occasioni: quando il DW viene popolato per la prima volta, e periodicamente quando il DW viene aggiornato.
 - ✓ estrazione
 - ✓ pulitura
 - ✓ trasformazione
 - ✓ caricamento

32

Estrazione

- I dati rilevanti vengono estratti dalle sorgenti
 - ✓ L'estrazione **statica** viene effettuata quando il DW deve essere popolato per la prima volta e consiste concettualmente in una fotografia dei dati operazionali
 - ✓ L'estrazione **incrementale** viene usata per l'aggiornamento periodico del DW, e cattura solamente i cambiamenti avvenuti nelle sorgenti dall'ultima estrazione
 - basata sul log mantenuto dal DBMS operazionale
 - basata su time-stamp
 - guidata dalle sorgenti
- La scelta dei dati da estrarre avviene principalmente in base alla loro qualità



33

Pulitura

- Si incarica di migliorare la qualità dei dati delle sorgenti
 - ✓ dati duplicati
 - ✓ inconsistenza tra valori logicamente associati
 - ✓ dati mancanti
 - ✓ uso non previsto di un campo
 - ✓ valori impossibili o errati
 - ✓ valori inconsistenti per la stessa entità dovuti a errori di battitura



34

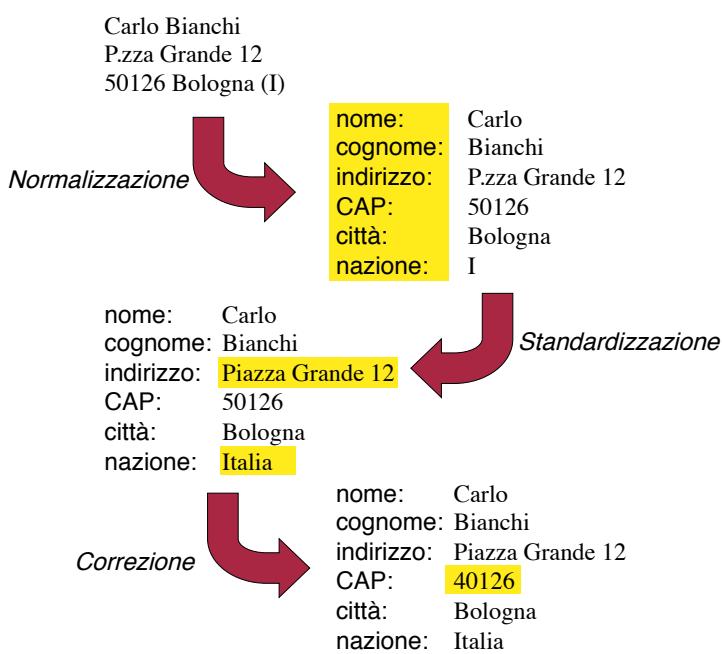
Trasformazione

- Converte i dati dal formato operazionale sorgente a quello del DW. La corrispondenza con il livello sorgente è complicata dalla presenza di fonti distinte eterogenee, che richiede una complessa fase di integrazione
 - ✓ presenza di testi liberi che nascondono informazioni importanti
 - ✓ utilizzo di formati e convenzioni differenti per lo stesso dato
- Per l'alimentazione dei dati riconciliati:
 - ✓ conversione e normalizzazione (operano a livello di formato di memorizzazione e di unità di misura per uniformare i dati)
 - ✓ matching (stabilisce corrispondenze tra campi equivalenti in sorgenti diverse)
 - ✓ selezione (riduce il numero di campi e di record rispetto alle sorgenti)
- Per l'alimentazione del DW:
 - ✓ la normalizzazione è sostituita dalla denormalizzazione
 - ✓ si introduce l'aggregazione, che realizza le opportune sintesi dei dati



35

Pulitura e trasformazione



36

Caricamento

- Il caricamento dei dati nel DW
 - ✓ Refresh: i dati del DW vengono riscritti integralmente, sostituendo quelli precedenti (tecnica utilizzata per popolare inizialmente il DW)
 - ✓ Update: i soli cambiamenti occorsi nei dati sorgente vengono aggiunti nel DW (tecnica utilizzata per l'aggiornamento periodico del DW)



37

Verso il modello multidimensionale

“Che incassi sono stati registrati l’anno passato per ciascuna regione e ciascuna categoria di prodotto?”

“Che rapporto c’è tra l’andamento dei titoli azionari dei produttori di PC e i profitti trimestrali lungo gli ultimi 5 anni?”

“Quali sono le tipologie di ordini che massimizzano gli incassi?”

“Quale di due nuove terapie risulta più efficace ai fini della diminuzione della durata media di un ricovero?”

“Che rapporto c’è tra i profitti realizzati con spedizioni di meno di 10 elementi e quelli realizzati con spedizioni di più di 10 elementi?”

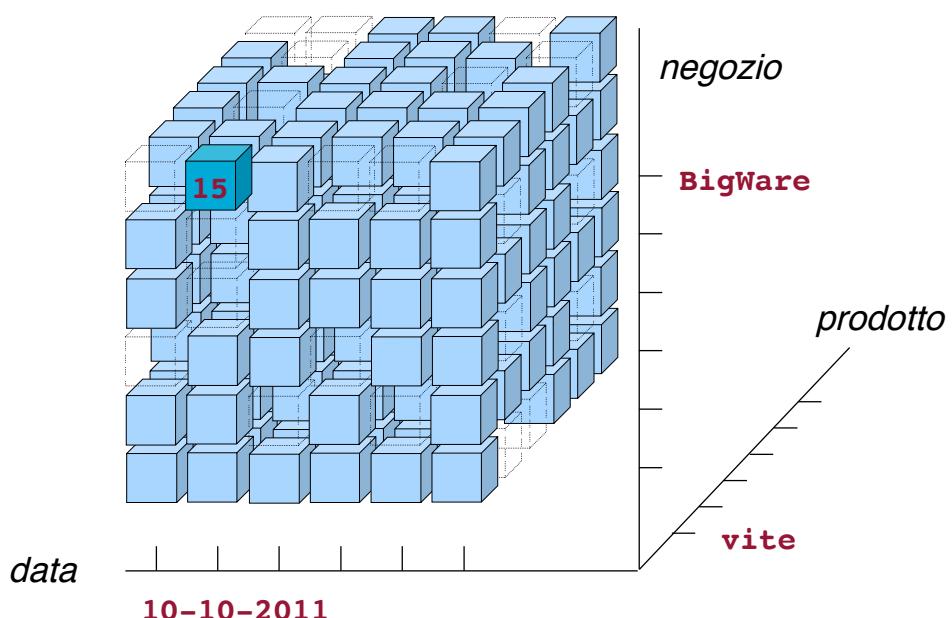
38

Il modello multidimensionale

- È il fondamento per la rappresentazione e l'interrogazione dei dati nei data warehouse.
- I *fatti* di interesse sono rappresentati in *cubi* in cui:
 - ✓ ogni cella contiene *misure* numeriche che quantificano il fatto da diversi punti di vista;
 - ✓ ogni asse rappresenta una *dimensione* di interesse per l'analisi;
 - ✓ ogni dimensione può essere la radice di una *gerarchia* di attributi usati per aggregare i dati memorizzati nei cubi base.

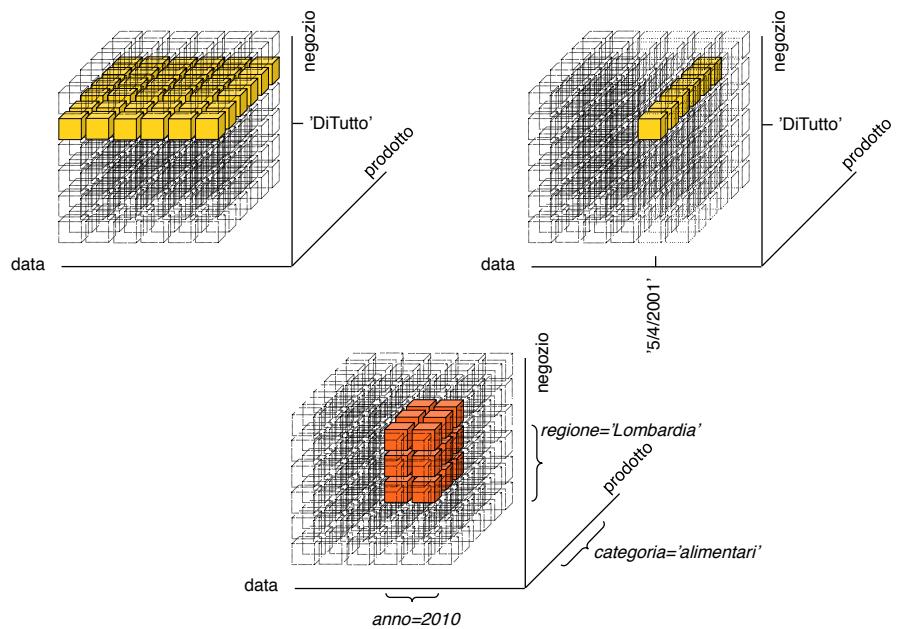
39

Il cubo delle vendite



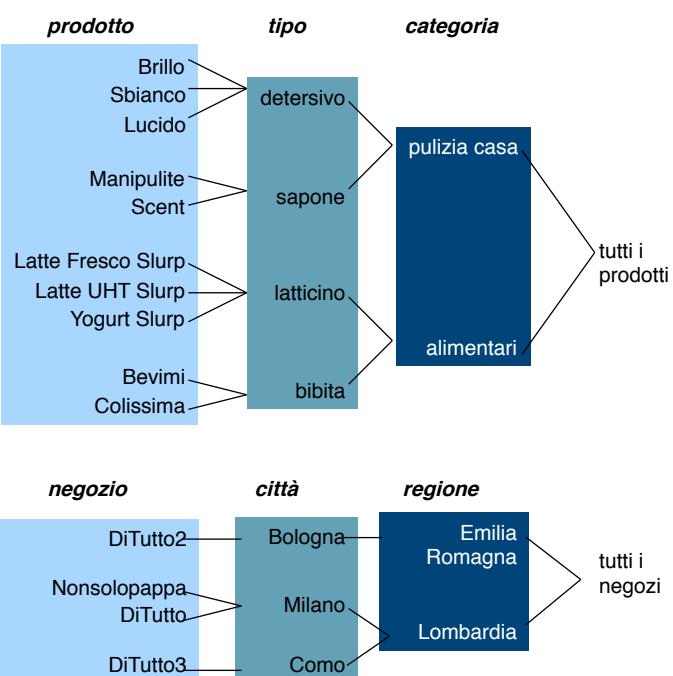
40

Slicing and dicing



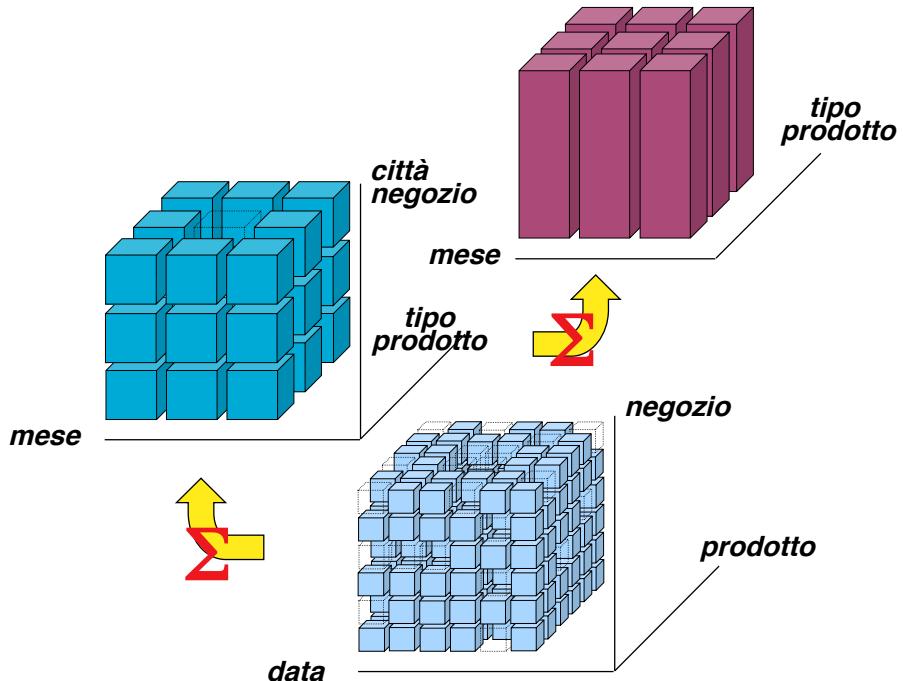
41

Le gerarchie



42

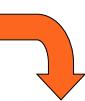
Aggregazione



43

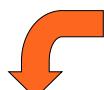
Aggregazione

	DiTutto	DiTutto2	Nonsolopappa
1/1/2000	—	—	—
2/1/2000	10	15	5
3/1/2000	20	—	5
.....
1/1/2001	—	—	—
2/1/2001	15	10	20
3/1/2001	20	20	25
.....
1/1/2002	—	—	—
2/1/2002	20	8	25
3/1/2002	20	12	20
.....



	DiTutto	DiTutto2	Nonsolopappa
Gennaio 2000	200	180	150
Febbraio 2000	180	150	120
Marzo 2000	220	180	160
.....
Gennaio 2001	350	220	200
Febbraio 2001	300	200	250
Marzo 2001	310	180	300
.....
Gennaio 2002	380	200	220
Febbraio 2002	310	200	250
Marzo 2002	300	160	280
.....

	DiTutto	DiTutto2	Nonsolopappa
2000	2400	2000	1600
2001	3200	2300	3000
2002	3400	2200	3200



Total:	DiTutto	DiTutto2	Nonsolopappa
	9000	6500	7800

44

Tecniche di analisi dei dati

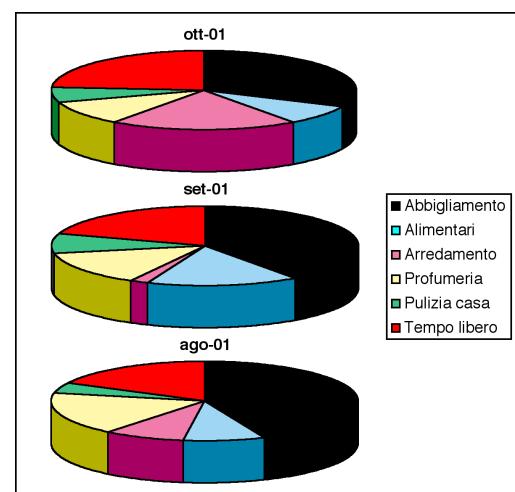
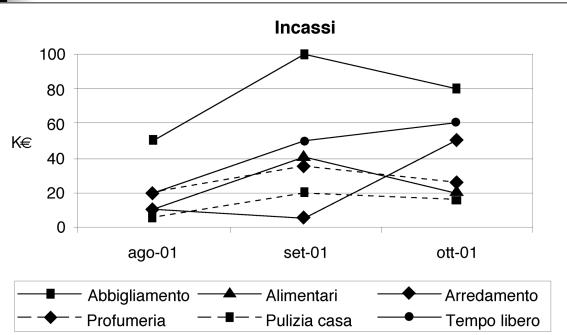
- Una volta che i dati sono stati ripuliti, integrati e trasformati, occorre capire come trarne il massimo vantaggio informativo
- Esistono due approcci differenti, supportati da altrettante categorie di strumenti, all'interrogazione di un DW da parte degli utenti finali:
 - ✓ *reportistica*: non richiede conoscenze informatiche
 - ✓ *OLAP*: richiede all'utente di ragionare in modo multidimensionale e di conoscere l'interfaccia dello strumento grafico utilizzato

45

Reportistica

orientato agli utenti
che hanno necessità
di accedere, a
intervalli di tempo
predefiniti, a
informazioni
strutturate in modo
pressoché invariabile

incassi (K€)	Ottobre 2001	Settembre 2001	Agosto 2001
Abbigliamento	80	100	50
Alimentari	20	40	10
Arredamento	50	5	10
Profumeria	25	35	20
Pulizia casa	15	20	5
Tempo libero	60	50	20



46

Reportistica



47

OLAP

- È la principale modalità di fruizione delle informazioni contenute in un DW
- Consente, a utenti le cui necessità di analisi non siano facilmente identificabili a priori, di analizzare ed esplorare interattivamente i dati sulla base del modello multidimensionale
- Mentre gli utenti degli strumenti di reportistica svolgono un ruolo essenzialmente passivo, gli utenti OLAP sono in grado di costruire attivamente una sessione di analisi complessa in cui ciascun passo effettuato è conseguenza dei risultati ottenuti al passo precedente
 - ✓ estemporaneità delle sessioni di lavoro
 - ✓ richiesta approfondita conoscenza dei dati
 - ✓ complessità delle interrogazioni formulabili
 - ✓ orientamento verso utenti non esperti di informatica



interfaccia flessibile, facile
da usare ed efficace

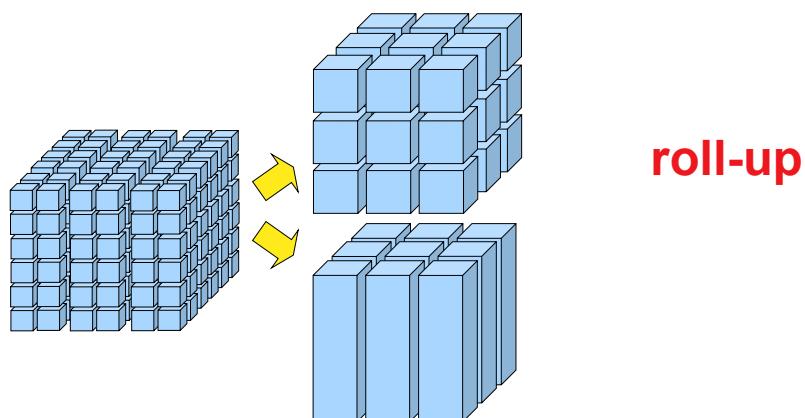
48

OLAP: sessione

- Una sessione OLAP consiste in un *percorso di navigazione* che riflette il procedimento di analisi di uno o più fatti di interesse sotto diversi aspetti e a diversi livelli di dettaglio. Questo percorso si concretizza in una sequenza di interrogazioni spesso formulate non direttamente, ma per differenza rispetto all'interrogazione precedente
- Ogni passo della sessione di analisi è scandito dall'applicazione di un **operatore OLAP** che trasforma l'ultima interrogazione formulata in una nuova interrogazione
- Il risultato delle interrogazioni è di tipo multidimensionale; gli strumenti OLAP rappresentano tipicamente i dati in modo tabellare evidenziando le diverse dimensioni mediante intestazioni multiple, colori ecc.

49

OLAP: operatori



50

OLAP: operatori

	Metrics Customer Region	Dollar Sales										
Month		North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germany	Canada
Jan 97		\$ 620	\$ 753	\$ 30	\$ 660	\$ 2,405	\$ 1,312	\$ 440	\$ 1,002	\$ 1,002	\$ 383	\$ 210
Feb 97		\$ 258	\$ 252	\$ 800	\$ 975	\$ 160	\$ 582	\$ 744	\$ 310	\$ 799	\$ 118	\$ 357
Mar 97		\$ 648	\$ 244	\$ 148	\$ 250	\$ 1,085	\$ 2,961	\$ 650	\$ 1,240	\$ 119	\$ 142	\$ 96
Apr 97		\$ 787	\$ 588	\$ 447	\$ 486	\$ 226	\$ 506	\$ 601	\$ 119	\$ 550	\$ 85	
May 97		\$ 1,350	\$ 245	\$ 936	\$ 159	\$ 664	\$ 626	\$ 107	\$ 135	\$ 200	\$ 177	\$ 230
Jun 97		\$ 842	\$ 582	\$ 1,281	\$ 937	\$ 240	\$ 774	\$ 176	\$ 1,139	\$ 652	\$ 254	\$ 745
Jul 97		\$ 652	\$ 690	\$ 486	\$ 1,293	\$ 605	\$ 303	\$ 818	\$ 103	\$ 124	\$ 173	\$ 66
Aug 97		\$ 1,783	\$ 304	\$ 1,032	\$ 170	\$ 398	\$ 356	\$ 432	\$ 190	\$ 241	\$ 407	\$ 259
Sep 97		\$ 581	\$ 778	\$ 3,558	\$ 587	\$ 440	\$ 1,652	\$ 1,071	\$ 315	\$ 210	\$ 202	
Oct 97		\$ 2,291	\$ 1,840	\$ 600	\$ 656	\$ 1,300	\$ 718	\$ 1,210	\$ 427	\$ 220	\$ 520	\$ 65
Nov 97		\$ 39	\$ 1,602	\$ 1,082	\$ 1,187	\$ 842	\$ 759	\$ 745	\$ 232	\$ 101	\$ 1,037	\$ 37
Dec 97		\$ 381	\$ 1,586	\$ 343	\$ 118	\$ 1,459	\$ 635	\$ 2,021	\$ 259	\$ 210	\$ 119	\$ 189
Jan 98		\$ 311	\$ 1,174	\$ 2,634	\$ 3,130	\$ 954	\$ 2,083	\$ 1,351	\$ 747	\$ 426	\$ 447	\$ 1,141
Feb 98		\$ 2,518	\$ 702	\$ 1,123	\$ 1,336	\$ 1,227	\$ 3,887	\$ 545	\$ 268	\$ 277	\$ 282	
Mar 98		\$ 2,459	\$ 1,523	\$ 1,178	\$ 4,708	\$ 1,420	\$ 1,948	\$ 1,705	\$ 276	\$ 1,168	\$ 63	
Apr 98		\$ 407	\$ 841	\$ 524	\$ 712	\$ 133	\$ 2,486	\$ 49	\$ 390	\$ 1,298	\$ 221	\$ 46
May 98		\$ 667	\$ 1,721	\$ 440	\$ 148	\$ 80	\$ 1,310	\$ 303	\$ 104	\$ 657	\$ 65	
Jun 98		\$ 699	\$ 1,096	\$ 898	\$ 353	\$ 902	\$ 839	\$ 230	\$ 155	\$ 105	\$ 75	
Jul 98		\$ 586	\$ 1,897	\$ 412	\$ 226	\$ 406	\$ 361	\$ 1,628	\$ 267	\$ 1,011	\$ 41	\$ 184
Aug 98		\$ 894	\$ 326	\$ 792	\$ 1,832	\$ 1,199	\$ 295	\$ 1,816	\$ 277	\$ 102	\$ 118	\$ 115
Sep 98		\$ 338	\$ 3,179	\$ 505	\$ 427	\$ 99	\$ 2,976	\$ 885	\$ 135	\$ 85	\$ 1,110	\$ 510
Oct 98		\$ 544	\$ 413	\$ 1,467	\$ 209	\$ 679	\$ 706	\$ 556	\$ 480	\$ 485	\$ 99	\$ 160
Nov 98		\$ 671	\$ 459	\$ 1,471	\$ 2,066	\$ 701	\$ 716	\$ 986	\$ 1,217	\$ 154	\$ 440	\$ 361
Dec 98		\$ 836	\$ 2,096	\$ 1,726	\$ 3,642	\$ 395	\$ 1,740	\$ 1,943	\$ 1,143	\$ 366	\$ 307	\$ 118

roll-up

	Metrics Customer Region	Dollar Sales										
Quarter		North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germany	Canada
Q1 1997		\$ 1,526	\$ 1,249	\$ 978	\$ 1,895	\$ 3,650	\$ 4,855	\$ 1,834	\$ 2,552	\$ 1,920	\$ 643	\$ 663
Q2 1997		\$ 2,979	\$ 1,415	\$ 2,664	\$ 1,582	\$ 1,130	\$ 1,906	\$ 884	\$ 1,393	\$ 1,402	\$ 516	\$ 975
Q3 1997		\$ 3,016	\$ 1,772	\$ 5,076	\$ 2,050	\$ 1,443	\$ 2,311	\$ 2,321	\$ 608	\$ 575	\$ 782	\$ 325
Q4 1997		\$ 2,711	\$ 5,030	\$ 2,025	\$ 1,961	\$ 3,601	\$ 2,112	\$ 3,976	\$ 918	\$ 531	\$ 1,676	\$ 291
Q1 1998		\$ 5,288	\$ 3,399	\$ 4,935	\$ 9,174	\$ 3,601	\$ 9,484	\$ 3,844	\$ 2,720	\$ 979	\$ 1,897	\$ 1,204
Q2 1998		\$ 1,773	\$ 3,658	\$ 1,862	\$ 2,123	\$ 1,115	\$ 4,635	\$ 352	\$ 724	\$ 2,110	\$ 391	\$ 121
Q3 1998		\$ 1,818	\$ 5,402	\$ 1,709	\$ 2,485	\$ 1,704	\$ 3,632	\$ 4,329	\$ 679	\$ 1,198	\$ 1,269	\$ 809
Q4 1998		\$ 2,051	\$ 2,968	\$ 4,664	\$ 5,917	\$ 1,775	\$ 3,162	\$ 3,485	\$ 2,750	\$ 1,005	\$ 846	\$ 639

51

OLAP: operatori

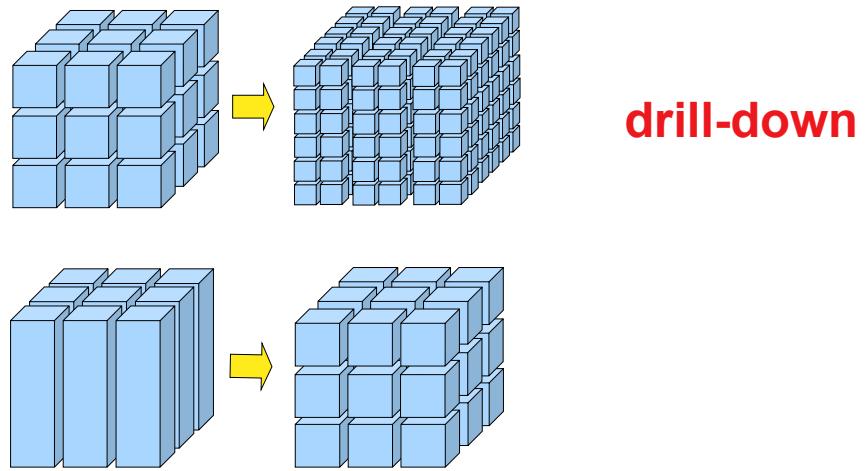
	Metrics Customer Region	Dollar Sales										
Category	Year	North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germany	Canada
Electronics	1997	\$ 138	\$ 1,774	\$ 384	\$ 138	\$ 2,346	\$ 2,554	\$ 2,184	\$ 566	\$ 199	\$	
	1998	\$ 1,184	\$ 4,529	\$ 1,892	\$ 7,232	\$ 651	\$ 9,488	\$ 476	\$ 2,683	\$ 462	\$ 7	
Food	1997	\$ 759	\$ 682	\$ 729	\$ 262	\$ 588	\$ 469	\$ 807	\$ 156	\$ 615	\$ 1	
	1998	\$ 538	\$ 925	\$ 959	\$ 677	\$ 213	\$ 1,503	\$ 261	\$ 165	\$ 175	\$ 1	
Gifts	1997	\$ 2,532	\$ 1,355	\$ 1,854	\$ 1,413	\$ 2,535	\$ 2,132	\$ 1,904	\$ 908	\$ 375	\$ 1,0	
	1998	\$ 1,955	\$ 2,785	\$ 2,800	\$ 2,695	\$ 1,813	\$ 2,844	\$ 1,778	\$ 1,158	\$ 717	\$ 6	
Health & Beauty	1997	\$ 624	\$ 640	\$ 1,317	\$ 647	\$ 588	\$ 754	\$ 654	\$ 143	\$ 292	\$ 3	
	1998	\$ 611	\$ 887	\$ 566	\$ 382	\$ 499	\$ 1,162	\$ 1,044	\$ 273	\$ 72		
Household	1997	\$ 5,354	\$ 4,112	\$ 5,410	\$ 4,446	\$ 3,058	\$ 3,974	\$ 2,654	\$ 3,545	\$ 2,875	\$ 1,9	
	1998	\$ 5,787	\$ 5,320	\$ 5,416	\$ 6,812	\$ 4,334	\$ 5,008	\$ 7,588	\$ 2,139	\$ 3,649	\$ 2,7	
Kid's Korner	1997	\$ 201	\$ 398	\$ 485	\$ 186	\$ 409	\$ 323	\$ 396	\$ 105	\$ 34	\$	
	1998	\$ 247	\$ 422	\$ 441	\$ 380	\$ 221	\$ 592	\$ 290	\$ 198	\$ 19	\$	
Travel	1997	\$ 624	\$ 505	\$ 564	\$ 386	\$ 300	\$ 978	\$ 416	\$ 48	\$ 38		
	1998	\$ 608	\$ 559	\$ 1,096	\$ 611	\$ 454	\$ 316	\$ 573	\$ 257	\$ 198	\$	

roll-up

	Metrics	Dollar Sales
Category	Year	
Electronics	1997	\$ 10,616
	1998	\$ 29,299
Food	1997	\$ 5,300
	1998	\$ 5,638
Gifts	1997	\$ 16,315
	1998	\$ 20,047
Health & Beauty	1997	\$ 6,042
	1998	\$ 5,665
Household	1997	\$ 38,383
	1998	\$ 50,391
Kid's Korner	1997	\$ 2,559
	1998	\$ 2,943
Travel	1997	\$ 4,497
	1998	\$ 4,792

52

OLAP: operatori



53

OLAP: operatori

	Metrics	Dollar Sales	North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germany	Canada
Quarter	Customer Region												
Q1 1997		\$ 1.526	\$ 1.249	\$ 978	\$ 1.885	\$ 3.650	\$ 4.855	\$ 1.834	\$ 2.552	\$ 1.920	\$ 643	\$ 663	
Q2 1997		\$ 2.979	\$ 1.415	\$ 2.664	\$ 1.582	\$ 1.130	\$ 1.906	\$ 884	\$ 1.393	\$ 1.402	\$ 516	\$ 975	
Q3 1997		\$ 3.016	\$ 1.772	\$ 5.076	\$ 2.050	\$ 1.443	\$ 2.311	\$ 2.321	\$ 608	\$ 575	\$ 782	\$ 325	
Q4 1997		\$ 2.711	\$ 5.030	\$ 2.025	\$ 1.961	\$ 3.601	\$ 2.112	\$ 3.976	\$ 918	\$ 531	\$ 1.676	\$ 291	
Q1 1998		\$ 5.288	\$ 3.399	\$ 4.935	\$ 9.174	\$ 3.601	\$ 9.484	\$ 3.844	\$ 2.720	\$ 979	\$ 1.897	\$ 1.201	
Q2 1998		\$ 1.773	\$ 3.658	\$ 1.862	\$ 1.213	\$ 1.115	\$ 4.635	\$ 352	\$ 724	\$ 2.110	\$ 391	\$ 12	
Q3 1998		\$ 1.818	\$ 5.402	\$ 1.709	\$ 2.485	\$ 1.704	\$ 3.632	\$ 4.329	\$ 679	\$ 1.198	\$ 1.269	\$ 809	
Q4 1998		\$ 2.051	\$ 2.968	\$ 4.664	\$ 5.917	\$ 1.775	\$ 3.162	\$ 3.485	\$ 2.750	\$ 1.005	\$ 846	\$ 639	

drill-down



	Metrics	Dollar Sales	Arlin	San Pedro	Springfield	Chappel Hill	Scranburg	Pebble Beach	Martinsville	Maddon	Peoria	Pecos	Lake Barkley	Alameda	Fingers Lake
Quarter	Customer City														
Q1 1997		\$ 675											\$ 39		
Q2 1997														\$ 135	
Q3 1997														\$ 252	\$ 63
Q4 1997		\$ 215	\$ 124											\$ 79	\$ 98
Q1 1998														\$ 237	\$ 30
Q2 1998														\$ 30	\$ 119
Q3 1998		\$ 734													
Q4 1998															

54

OLAP: operatori

Category	Metrics		Dollar Sales	
	Year	1997	1998	
Electronics	\$ 10,616	\$ 29,299		
Food	\$ 5,900	\$ 5,638		
Gifts	\$ 16,315	\$ 20,047		
Health & Beauty	\$ 6,042	\$ 5,665		
Household	\$ 38,383	\$ 50,391		
Kid's Korner	\$ 2,559	\$ 2,943		
Travel	\$ 4,497	\$ 4,792		

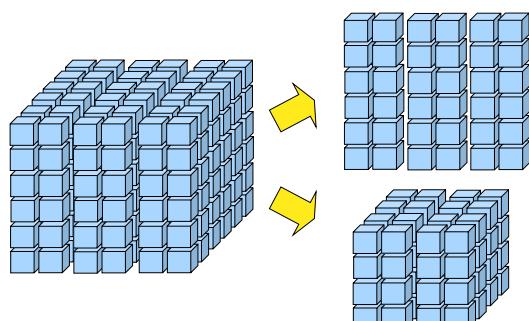
drill-down



Category	Metrics		Dollar Sales		North-East		Mid-Atlantic		South-East		Central		South		North-West		West		
	Customer Region	Year	North-East		Mid-Atlantic		South-East		Central		South		North-West		West		Europe		
			1997	1998	1997	1998	1997	1998	1997	1998	1997	1998	1997	1998	1997	1998	1997	1998	
Electronics	\$ 138	\$ 1,184	\$ 1,774	\$ 4,529	\$ 384	\$ 1,892	\$ 138	\$ 7,232	\$ 2,346	\$ 651	\$ 2,554	\$ 9,488	\$ 1,184	\$ 1,774	\$ 4,529	\$ 384	\$ 1,892	\$ 138	
Food	\$ 759	\$ 538	\$ 682	\$ 925	\$ 729	\$ 959	\$ 262	\$ 677	\$ 588	\$ 213	\$ 469	\$ 1,503	\$ 682	\$ 925	\$ 729	\$ 959	\$ 262	\$ 677	\$ 588
Gifts	\$ 2,532	\$ 1,955	\$ 1,355	\$ 2,785	\$ 1,854	\$ 2,800	\$ 1,413	\$ 2,695	\$ 2,535	\$ 1,813	\$ 2,132	\$ 2,844	\$ 1,355	\$ 2,785	\$ 1,854	\$ 2,800	\$ 1,413	\$ 2,695	\$ 2,535
Health & Beauty	\$ 624	\$ 611	\$ 640	\$ 887	\$ 1,317	\$ 566	\$ 647	\$ 382	\$ 588	\$ 499	\$ 754	\$ 1,162	\$ 611	\$ 640	\$ 887	\$ 1,317	\$ 566	\$ 647	\$ 382
Household	\$ 5,354	\$ 5,787	\$ 4,112	\$ 5,320	\$ 5,410	\$ 5,416	\$ 4,446	\$ 6,012	\$ 3,058	\$ 4,334	\$ 3,974	\$ 5,008	\$ 5,787	\$ 4,112	\$ 5,320	\$ 5,410	\$ 5,416	\$ 4,446	\$ 6,012
Kid's Korner	\$ 201	\$ 247	\$ 398	\$ 422	\$ 485	\$ 441	\$ 186	\$ 380	\$ 409	\$ 221	\$ 323	\$ 592	\$ 247	\$ 398	\$ 422	\$ 485	\$ 441	\$ 186	\$ 380
Travel	\$ 624	\$ 608	\$ 505	\$ 559	\$ 564	\$ 1,096	\$ 386	\$ 611	\$ 300	\$ 464	\$ 978	\$ 316	\$ 505	\$ 559	\$ 611	\$ 300	\$ 464	\$ 978	\$ 316

55

OLAP: operatori



slice-and-dice

56

OLAP: operatori

A 3D bar chart illustrating sales data. The bars are color-coded by category: Electronics (blue), Food (orange), Gifts (green), Health & Beauty (red), Household (purple), Kid's Korner (yellow), and Travel (pink). The chart shows sales figures for each category across different years (1997-1999) and regions (North-East, Mid-Atlantic, South-East, Central, South, North-West, South-West, England, France, Germany).

Category	Year	Metrics Customer Region									
		North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germany
Electronics	1997	\$ 138	\$ 1.774	\$ 384	\$ 138	\$ 2.346	\$ 2.554	\$ 2.184	\$ 566	\$ 199	\$ 7
Food	1997	\$ 1.184	\$ 4.529	\$ 1.892	\$ 7.232	\$ 651	\$ 9.488	\$ 476	\$ 2.683	\$ 462	\$ 1
Gifts	1997	\$ 759	\$ 682	\$ 729	\$ 262	\$ 588	\$ 469	\$ 807	\$ 156	\$ 615	\$ 1
Health & Beauty	1997	\$ 538	\$ 925	\$ 959	\$ 677	\$ 213	\$ 1.503	\$ 261	\$ 165	\$ 175	\$ 1
Household	1997	\$ 2.532	\$ 1.355	\$ 1.854	\$ 1.413	\$ 2.535	\$ 2.132	\$ 1.904	\$ 908	\$ 375	\$ 1.0
Kid's Korner	1997	\$ 1.955	\$ 2.785	\$ 2.800	\$ 2.695	\$ 1.813	\$ 2.844	\$ 1.778	\$ 1.158	\$ 717	\$ 6
Travel	1997	\$ 624	\$ 640	\$ 1.317	\$ 647	\$ 588	\$ 754	\$ 654	\$ 143	\$ 292	\$ 3
Electronics	1998	\$ 611	\$ 887	\$ 566	\$ 382	\$ 499	\$ 1.162	\$ 1.044	\$ 273	\$ 72	
Food	1998	\$ 538	\$ 925	\$ 959	\$ 677	\$ 213	\$ 1.503	\$ 261	\$ 165	\$ 175	\$ 100
Gifts	1998	\$ 1.955	\$ 2.785	\$ 2.800	\$ 2.695	\$ 1.813	\$ 2.844	\$ 1.778	\$ 1.158	\$ 717	\$ 6
Health & Beauty	1998	\$ 624	\$ 640	\$ 1.317	\$ 647	\$ 588	\$ 754	\$ 654	\$ 143	\$ 292	\$ 3
Household	1998	\$ 5.787	\$ 5.320	\$ 5.416	\$ 6.812	\$ 4.334	\$ 5.008	\$ 7.588	\$ 2.139	\$ 3.649	\$ 1.9
Kid's Korner	1998	\$ 201	\$ 398	\$ 485	\$ 186	\$ 409	\$ 323	\$ 396	\$ 105	\$ 34	
Travel	1998	\$ 247	\$ 422	\$ 441	\$ 380	\$ 221	\$ 592	\$ 290	\$ 198	\$ 19	\$ 69
Electronics	1999	\$ 624	\$ 505	\$ 564	\$ 386	\$ 300	\$ 978	\$ 416	\$ 48	\$ 38	
Food	1999	\$ 608	\$ 559	\$ 1.096	\$ 611	\$ 464	\$ 316	\$ 573	\$ 257	\$ 198	\$ 55

slice-and-dice

57

OLAP: operatori

A 3D bar chart illustrating sales data. The bars are color-coded by subcategory: Audio (blue), Automotive (orange), Chocolate (green), Christmas (red), Classic Toys (purple), Coffee (yellow), Comfort (pink), Gadgets (light blue), Games & Puzzles (light green), Gift Baskets (light orange), Golf (light purple), Hearth (light pink), Jewelry (light green), Kitchen (light blue), Lawn & Garden (light orange), Learning (light green), Meat & Cheese (light blue), Miscellaneous (light green), Natural Remedies (light orange), Pets (light green), Plants & Flowers (light blue), Safety & Security (light green), Skin Care (light blue), Sleeping (light green), and Toys & Accessories (light orange). The chart shows sales figures for each subcategory across different cities (Afton, Akron, Albon, Alcameda, Alka, Allagash, Alta, Altoola, Amestra, Amsterdam, Andersonville, Annap) and metrics.

Subcategory	Customer City	Metrics											
		Dollar Sales	Afton	Akron	Albon	Alcameda	Alka	Allagash	Alta	Altoola	Amestra	Amsterdam	Andersonville
Audio							\$ 85						
Automotive								\$ 30					
Chocolate	\$ 42	\$ 42			\$ 50		\$ 20	\$ 22	\$ 44				
Christmas	\$ 30						\$ 25	\$ 30	\$ 15				
Classic Toys				\$ 9			\$ 7	\$ 26					\$ 38
Coffee							\$ 59						
Comfort								\$ 485					
Furniture								\$ 199	\$ 79	\$ 79			
Gadgets									\$ 17	\$ 45			\$ 45
Games & Puzzles													
Gift Baskets					\$ 55	\$ 43							
Golf	\$ 25												\$ 25
Hearth													
Jewelry	\$ 75												
Kitchen													
Lawn & Garden	\$ 75		\$ 100			\$ 15		\$ 63	\$ 100				
Learning	\$ 16												
Meat & Cheese													
Miscellaneous													
Natural Remedies	\$ 13												
Pets	\$ 215												
Plants & Flowers	\$ 65	\$ 65	\$ 65										
Safety & Security													
Skin Care													
Sleeping													
Toys & Accessories													

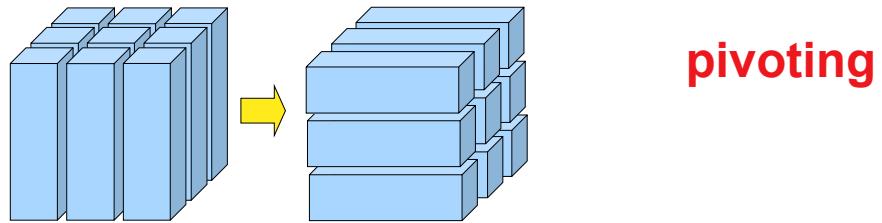
slice-and-dice

Filter Details:
Category = Electronics
AND
Dollar Sales > 80
AND
Customer Region = North-West
AND
Year = 1997

Subcategory	Customer City	Metrics					
		Dollar Sales	Alta	Armstrong	Avery Heights	Lane	Mt. Everest
Audio					\$ 98		
Comfort						\$ 118	\$ 1.495
Gadgets					\$ 199		

58

OLAP: operatori



59

OLAP: operatori

Category	Metrics	Dollar Sales	
		1997	1998
Electronics	1997	\$ 10.616	
	1998	\$ 29.299	
Food	1997	\$ 5.300	
	1998	\$ 5.638	
Gifts	1997	\$ 16.315	
	1998	\$ 20.047	
Health & Beauty	1997	\$ 6.042	
	1998	\$ 5.665	
Household	1997	\$ 38.383	
	1998	\$ 50.391	
Kid's Korner	1997	\$ 2.559	
	1998	\$ 2.943	
Travel	1997	\$ 4.497	
	1998	\$ 4.792	

pivoting

Category	Metrics	Dollar Sales	
		1997	1998
Electronics	Year	\$ 10.616	\$ 29.299
Food		\$ 5.300	\$ 5.638
Gifts		\$ 16.315	\$ 20.047
Health & Beauty		\$ 6.042	\$ 5.665
Household		\$ 38.383	\$ 50.391
Kid's Korner		\$ 2.559	\$ 2.943
Travel		\$ 4.497	\$ 4.792

60

OLAP: operatori

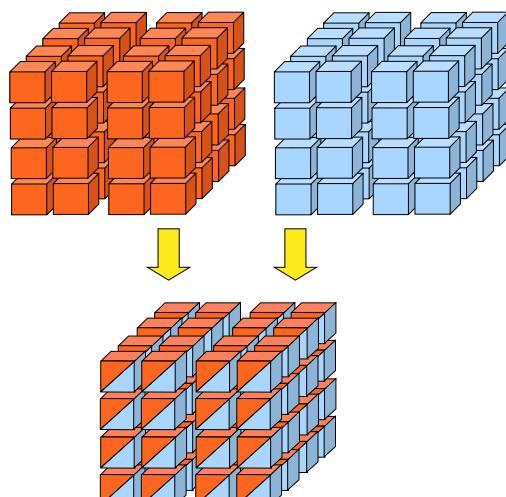
Category	Year	Metrics									
		Customer Region									
		Dollar Sales									
North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germany		
Electronics	1997	\$ 138	\$ 1.774	\$ 384	\$ 139	\$ 2.346	\$ 2.554	\$ 2.184	\$ 566	\$ 199	\$
	1998	\$ 1.184	\$ 4.529	\$ 1.892	\$ 7.232	\$ 651	\$ 9.488	\$ 476	\$ 2.683	\$ 462	\$ 7
Food	1997	\$ 759	\$ 682	\$ 729	\$ 262	\$ 588	\$ 469	\$ 807	\$ 156	\$ 615	\$ 1
	1998	\$ 538	\$ 925	\$ 959	\$ 677	\$ 213	\$ 1.503	\$ 261	\$ 165	\$ 175	\$ 1
Gifts	1997	\$ 2.532	\$ 1.355	\$ 1.854	\$ 1.413	\$ 2.535	\$ 2.132	\$ 1.904	\$ 908	\$ 375	\$ 1.0
	1998	\$ 1.955	\$ 2.785	\$ 2.800	\$ 2.695	\$ 1.813	\$ 2.844	\$ 1.778	\$ 1.158	\$ 717	\$ 6
Health & Beauty	1997	\$ 624	\$ 640	\$ 1.317	\$ 647	\$ 588	\$ 754	\$ 654	\$ 143	\$ 292	\$ 3
	1998	\$ 611	\$ 887	\$ 566	\$ 382	\$ 499	\$ 1.162	\$ 1.044	\$ 273	\$ 72	
Household	1997	\$ 5.354	\$ 4.112	\$ 5.410	\$ 4.446	\$ 3.058	\$ 3.974	\$ 2.654	\$ 3.545	\$ 2.875	\$ 1.9
	1998	\$ 5.787	\$ 5.320	\$ 5.416	\$ 6.812	\$ 4.334	\$ 5.008	\$ 7.588	\$ 2.139	\$ 3.649	\$ 2.7
Kid's Korner	1997	\$ 201	\$ 398	\$ 485	\$ 186	\$ 409	\$ 323	\$ 396	\$ 105	\$ 34	\$
	1998	\$ 247	\$ 422	\$ 441	\$ 380	\$ 221	\$ 592	\$ 290	\$ 198	\$ 19	\$
Travel	1997	\$ 624	\$ 505	\$ 564	\$ 386	\$ 300	\$ 978	\$ 416	\$ 48	\$ 38	
	1998	\$ 608	\$ 559	\$ 1.096	\$ 611	\$ 464	\$ 316	\$ 573	\$ 257	\$ 198	\$

pivoting

Category	Year	Metrics									
		Customer Region									
		Dollar Sales									
North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germany		
Electronics	1997	\$ 138	\$ 1.184	\$ 1.774	\$ 4.529	\$ 384	\$ 1.892	\$ 138	\$ 7.232	\$ 2.346	\$ 651
	1998	\$ 538	\$ 682	\$ 925	\$ 729	\$ 959	\$ 262	\$ 677	\$ 588	\$ 213	\$ 469
Food	1997	\$ 759	\$ 538	\$ 682	\$ 925	\$ 729	\$ 959	\$ 262	\$ 677	\$ 588	\$ 213
	1998	\$ 538	\$ 682	\$ 925	\$ 729	\$ 959	\$ 262	\$ 677	\$ 588	\$ 213	\$ 469
Gifts	1997	\$ 2.532	\$ 1.955	\$ 1.355	\$ 2.785	\$ 1.854	\$ 2.800	\$ 1.413	\$ 2.695	\$ 2.535	\$ 1.813
	1998	\$ 1.955	\$ 2.785	\$ 1.854	\$ 2.800	\$ 1.413	\$ 2.695	\$ 2.535	\$ 1.813	\$ 2.132	\$ 2.844
Health & Beauty	1997	\$ 624	\$ 611	\$ 640	\$ 887	\$ 1.317	\$ 566	\$ 647	\$ 382	\$ 588	\$ 754
	1998	\$ 611	\$ 640	\$ 887	\$ 1.317	\$ 566	\$ 647	\$ 382	\$ 588	\$ 754	\$ 1.162
Household	1997	\$ 5.354	\$ 5.787	\$ 4.112	\$ 5.320	\$ 5.410	\$ 5.416	\$ 4.446	\$ 6.812	\$ 3.058	\$ 4.334
	1998	\$ 5.787	\$ 4.112	\$ 5.320	\$ 5.410	\$ 5.416	\$ 4.446	\$ 6.812	\$ 3.058	\$ 4.334	\$ 3.974
Kid's Korner	1997	\$ 201	\$ 247	\$ 398	\$ 422	\$ 485	\$ 441	\$ 186	\$ 380	\$ 409	\$ 221
	1998	\$ 247	\$ 398	\$ 422	\$ 485	\$ 441	\$ 186	\$ 380	\$ 409	\$ 221	\$ 323
Travel	1997	\$ 624	\$ 508	\$ 559	\$ 564	\$ 1.096	\$ 386	\$ 611	\$ 300	\$ 464	\$ 978
	1998	\$ 508	\$ 559	\$ 564	\$ 1.096	\$ 611	\$ 464	\$ 316	\$ 300	\$ 464	\$ 316

61

OLAP: operatori



drill-across

62

OLAP: operatori

Category	Metrics	Dollar Sales							
		Quarter	Q1 1997	Q2 1997	Q3 1997	Q4 1997	Q1 1998	Q2 1998	Q3 1998
Electronics	\$ 4.383	\$ 817	\$ 827	\$ 4.589	\$ 13.770	\$ 2.977	\$ 4.226	\$ 8.326	
Food	\$ 1.546	\$ 1.310	\$ 1.268	\$ 1.176	\$ 2.676	\$ 1.120	\$ 953	\$ 889	
Gifts	\$ 3.398	\$ 3.893	\$ 4.682	\$ 4.342	\$ 7.879	\$ 4.145	\$ 4.378	\$ 3.645	
Health & Beauty	\$ 1.826	\$ 878	\$ 1.904	\$ 1.434	\$ 2.156	\$ 898	\$ 1.207	\$ 1.404	
Household	\$ 9.314	\$ 8.124	\$ 9.331	\$ 11.614	\$ 17.453	\$ 7.604	\$ 12.898	\$ 12.436	
Kid's Korner	\$ 685	\$ 531	\$ 811	\$ 532	\$ 1.084	\$ 491	\$ 532	\$ 836	
Travel	\$ 603	\$ 1.293	\$ 1.456	\$ 1.145	\$ 1.507	\$ 719	\$ 840	\$ 1.726	

drill-across



Category	Metrics	Quarter		Q1 1997		Q2 1997		Q3 1997		Q4 1997		Q1 1998		Q2 1998		Q3 1998		Q4 1998	
		Discount	Dollar Sales	Discount	Dollar Sales	Discount	Dollar Sales	Discount	Dollar Sales	Discount	Dollar Sales	Discount	Dollar Sales	Discount	Dollar Sales	Discount	Dollar Sales	Discount	Dollar Sales
Electronics	\$ 0	\$ 4.383	\$ 0	\$ 817	\$ 0	\$ 827	\$ 300	\$ 4.589	\$ 15	\$ 13.770	\$ 0	\$ 2.977							
Food	\$ 25	\$ 1.546	\$ 0	\$ 1.310	\$ 0	\$ 1.268	\$ 38	\$ 1.176	\$ 0	\$ 2.676	\$ 0	\$ 1.120							
Gifts	\$ 31	\$ 3.398	\$ 0	\$ 3.893	\$ 5	\$ 4.682	\$ 0	\$ 4.342	\$ 15	\$ 7.879	\$ 0	\$ 4.145							
Health & Beauty	\$ 0	\$ 1.826	\$ 0	\$ 878	\$ 0	\$ 1.904	\$ 0	\$ 1.434	\$ 229	\$ 2.156	\$ 0	\$ 898							
Household	\$ 0	\$ 9.314	\$ 228	\$ 8.124	\$ 175	\$ 9.331	\$ 35	\$ 11.614	\$ 5	\$ 17.453	\$ 211	\$ 7.604							
Kid's Korner	\$ 0	\$ 685	\$ 0	\$ 531	\$ 32	\$ 811	\$ 40	\$ 532	\$ 0	\$ 1.084	\$ 0	\$ 491							
Travel	\$ 0	\$ 603	\$ 0	\$ 1.293	\$ 200	\$ 1.456	\$ 0	\$ 1.145	\$ 0	\$ 1.507	\$ 0	\$ 719							

63

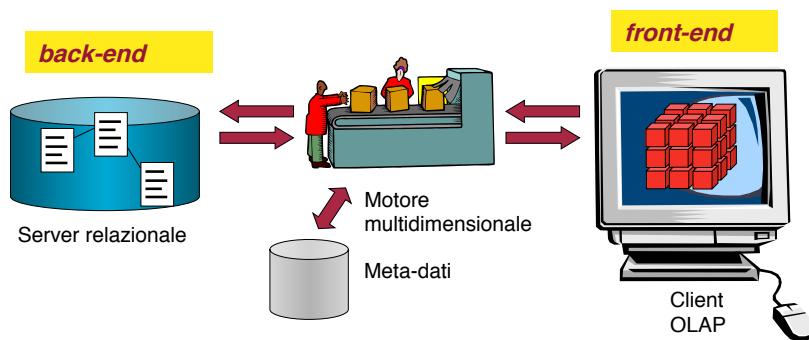
Reportistica semi-statica

- In molti contesti applicativi, è utile un approccio intermedio tra reportistica statica e OLAP: la *reportistica semi-statica*
 - ✓ Un rapporto semi-statico, pur essendo focalizzato su un'insieme di informazioni predefinite, permette all'analista alcuni gradi di libertà, che si concretizzano nella possibilità di eseguire un ristretto insieme di percorsi di navigazione
- Vantaggi:
 - ✓ agli utenti è richiesta una minor competenza sul modello dei dati e sullo strumento di analisi rispetto al caso dell'OLAP
 - ✓ si elimina il rischio di creare risultati d'analisi inconsistenti o scorretti a causa dell'uso improprio dei meccanismi di aggregazione
 - ✓ vincolando i tipi di analisi permessi si evita che l'utente possa involontariamente rallentare il sistema formulando interrogazioni eccessivamente pesanti

64

ROLAP (Relational OLAP)

- Giustificato dall'enorme lavoro svolto in letteratura sul modello relazionale, dalla diffusa esperienza aziendale sull'utilizzo e l'amministrazione di basi di dati relazionali e dall'elevato livello di prestazioni e flessibilità raggiunto dai DBMS relazionali
 - ✓ I dati sono memorizzati su un DBMS relazionale, in forma dettagliata e pre-aggregata
 - ✓ Occorre elaborare tipologie specifiche di schemi che permettano di traslare il modello multidimensionale sul modello relazionale: *schema a stella*
 - ✓ Il problema delle prestazioni porta a *denormalizzazione* per evitare costosi join



65

MOLAP (Multidimensional OLAP)

- Basato su un modello logico ad hoc sul quale i dati e le operazioni multidimensionali possono essere direttamente rappresentati
- I dati vengono fisicamente memorizzati in vettori e l'accesso è di tipo posizionale
 - ✓ Il grosso vantaggio dell'approccio MOLAP rispetto a quello ROLAP è che le operazioni multidimensionali sono realizzabili in modo semplice e naturale, senza necessità di ricorrere a join; le prestazioni risultano pertanto ottime
 - ✓ Non esistendo ancora uno standard per il modello logico multidimensionale, le diverse implementazioni MOLAP hanno veramente poco in comune: in genere, solo l'utilizzo di tecnologie di ottimizzazione specifiche per trattare il problema della sparsità

66

HOLAP (Hybrid OLAP)

- Sistemi di questo tipo combinano in un'unica architettura elementi di ROLAP e MOLAP
 - ✓ Tipicamente i dati di dettaglio sono memorizzati su DBMS relazionale, i pre-aggregati su strutture multidimensionali proprietarie
 - ✓ Oppure, i sottocubi densi sono memorizzati in forma multidimensionale, quelli sparsi in forma relazionale

67

La qualità



La qualità di un processo misura la sua aderenza agli obiettivi degli utenti

- Fattori che caratterizzano la qualità dei dati in un DW:
 1. **Accuratezza:** la conformità tra il valore memorizzato e quello reale.
 2. **Attualità:** il dato memorizzato non è obsoleto.
 3. **Completezza:** non mancano informazioni.
 4. **Consistenza:** la rappresentazione dei dati è uniforme.
 5. **Disponibilità:** i dati sono facilmente disponibili all'utente.
 6. **Tracciabilità:** è possibile risalire alla fonte di ciascun dato.
 7. **Chiarezza:** i dati sono facilmente interpretabili

68



La qualità

- Un ruolo basilare nel raggiungimento degli obiettivi di qualità dei dati è di pertinenza dell'organizzazione aziendale, e potrà essere efficacemente svolto solo mettendo a punto un adeguato e puntuale meccanismo di **certificazione** che individui un ristretto insieme di utenti cui affidare la responsabilità dei dati
- È pertanto un preciso dovere del progettista sensibilizzare i vertici aziendali sull'importanza dell'argomento, e stimolarli affinché mettano a punto un corretto iter di certificazione, opportunamente differenziato per aree aziendali

69

La sicurezza



- La sicurezza dell'informazione è un requisito fondamentale per un sistema, da considerare attentamente nell'ingegneria del software attraverso tutti gli stadi del ciclo di sviluppo
- Il problema della sicurezza è ancora più sentito per i DW, poiché
 - ✓ i DW gestiscono informazione cruciale per il processo decisionale strategico
 - ✓ la multidimensionalità e l'aggregazione introducono addizionali problemi di sicurezza poiché implicitamente consentono inferenze indesiderate sui dati
 - ✓ l'elevata mole di comunicazione che ha luogo nei DW durante l'alimentazione crea specifici problemi relativi alla sicurezza di rete

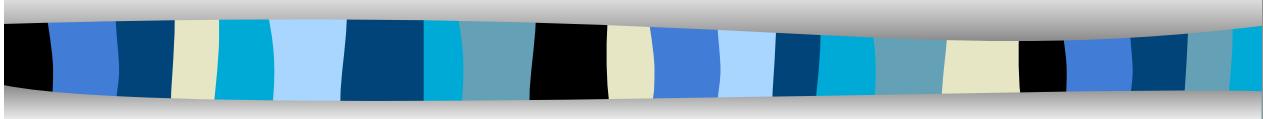
70

La sicurezza



- Controllo delle autorizzazioni
 - ✓ si svolge all'interno degli strumenti di front-end oppure utilizza i servizi messi a disposizione dai sistemi operativi
- Auditing
 - ✓ le tecniche fornite dai DBMS server non sono in genere sufficienti, e occorre appoggiarsi ai meccanismi implementati dai motori OLAP
- Accesso ai dati per profilo utente
 - ✓ i requisiti di base sono relativi alla mascheratura di interi cubi, di certe loro “fette”, di certe loro misure
 - ✓ in molti casi è necessario nascondere i dati di un cubo oltre un livello di dettaglio assegnato

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

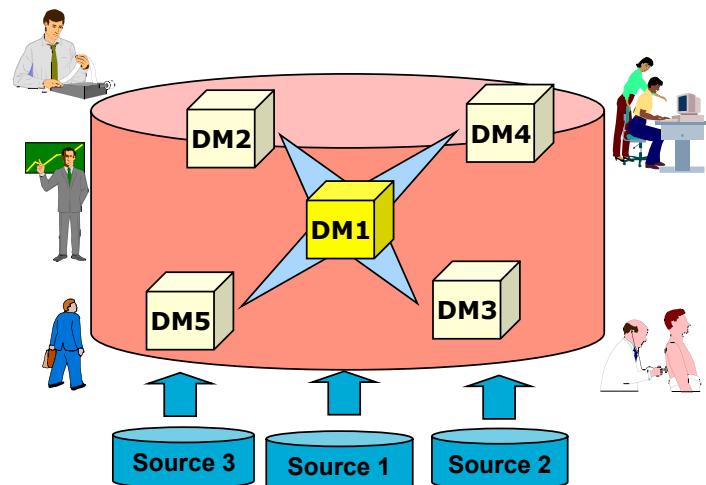
Approccio bottom-up

- 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

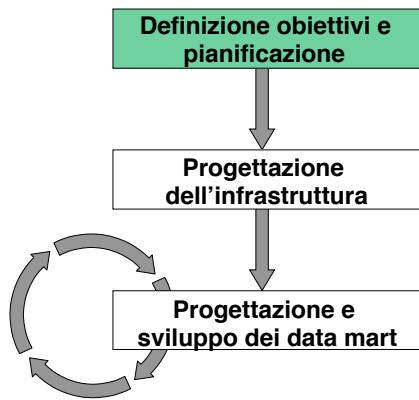
Il primo data mart da prototipare...

- ✓ 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

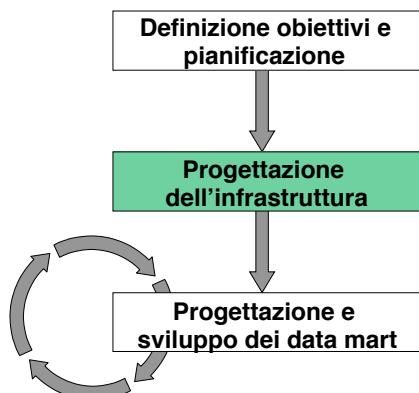
Il ciclo di sviluppo



- 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

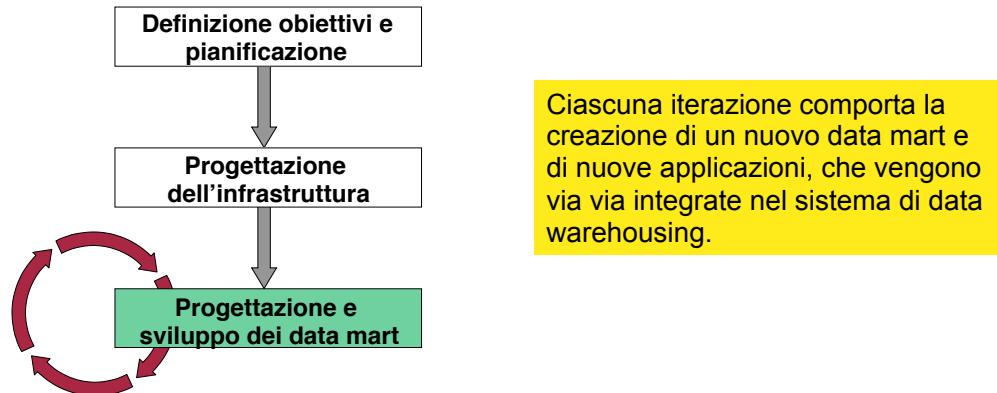
Il ciclo di sviluppo



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.

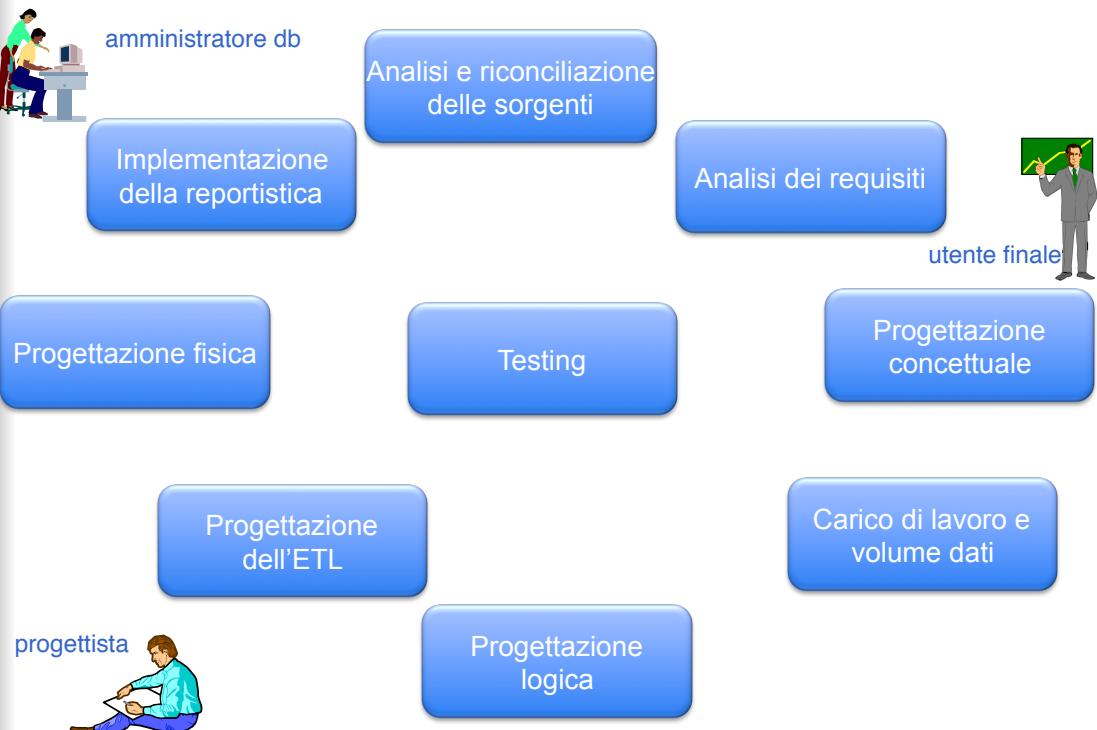
8

Il ciclo di sviluppo

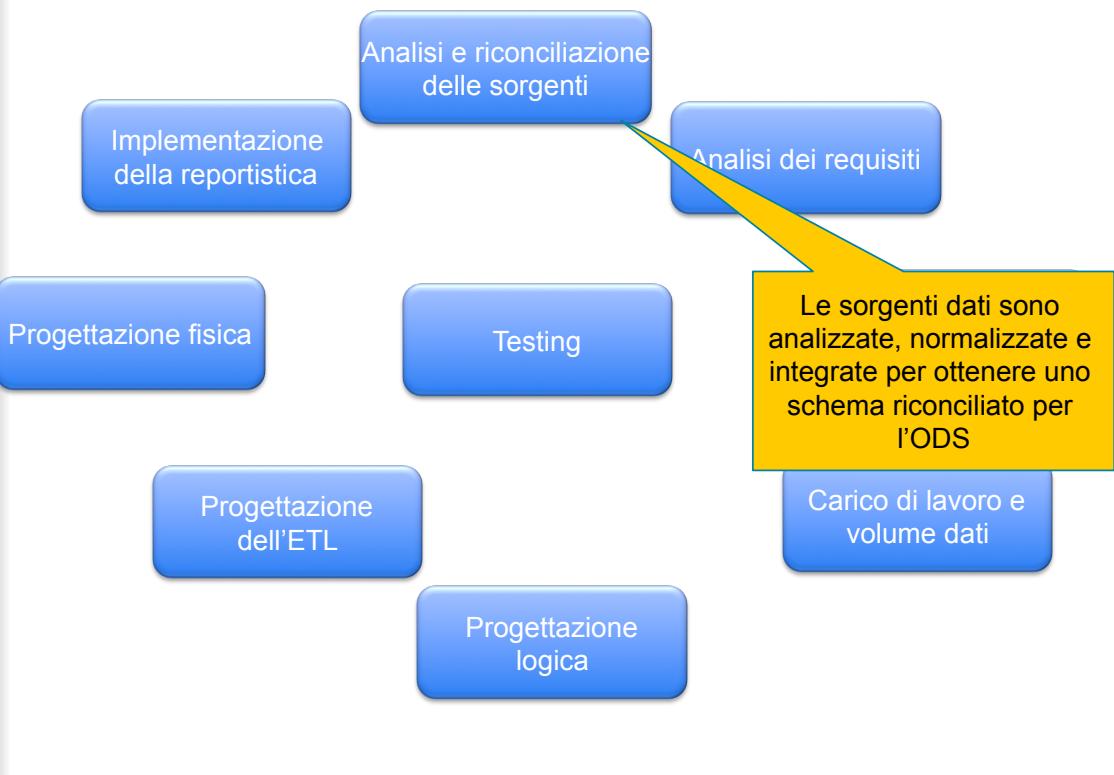


9

La progettazione di data mart



La progettazione di data mart



La progettazione di data mart



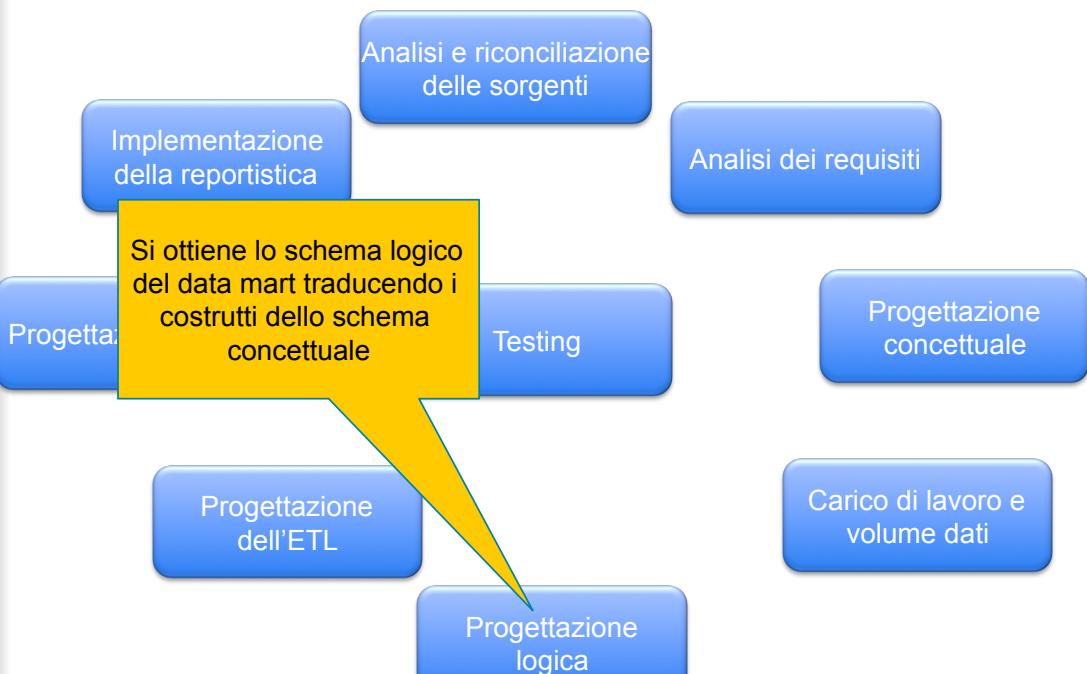
La progettazione di data mart



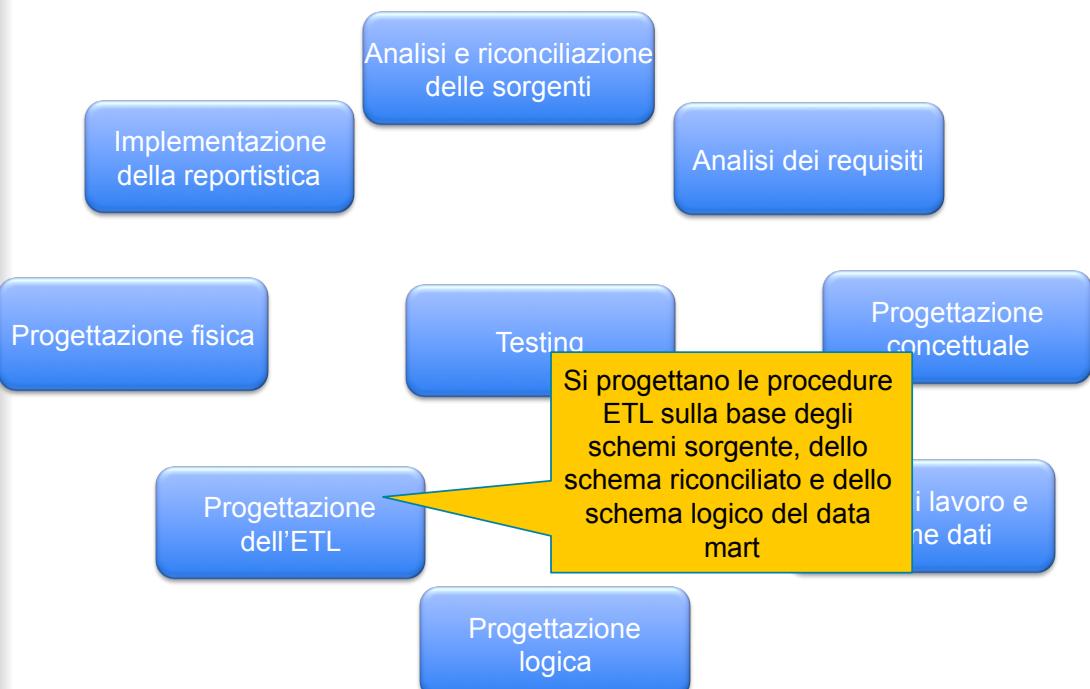
La progettazione di data mart



La progettazione di data mart



La progettazione di data mart



La progettazione di data mart



La progettazione di data mart



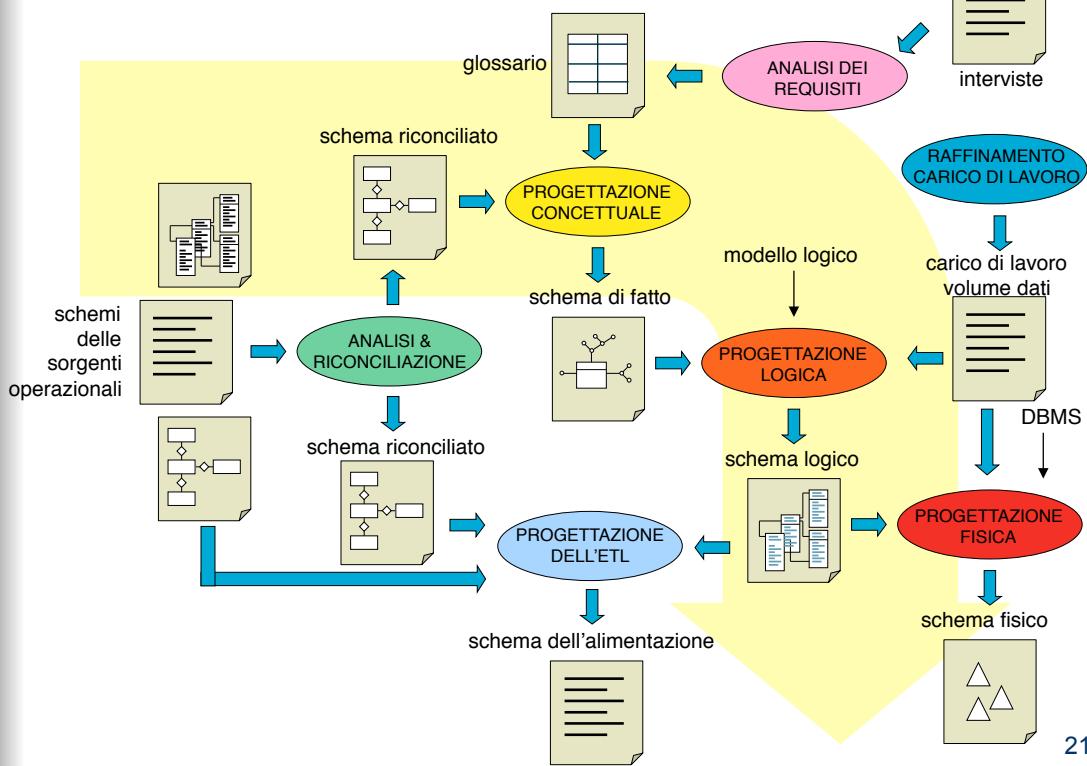
La progettazione di data mart



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

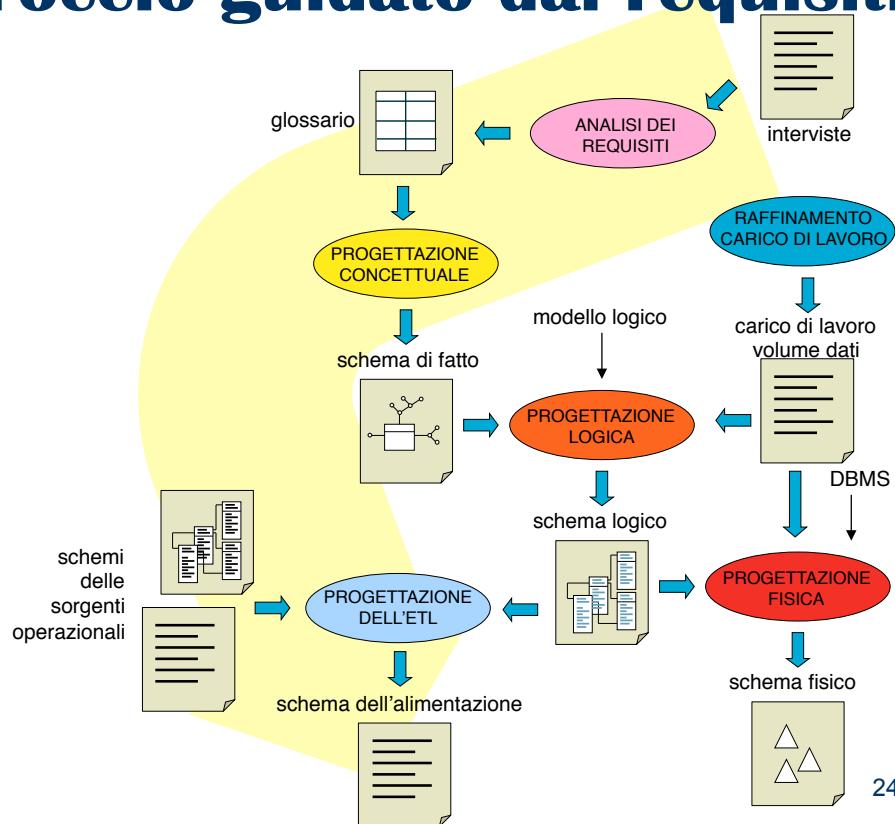
Approccio guidato dai dati

■ 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

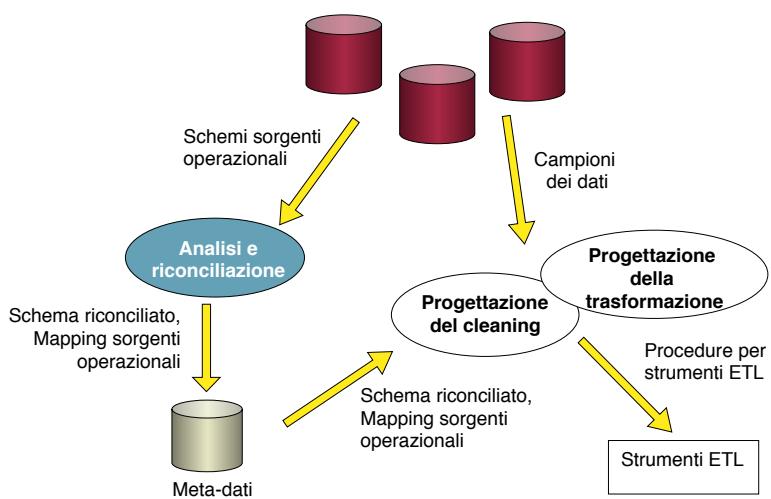
■ Applicabilità

- ✓ 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

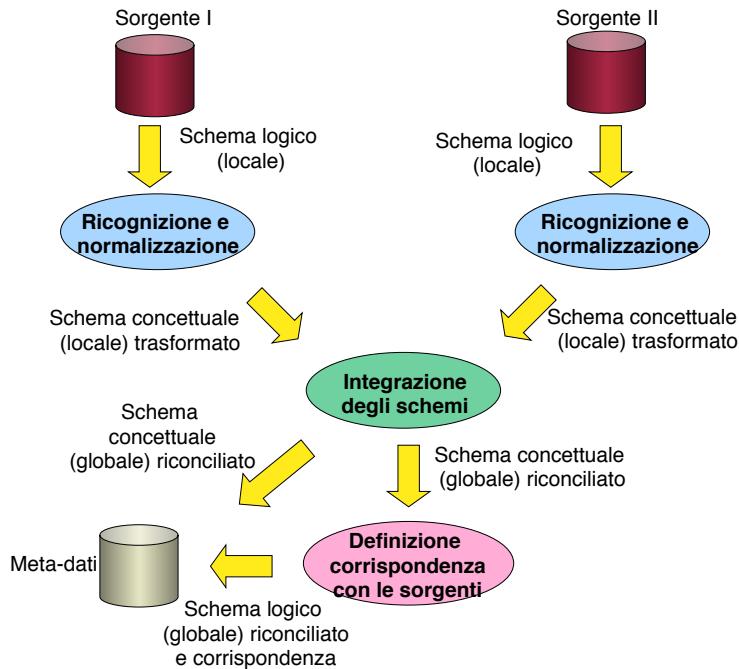
Analisi e riconciliazione delle sorgenti operazionali

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



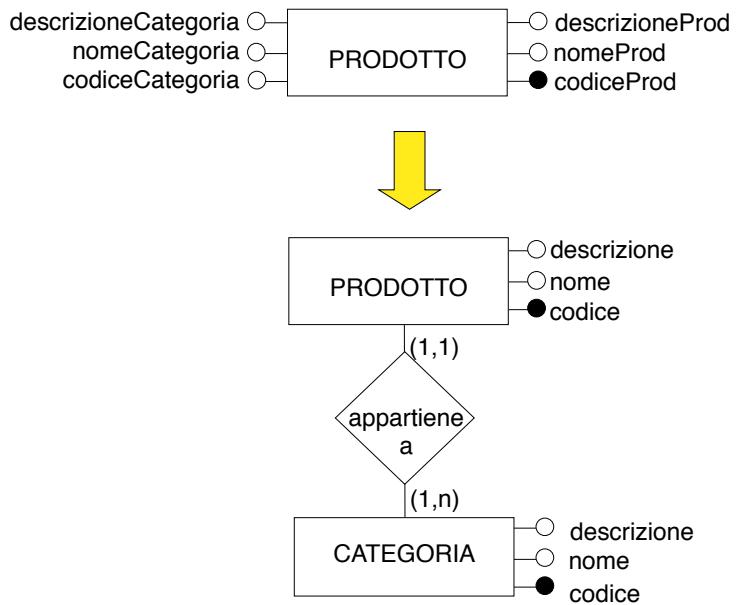
29

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



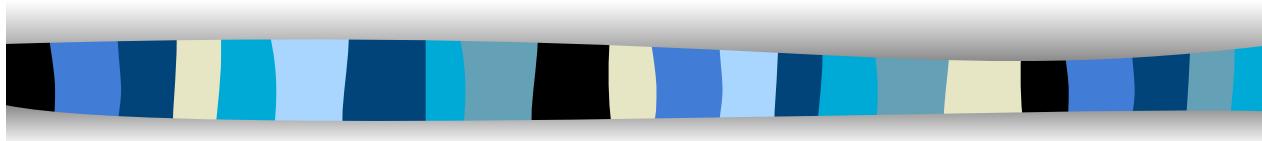
31

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



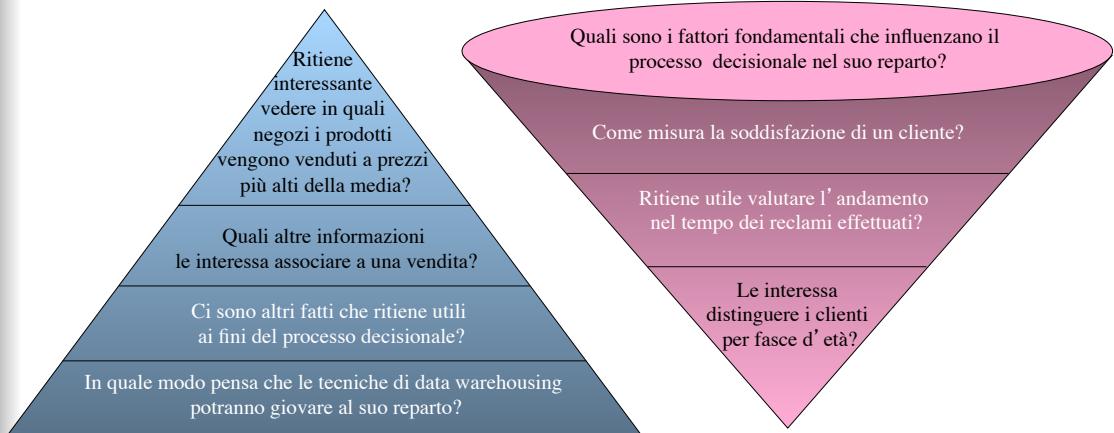
35

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.

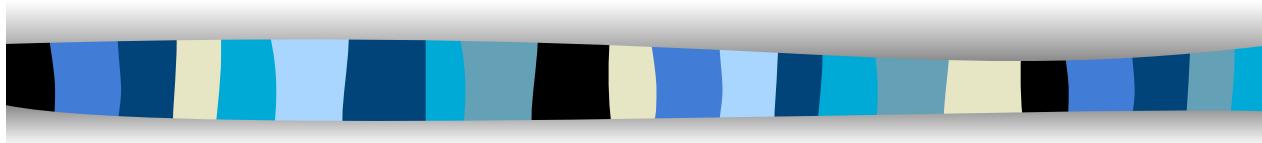
43

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!



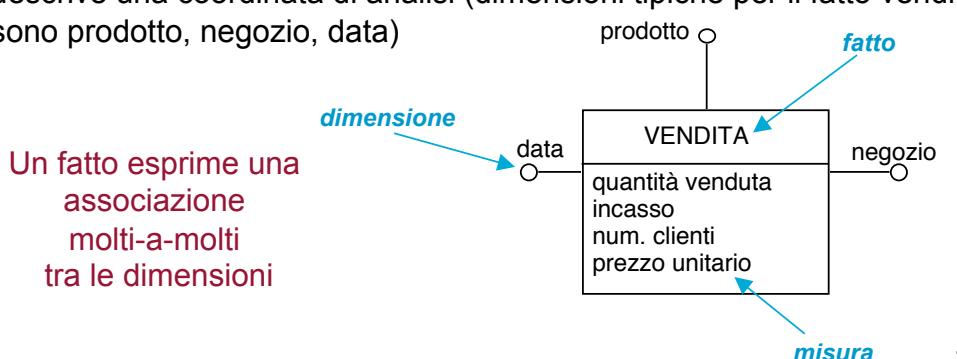
Il Dimensional Fact Model

- 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

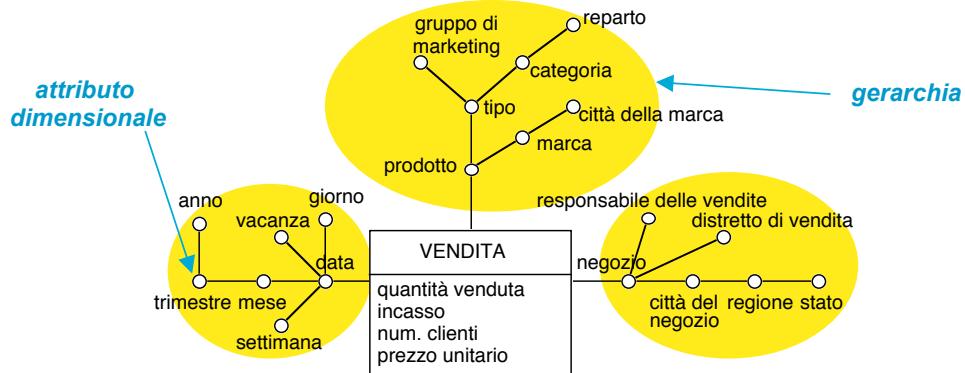
- Un **fatto** è un concetto di interesse per il processo decisionale; tipicamente modella un insieme di eventi che accadono nell'impresa (ad esempio: vendite, spedizioni, acquisti, ...). È essenziale che un fatto abbia aspetti dinamici, ovvero evolva nel tempo
- Una **misura** è una proprietà numerica 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

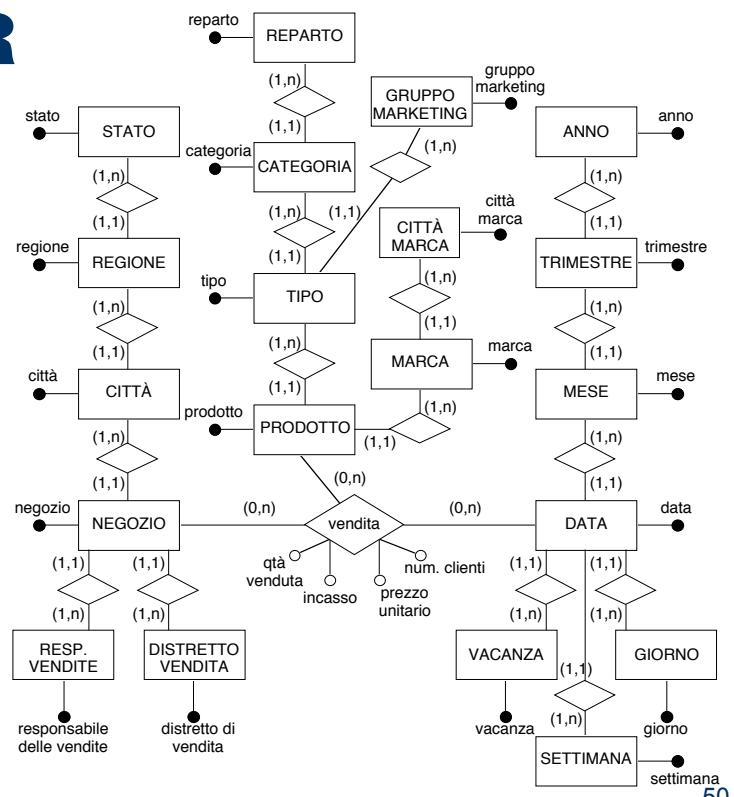
Il DFM: costrutti di base

- 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



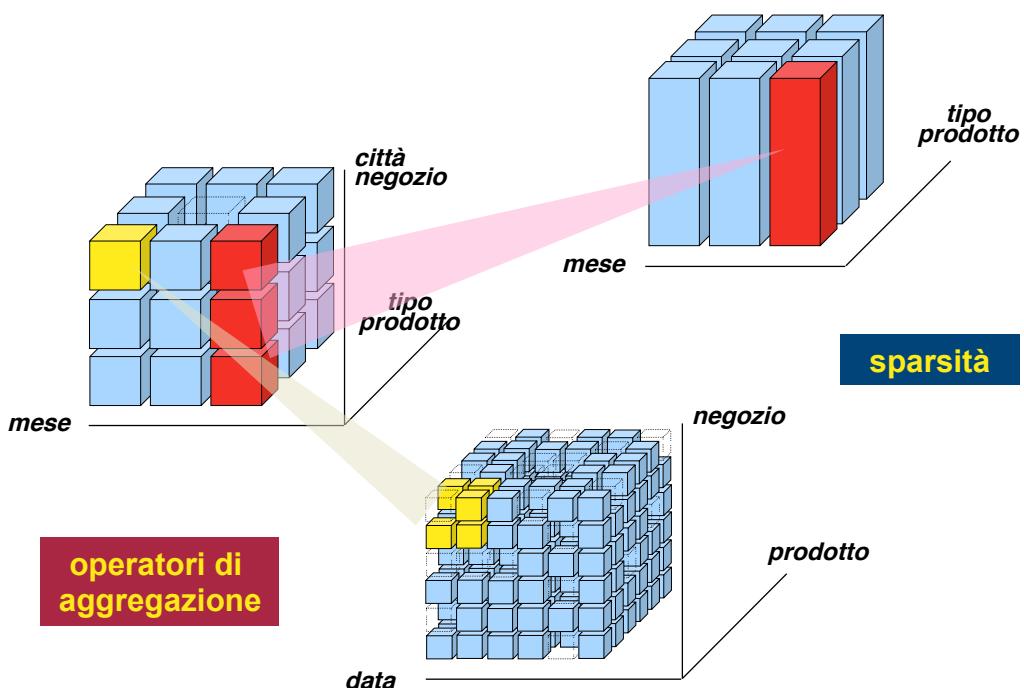
50

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

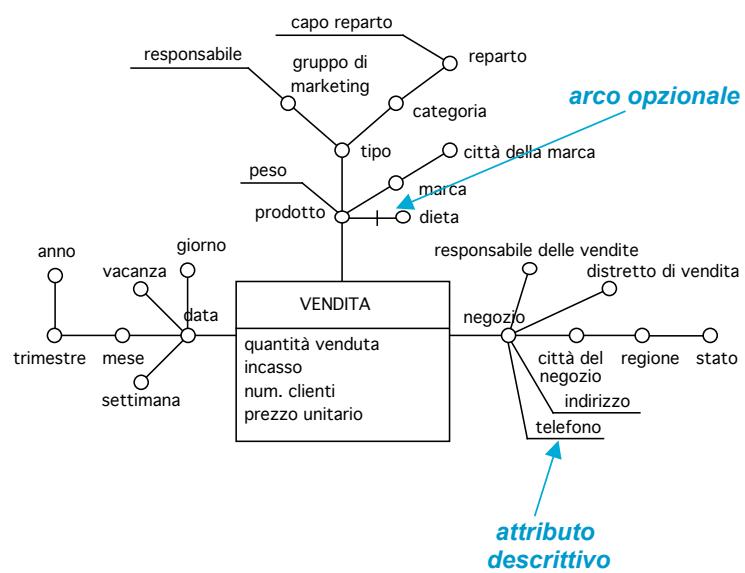
Eventi e aggregazione



52

II DFM: costrutti avanzati

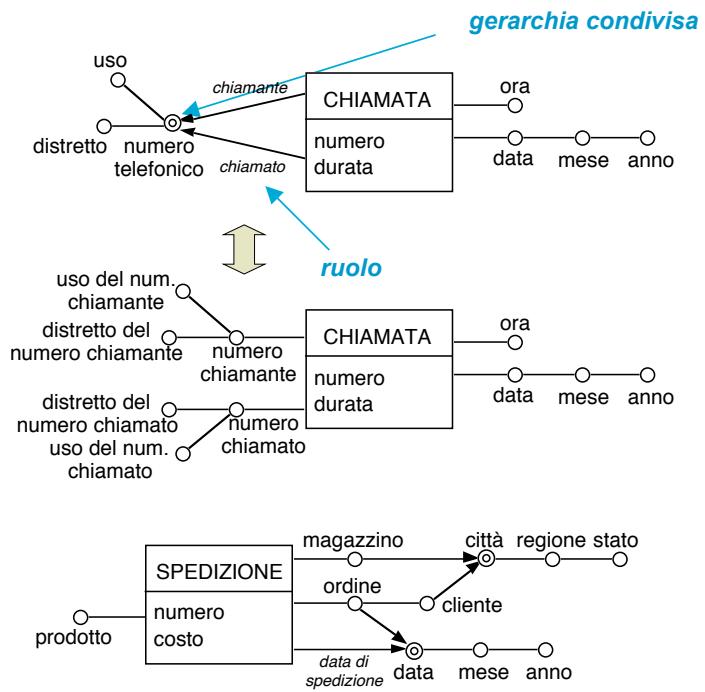
- Un *attributo descrittivo* contiene informazioni aggiuntive su un attributo dimensionale di una gerarchia, a cui è connesso da una associazione -a-uno. Non viene usato per l' aggregazione poiché ha valori continui e/o poiché deriva da un' associazione uno-a-uno
- Alcuni archi dello schema di fatto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



53

II DFM: costrutti avanzati

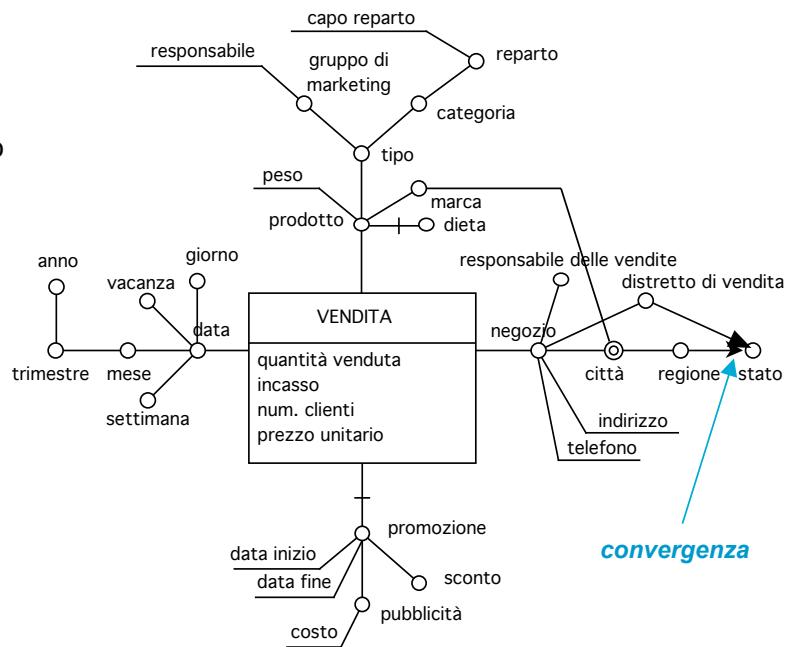
- La *gerarchia condivisa* è un' abbreviazione usata per denotare il fatto che una porzione di gerarchia è replicata più volte nello schema



54

Il DFM: costrutti avanzati

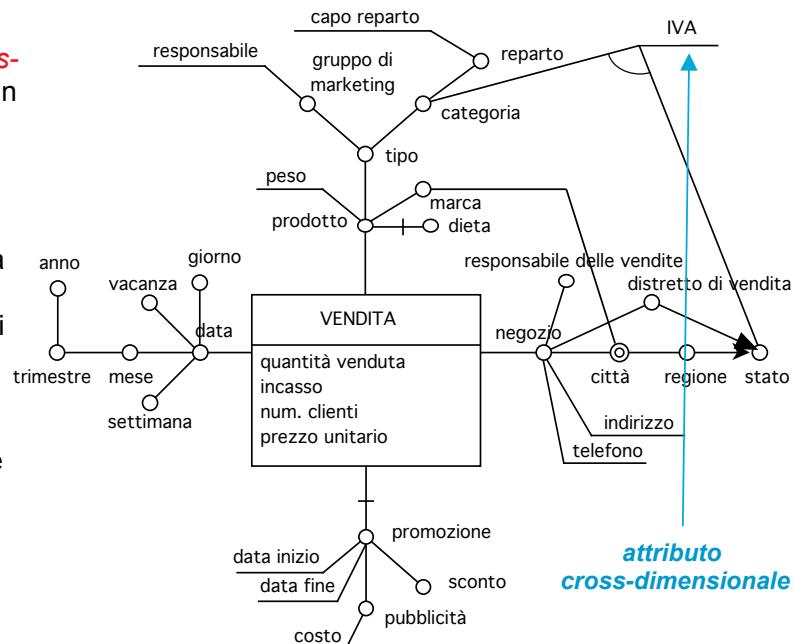
- Due attributi dimensionali possono essere connessi da due o più cammini direzionali distinti, a patto che ciascuno di essi rappresenti ancora una dipendenza funzionale (*convergenza*)



55

Il DFM: costrutti avanzati

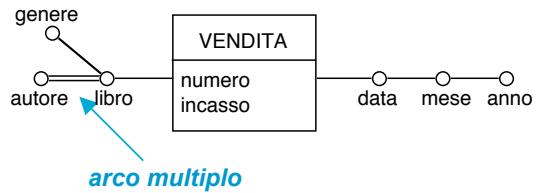
- Un *attributo cross-dimensionale* è un attributo, dimensionale o descrittivo, il cui valore è determinato dalla combinazione di due o più attributi dimensionali, eventualmente appartenenti a gerarchie distinte



56

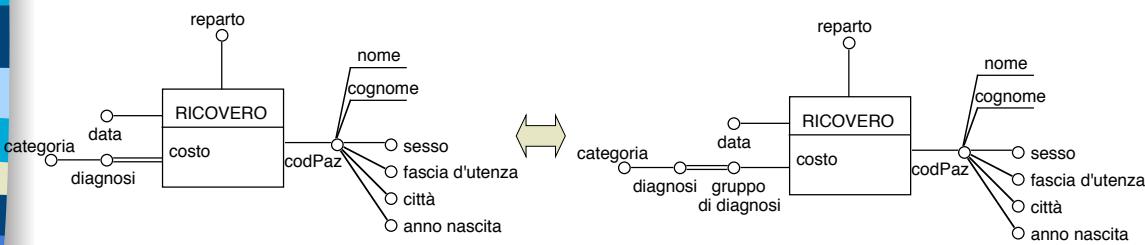
II DFM: costrutti avanzati

- 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

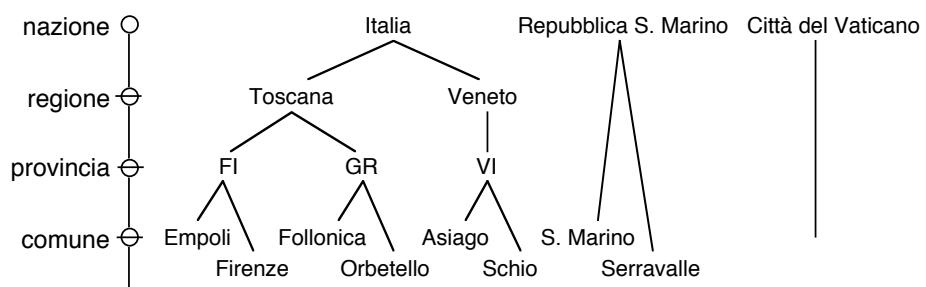
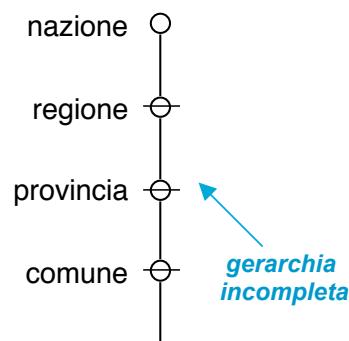
Quanto ha venduto Rizzi?



57

II DFM: costrutti avanzati

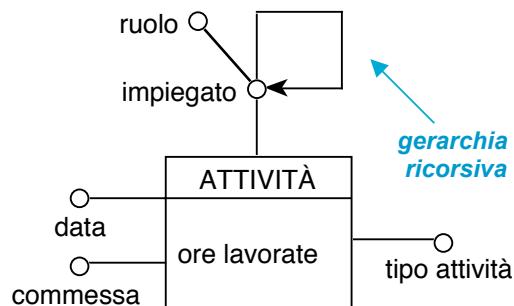
- Una *gerarchia incompleta* è una gerarchia in cui, per alcune istanze, risultano assenti (in quanto non noti oppure non definiti) uno o più livelli di aggregazione



58

II DFM: costrutti avanzati

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

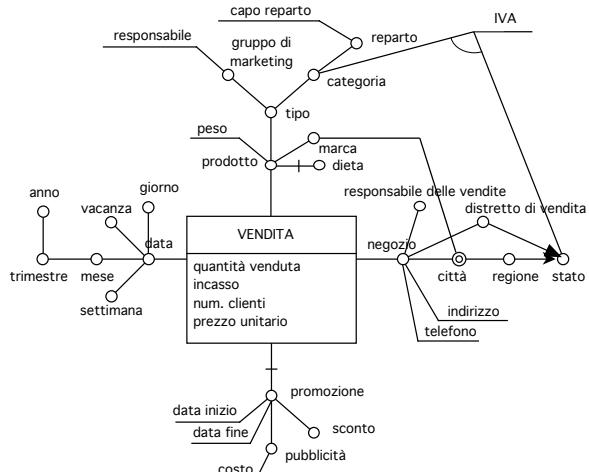


59

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)

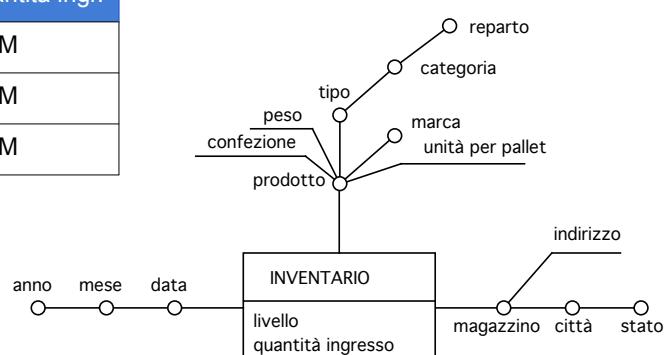
	Gerarchie temporali	Gerarchie non temporali
Misure di flusso	SUM, AVG, MIN, MAX	SUM, AVG, MIN, MAX
Misure di livello	Avg, MIN, MAX	SUM, AVG, MIN, MAX
Misure unitarie	Avg, MIN, MAX	Avg, MIN, MAX

61

Additività

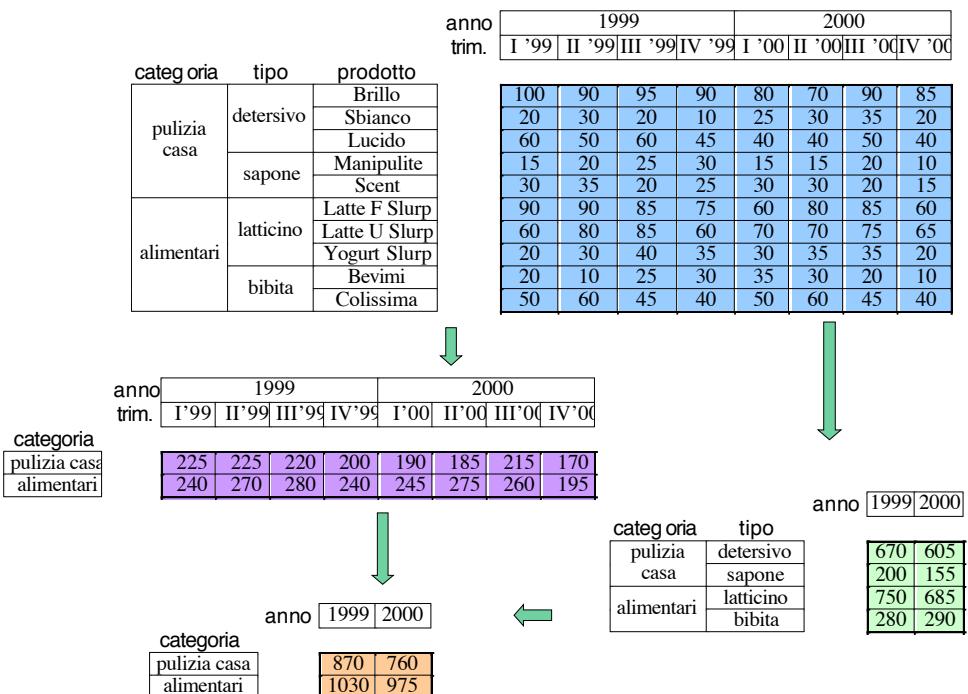
- Una misura è detta **additiva** su una dimensione se i suoi valori possono essere aggregati lungo la corrispondente gerarchia tramite l'operatore di somma, altrimenti è detta **non-additiva**. Una misura non-additiva è **non-aggregabile** se nessun operatore di aggregazione può essere usato su di essa

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



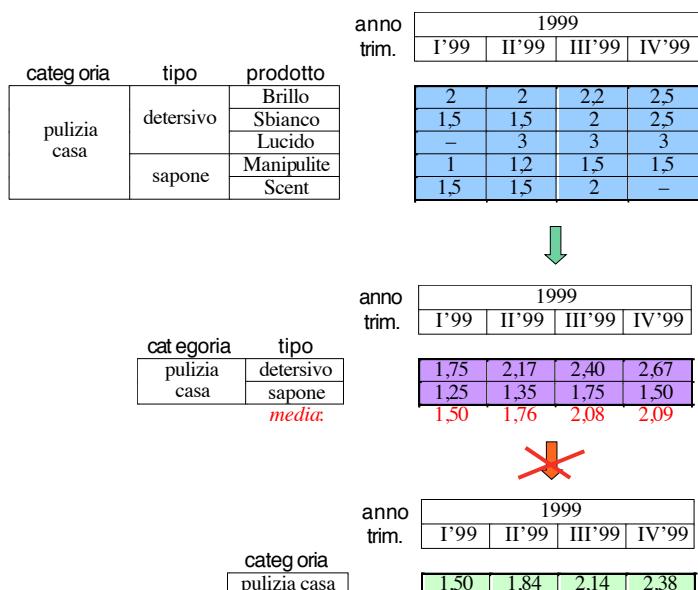
62

Misure additive



63

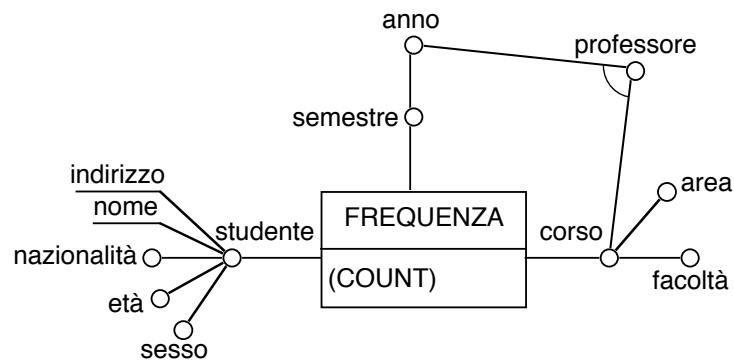
Misure non-additive



64

Schemi di fatto vuoti

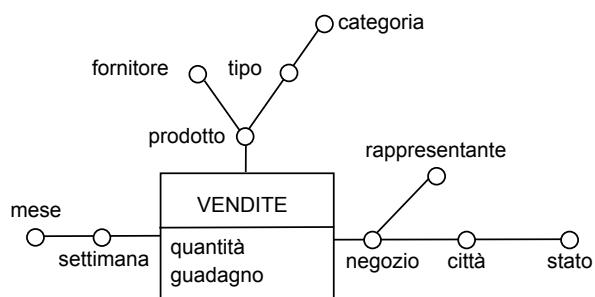
- Uno schema di fatto si dice **vuoto** se non ha misure
 - ✓ In questo caso, il fatto registra solo il verificarsi di un evento



65

Schemi di fatto transazionali

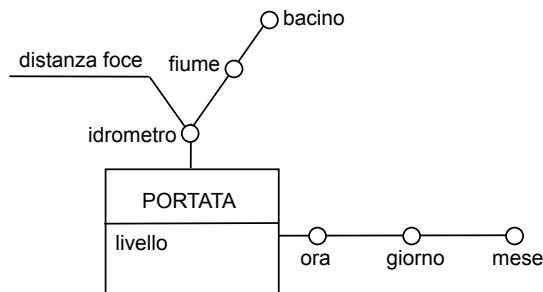
- Ciascun evento registra una singola transazione o riassume un insieme di transazioni che avvengono durante lo stesso intervallo di tempo
 - ✓ La maggior parte delle misure sono di flusso



66

Schemi di fatto istantanei

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



67

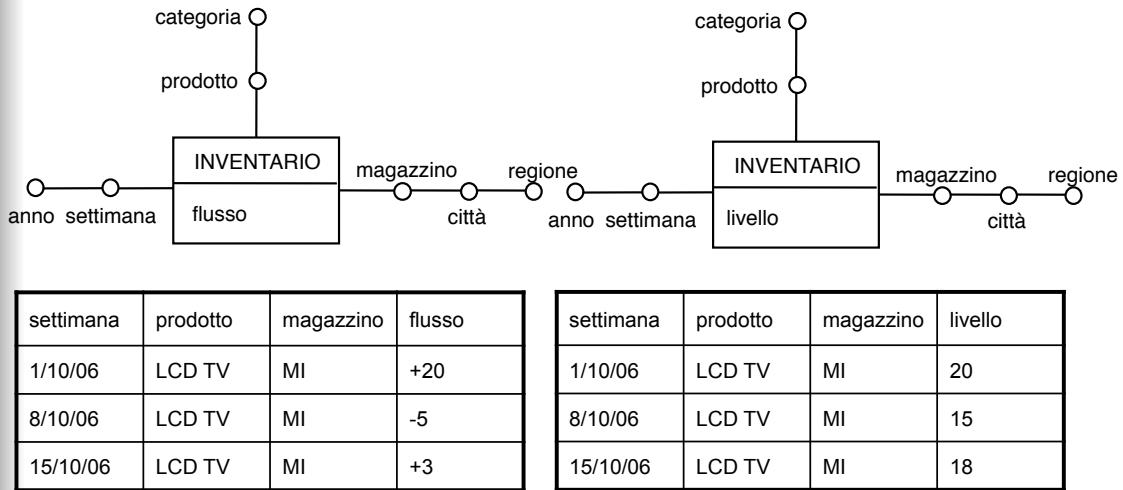
Transazionale vs. istantaneo

- 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:



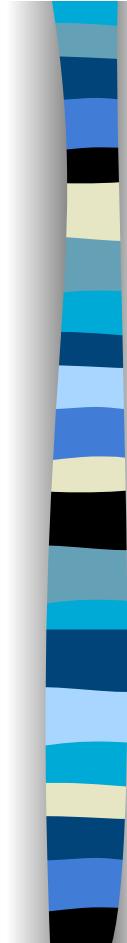
69

Il DFM in azione

Lauree universitarie



70



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

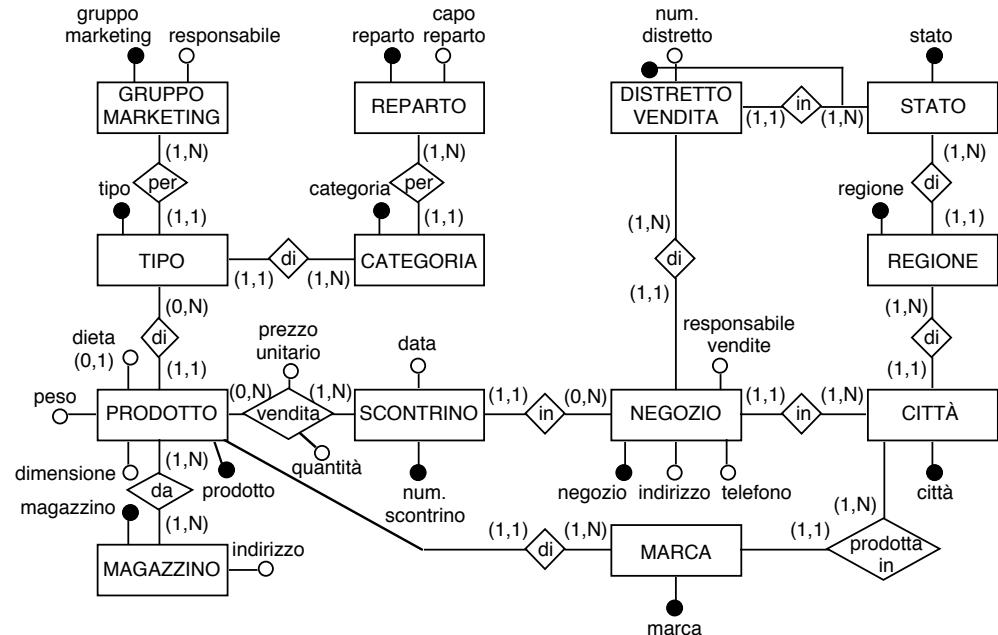


Progettazione concettuale: come

- 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)

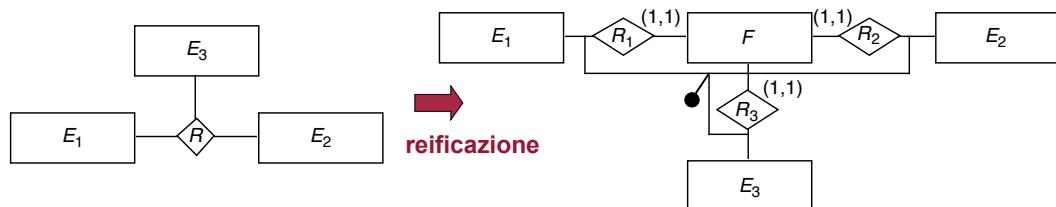
PRODOTTI (prodotto, peso, dimensione, dieta,
diMarca:**MARCHE**, diTipo:**TIPI**)
NEGOZI (negozio, indirizzo, telefono, respVendite,
numDistr, stato):**DISTRETTI**, inCittà:**CITTÀ**)
SCONTRINI (numScontrino, data, negozio:**NEGOZI**)
VENDITE (prodotto:**PRODOTTI**, numScontrino:**SCONTRINI**,
quantità, 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



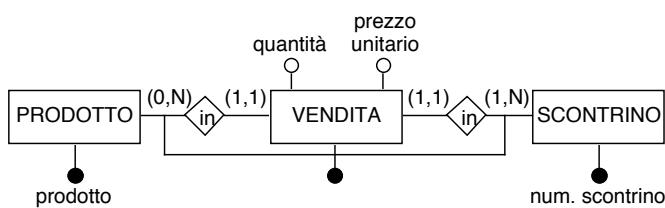
- Sullo schema relazionale un fatto corrisponde a una relazione F

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

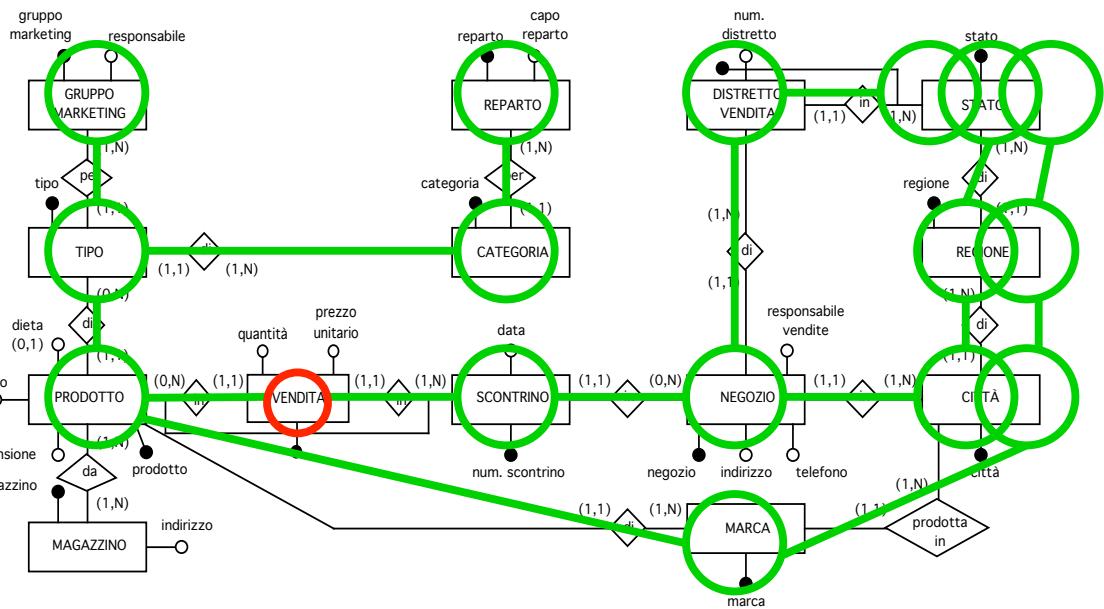
78

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

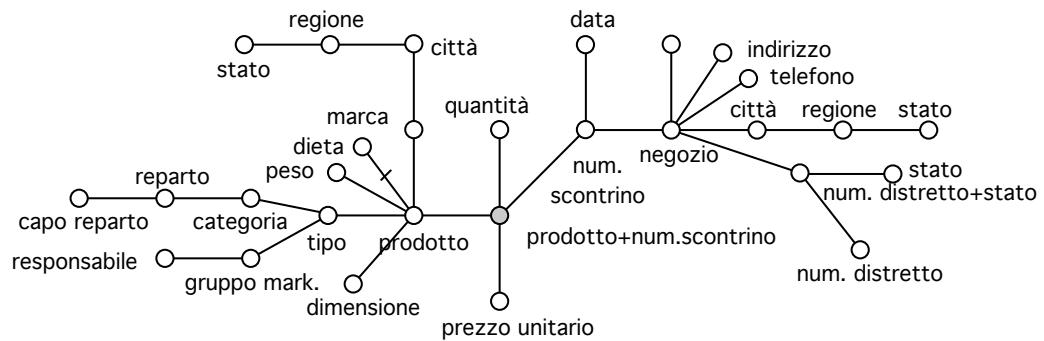
79

L' esempio delle vendite



80

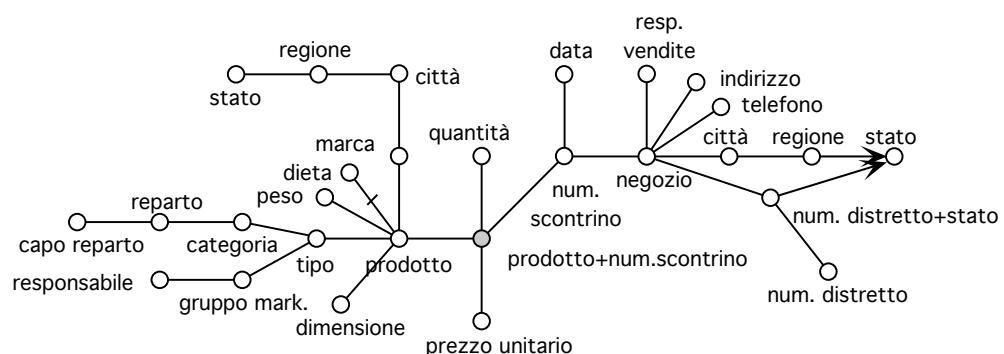
L' esempio delle vendite



81

Problemi

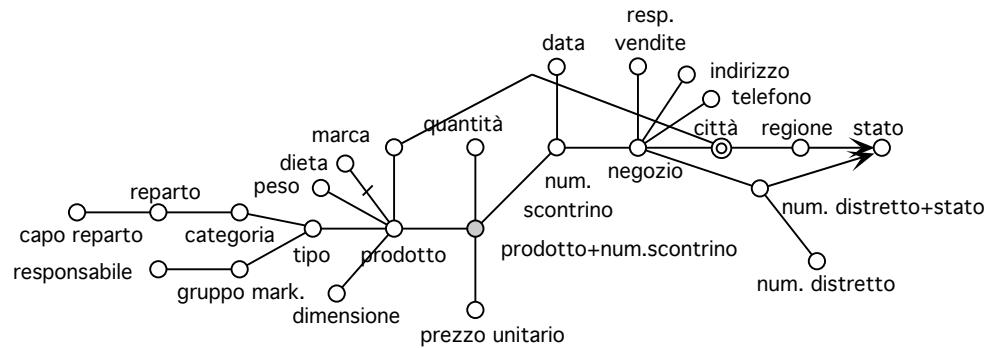
- Stessa entità raggiunta due volte
✓ convergenza



82

Problemi

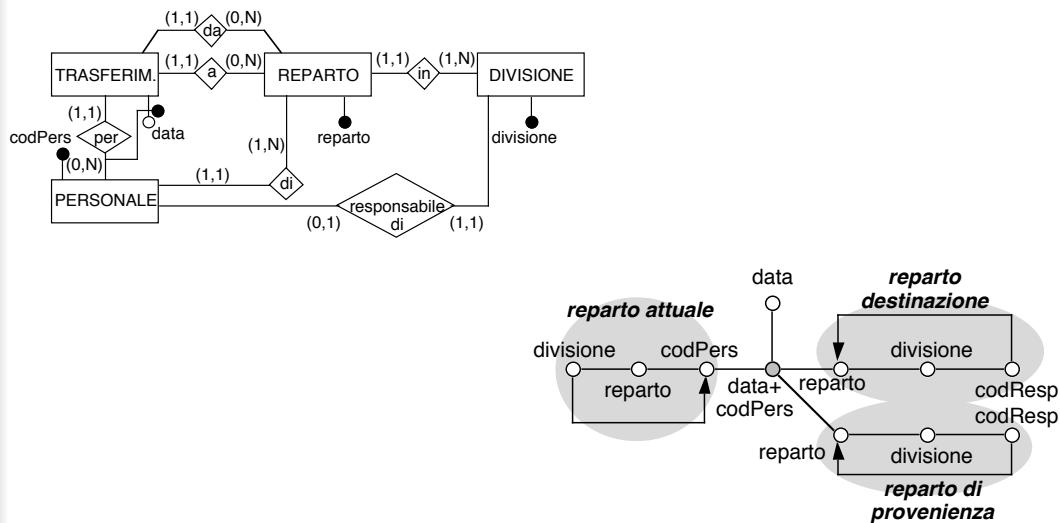
- Stessa entità raggiunta due volte
 - ✓ gerarchia condivisa



83

Problemi

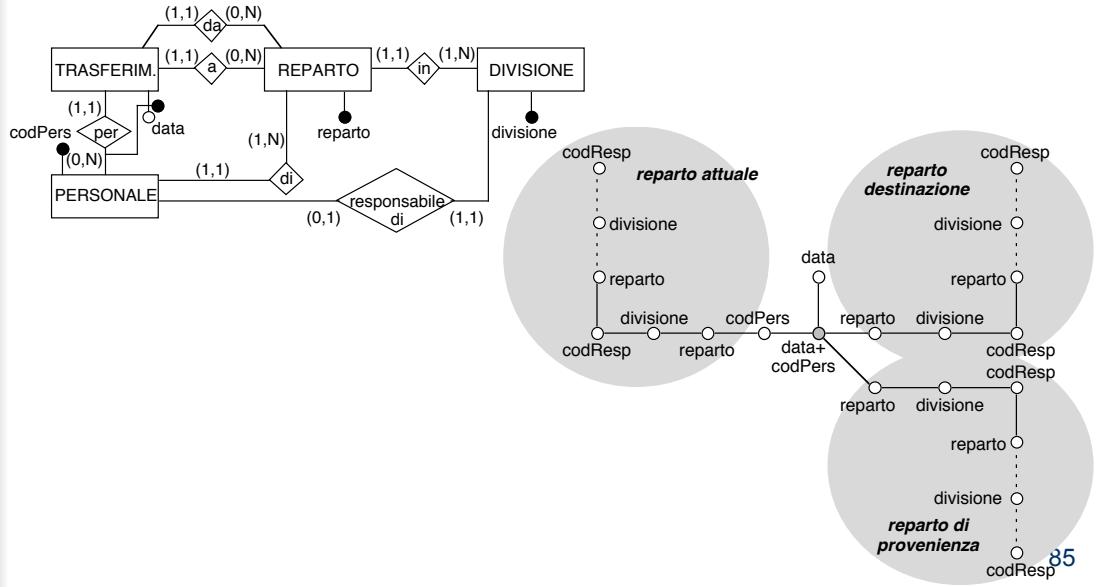
- Cicli di associazioni multi-a-uno
 - ✓ uso di gerarchie ricorsive



84

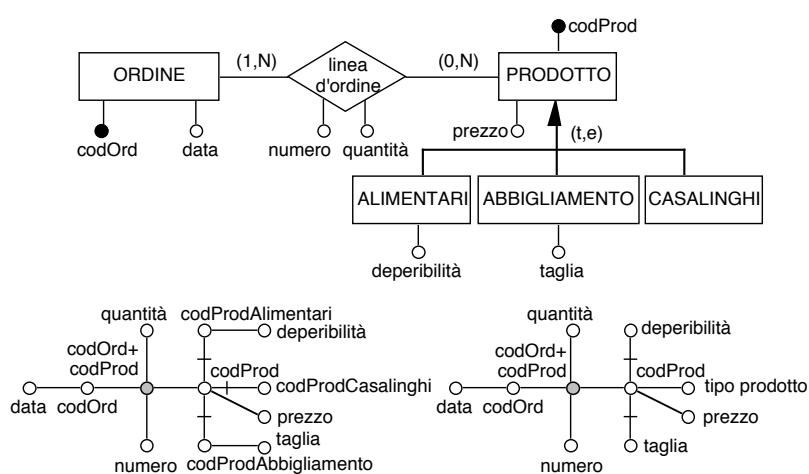
Problemi

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



Problemi

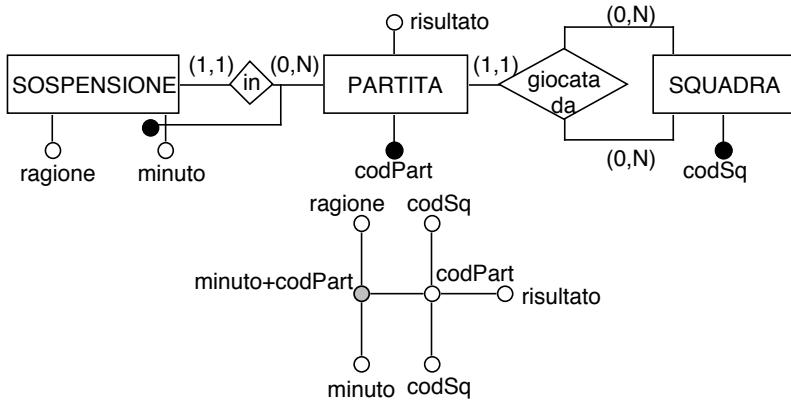
- Gerarchie di specializzazione
 - ✓ equivalenza con associazioni uno-a-uno opzionali



Problemi

■ Associazioni n-arie

✓ percorribili solo le “false n-arie”

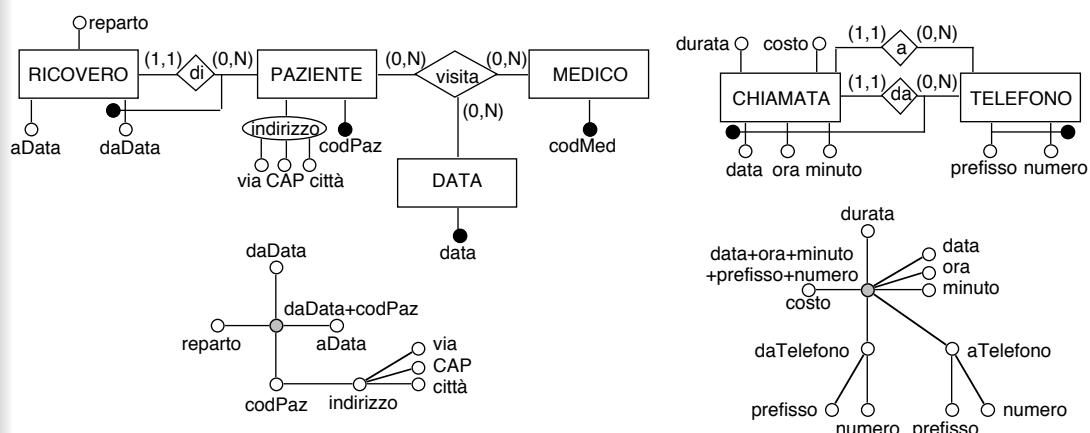


87

Problemi

■ Attributi composti

✓ generano due livelli nell’ albero



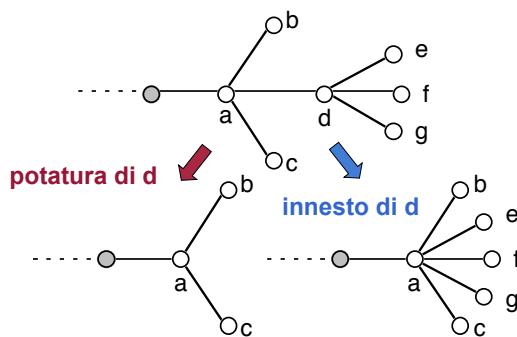
88

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

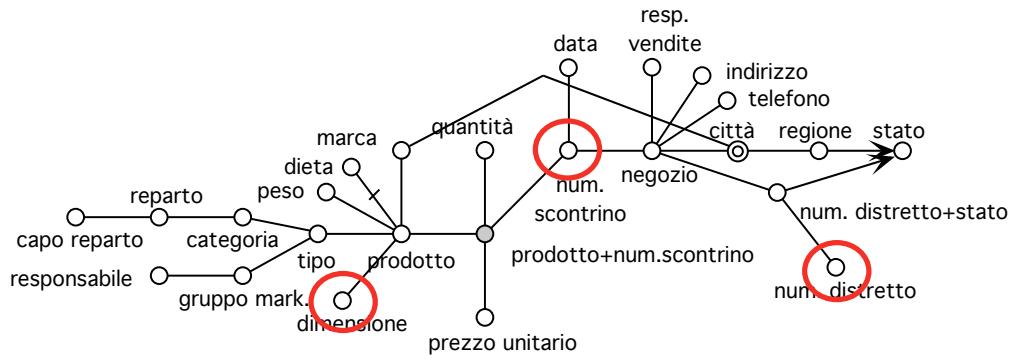
Editing dell' albero



- 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

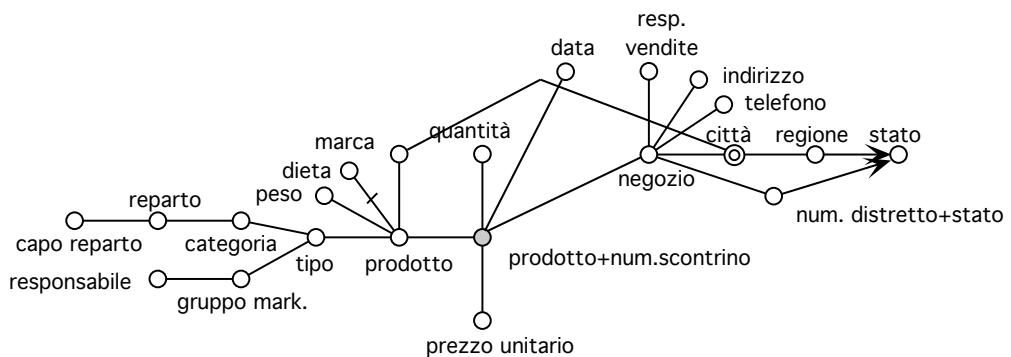
90

L' esempio delle vendite



91

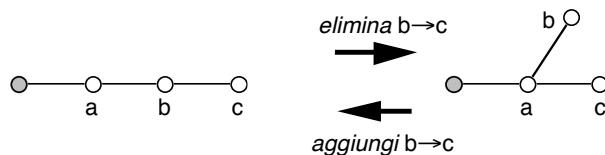
L' esempio delle vendite



92

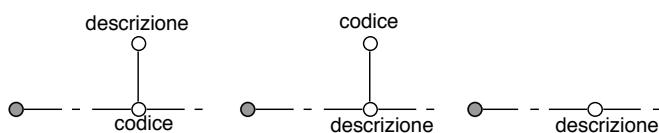
Editing dell' albero

- 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



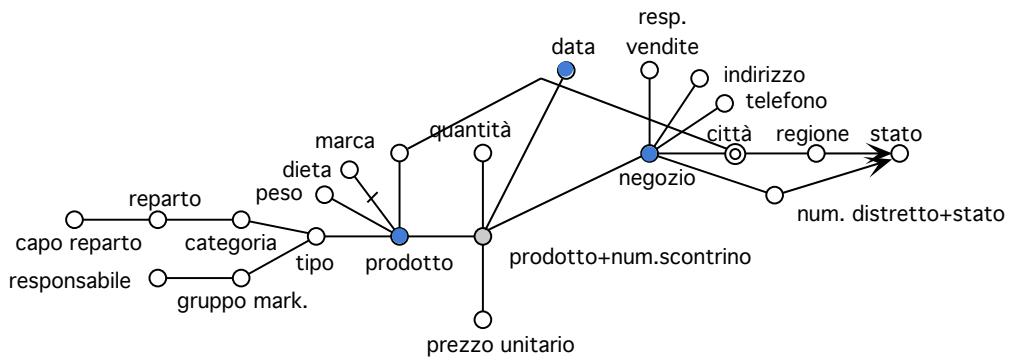
93

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

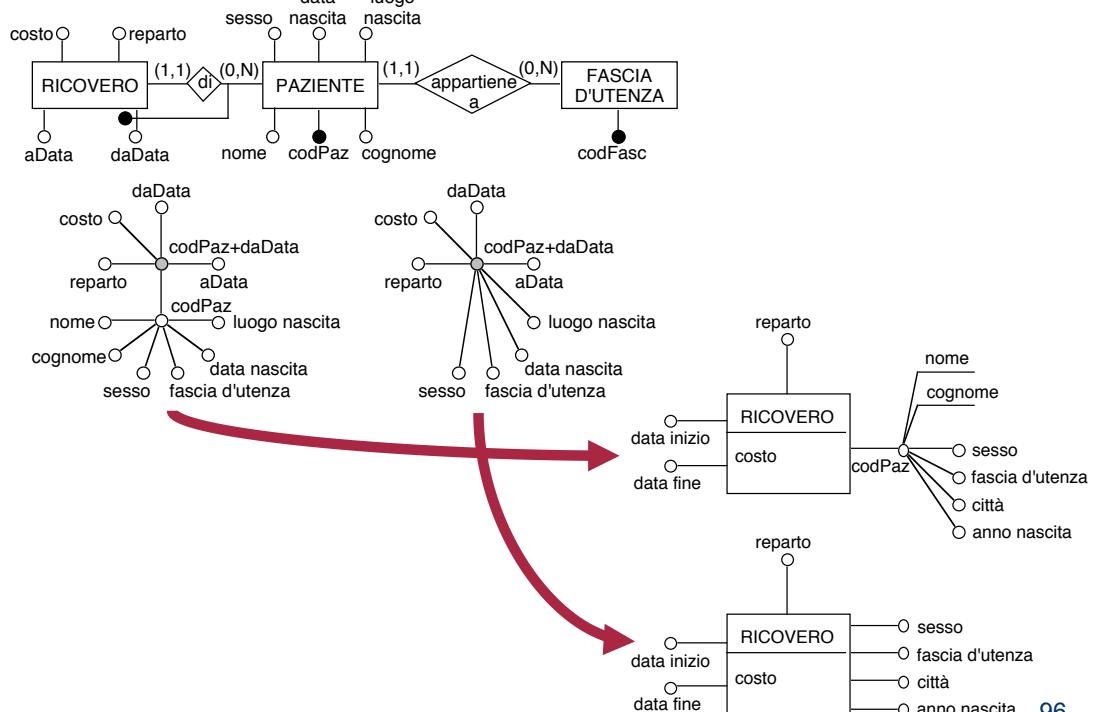
94

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

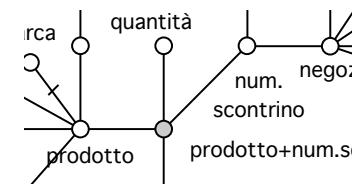
97

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, negozi: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

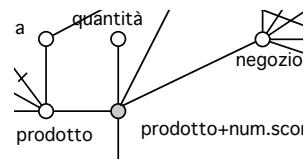
98

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)

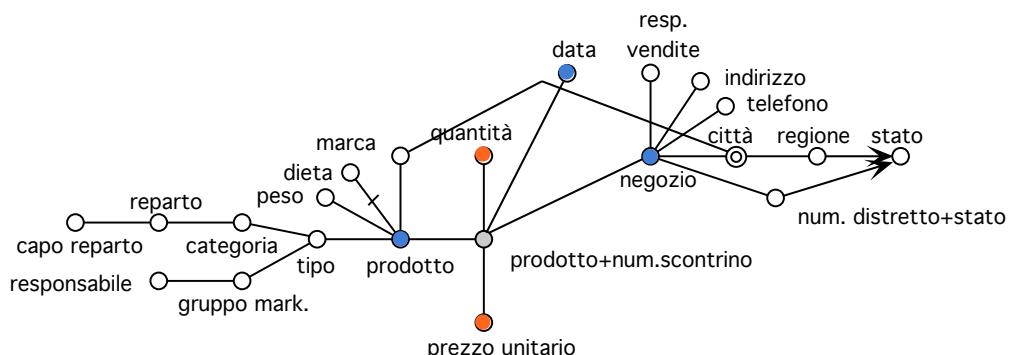
prodotto	numScontrino	quantità	data	negozi
vite	S1	10	2/2/2019	DiTutto
bullone	S1	5	2/2/2019	DiTutto
vite	S2	3	2/2/2019	DiTutto
dado	S2	8	2/2/2019	DiTutto
dado	S3	4	2/2/2019	DiTutto



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

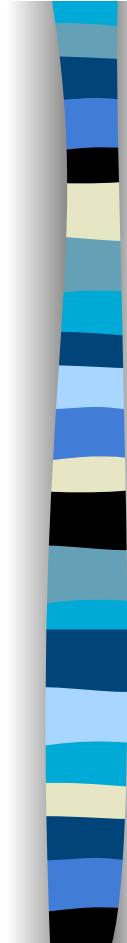
99

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

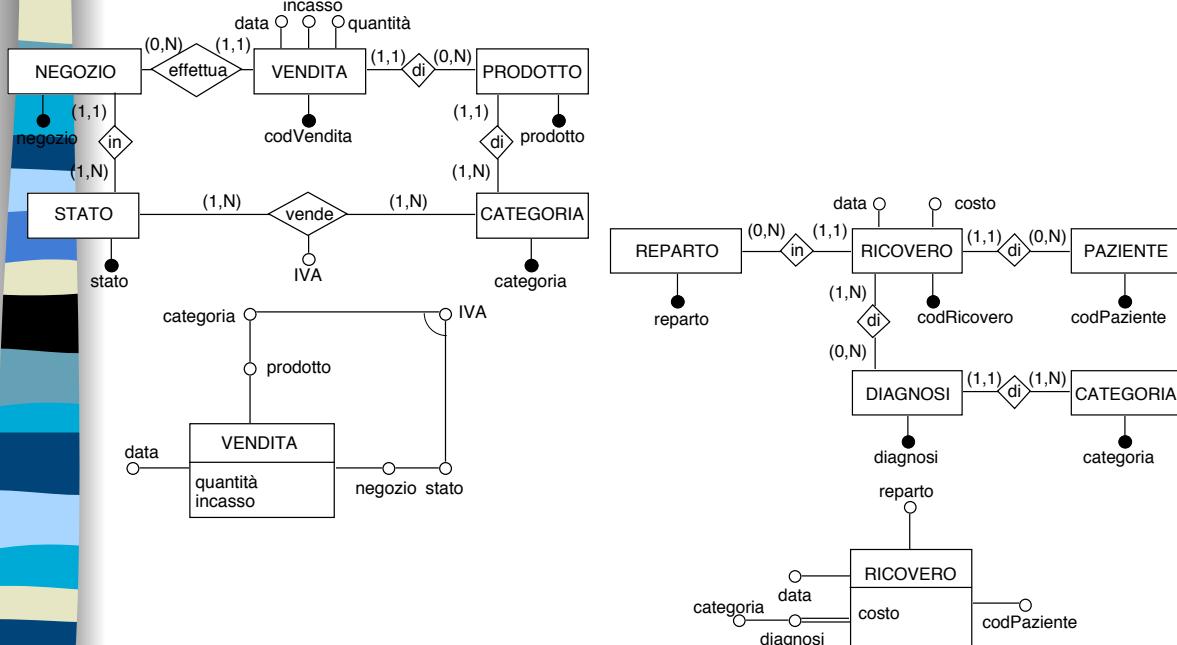


Creazione dello schema di fatto

- 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

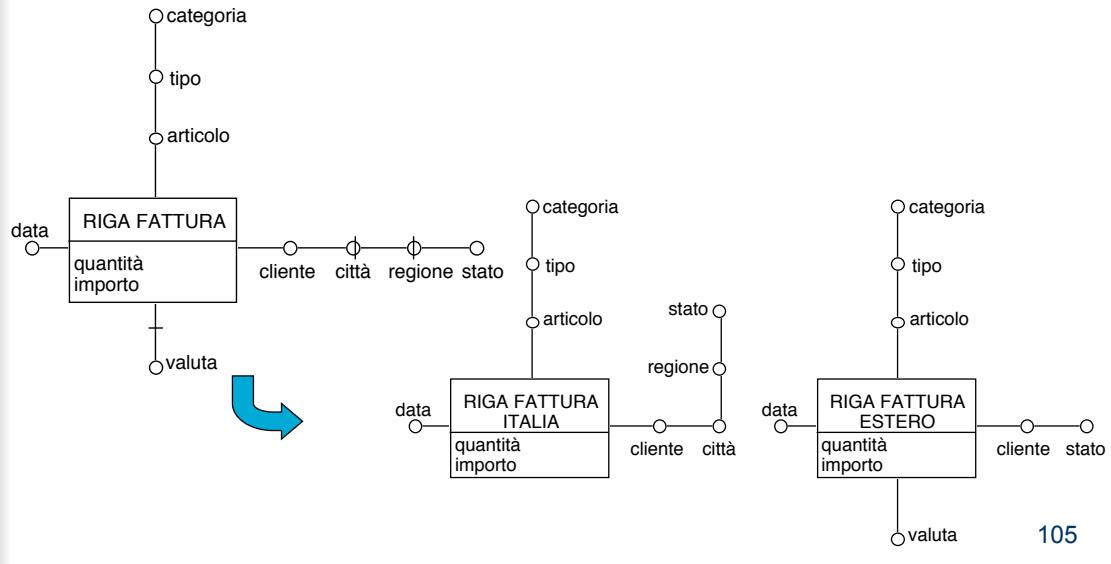
- In questa fase devono anche essere identificate le eventuali non-additività e non-aggregabilità presenti nello schema, considerando tutte le accoppiate dimensione-misura
- Dato uno schema di fatto n-dimensionale, per la dimensione d_i e la misura m_j , la domanda da porsi sarà:

“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

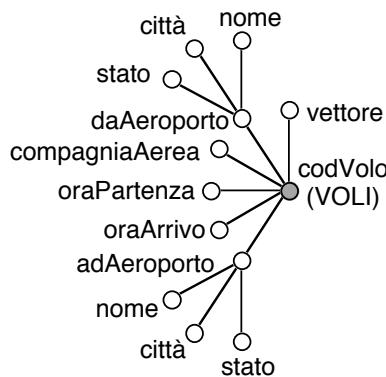
Frammentazione dello schema di fatto

- 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



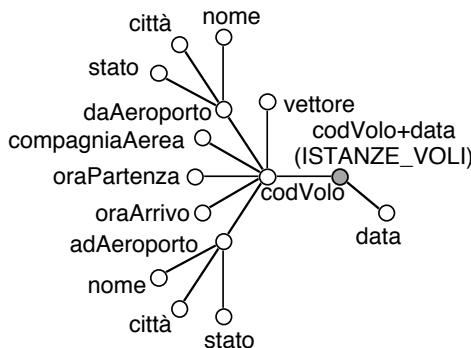
L'esempio dei voli

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

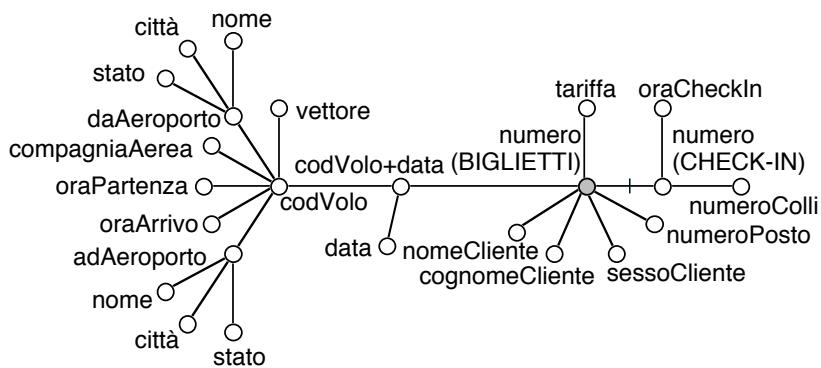
VOLI(codVolo, compagniaAerea, daAeroporto:AEROPORTI,
 adAeroporto:AEROPORTI, oraPartenza, oraArrivo, vettore)
 ISTANZE_VOLI(codVolo:VOLI, data)
 AEROPORTI(sigla, nome, città, stato)
 BIGLIETTI(numero, (codVolo, data):ISTANZE_VOLI, numPosto, tariffa,
 nomeCliente, cognomeCliente, sessoCliente)
 CHECK-IN(numero:BIGLIETTI,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



107

L' esempio dei voli

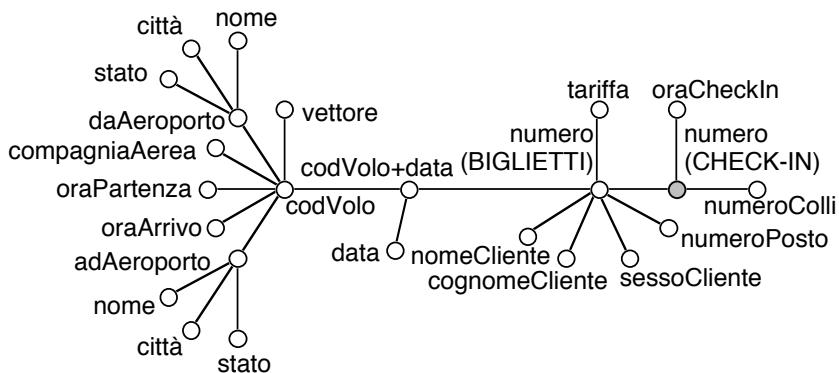
VOLI(codVolo, compagniaAerea, daAeroporto:AEROPORTI,
 adAeroporto:AEROPORTI, oraPartenza, oraArrivo, vettore)
 ISTANZE_VOLI(codVolo:VOLI, data)
 AEROPORTI(sigla, nome, città, stato)
 BIGLIETTI(numero, (codVolo, data):ISTANZE_VOLI, numPosto, tariffa,
 nomeCliente, cognomeCliente, sessoCliente)
 CHECK-IN(numero:BIGLIETTI,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



108

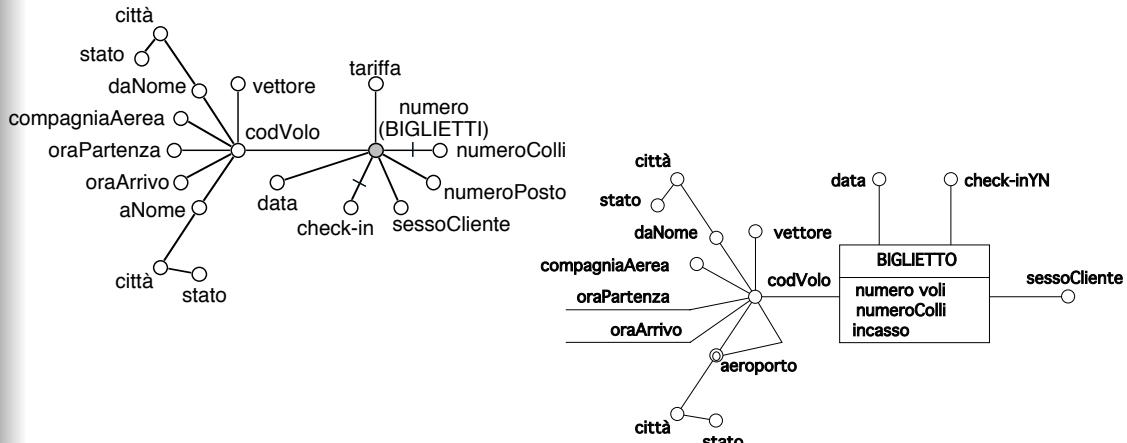
L' esempio dei voli

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

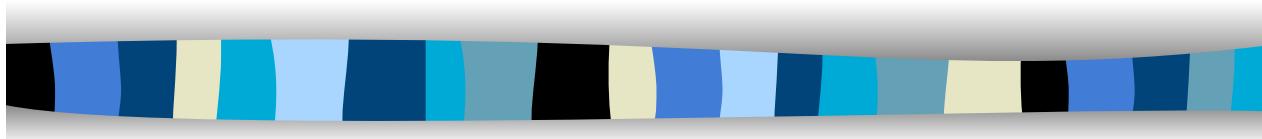


```

    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

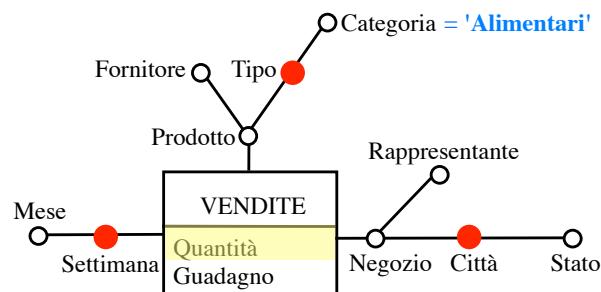


Il carico di lavoro

- 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

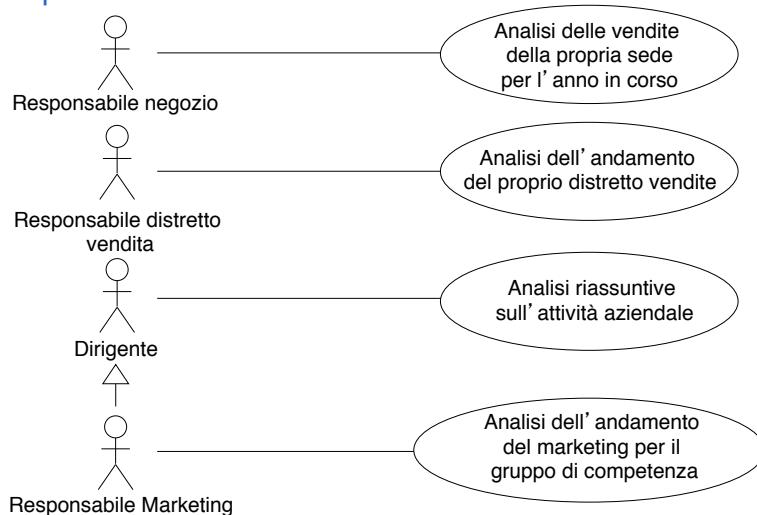
Il carico di lavoro e gli utenti

- 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

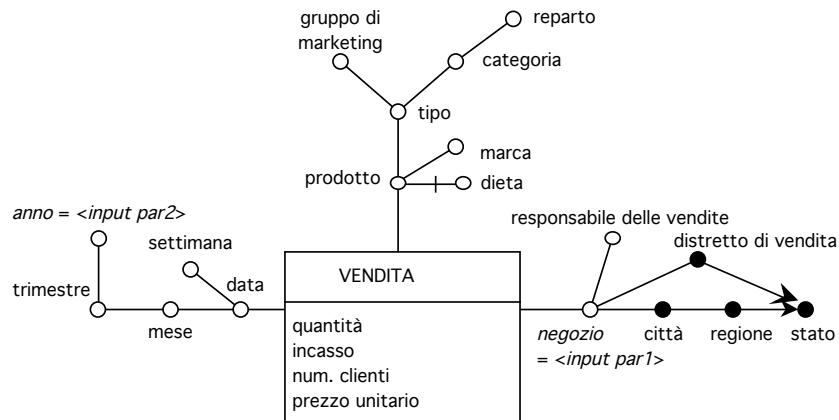


116

Il carico di lavoro e gli utenti

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

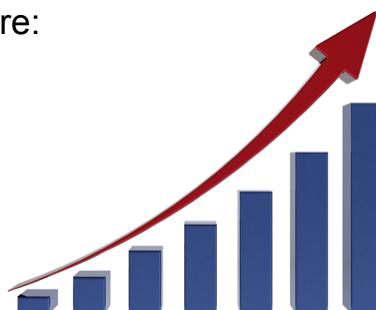
- ✓ Quali misure e quali attributi descrittivi sono visualizzabili
- ✓ Quali gerarchie e quali attributi dimensionali sono navigabili
- ✓ Quali restrizioni a livello di istanze è necessario applicare sui dati



117

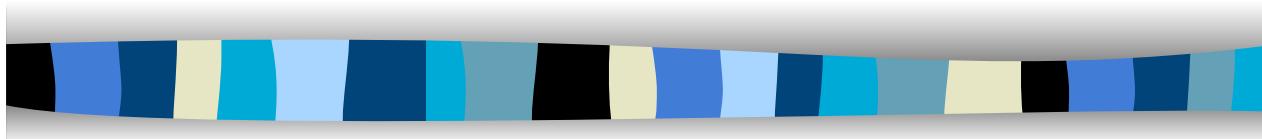
Il volume dati

- 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

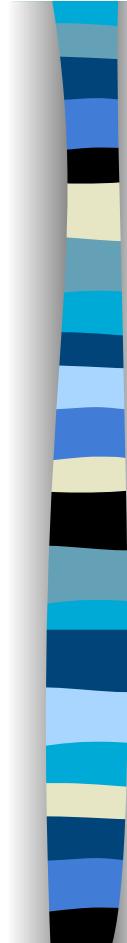
Progettazione logica



Modelli logici per il Data Mart

- Mentre la modellazione concettuale è indipendente dal modello logico prescelto per l' implementazione, evidentemente lo stesso non si può dire per i temi legati alla modellazione logica.
- La struttura multidimensionale dei dati può essere rappresentata utilizzando due distinti modelli logici:
 - ✓ **MOLAP** (*Multidimensional On-Line Analytical Processing*) memorizzano i dati utilizzando strutture intrinsecamente multidimensionali (es. vettori multidimensionali).
 - ✓ **ROLAP** (*Relational On-Line Analytical Processing*) utilizza il ben noto modello relazionale per la rappresentazione dei dati multidimensionali.

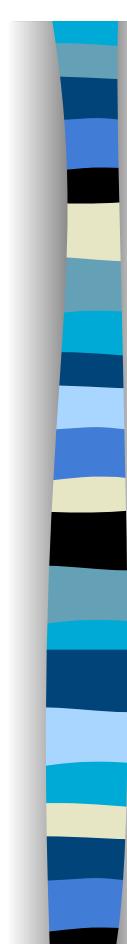




Sistemi MOLAP

- L' utilizzo di soluzioni MOLAP:
 - ✓ Rappresenta una soluzione naturale e può fornire ottime prestazioni poiché le operazioni non devono essere "simulate" mediante complesse istruzioni SQL.
 - ✓ Pone il problema della sparsità: in media solo il 20% delle celle dei cubi contiene effettivamente informazioni, mentre le restanti celle corrispondono a fatti non accaduti.
 - ✓ È frenato dalla mancanza di strutture dati standard: i diversi produttori di software utilizzano strutture proprietarie che li rendono difficilmente sostituibili e accessibili mediante strumenti di terze parti.
 - ✓ Progettisti e sistemisti sono riluttanti a rinunciare alla loro ormai ventennale esperienza sui sistemi relazionali.

3

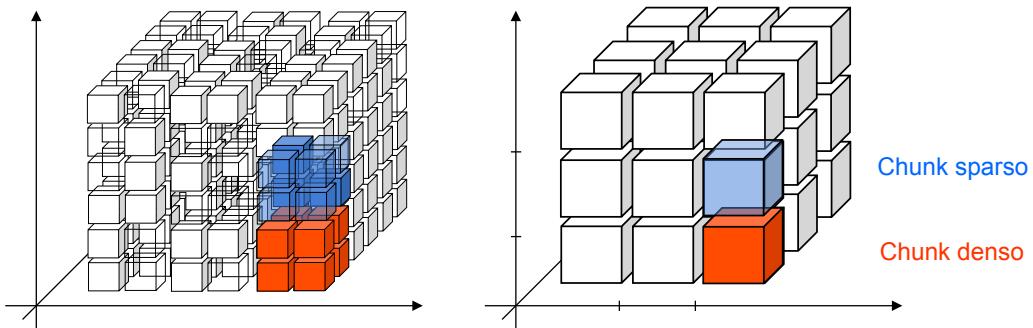


Sistemi MOLAP e sparsità

- Le tecniche di gestione della sparsità sono basate sui seguenti principi:
 - ✓ **Suddivisione delle dimensioni:** consiste nel partizionare un cubo n -dimensionale in più sottocubi n -dimensionali (*chunk*). I singoli chunk potranno essere caricati più agevolmente in memoria e potranno essere gestiti in modo differente a seconda che siano *densi* (la maggior parte delle celle contiene informazioni) oppure *sparsi* (la maggior parte delle celle non contiene informazioni).
 - ✓ **Compressione dei chunk:** i chunk sparsi vengono rappresentati in forma compressa al fine di evitare lo spreco di spazio dovuto alla rappresentazione di celle che non contengono informazioni.

4

Sistemi MOLAP e sparsità



Una struttura dati comunemente usata per la compressione dei chunk sparsi prevede un indice che riporti il solo offset delle celle che effettivamente contengono informazioni.

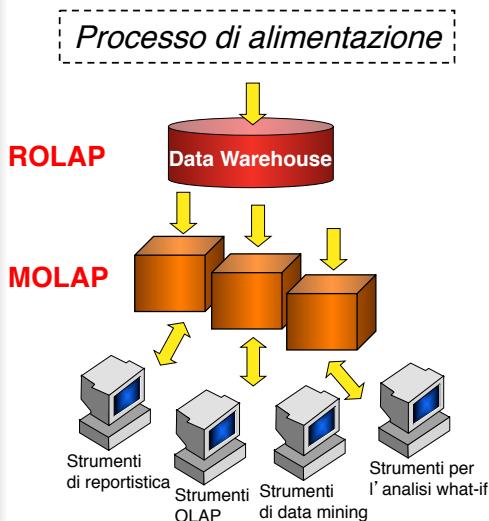
5

ROLAP, MOLAP e HOLAP

- I sistemi commerciali si differenziano in base al modello logico adottato.
- Sebbene la maggior parte dei sistemi, soprattutto di grandi dimensioni, sia realizzato con soluzioni ROLAP, cominciano ad essere proposte anche alcune soluzioni ibride (Hybrid-OLAP)
- Le soluzioni HOLAP sfruttano le proprietà di entrambi i modelli....

6

HOLAP



- Il DW ROLAP è ottimale per memorizzare enormi quantità di dati
- I DM MOLAP massimizzano la velocità di accesso ai dati
- I cubi MOLAP possono anche essere creati ‘al volo’ per svolgere specifiche sessioni di analisi (report semi-statici)

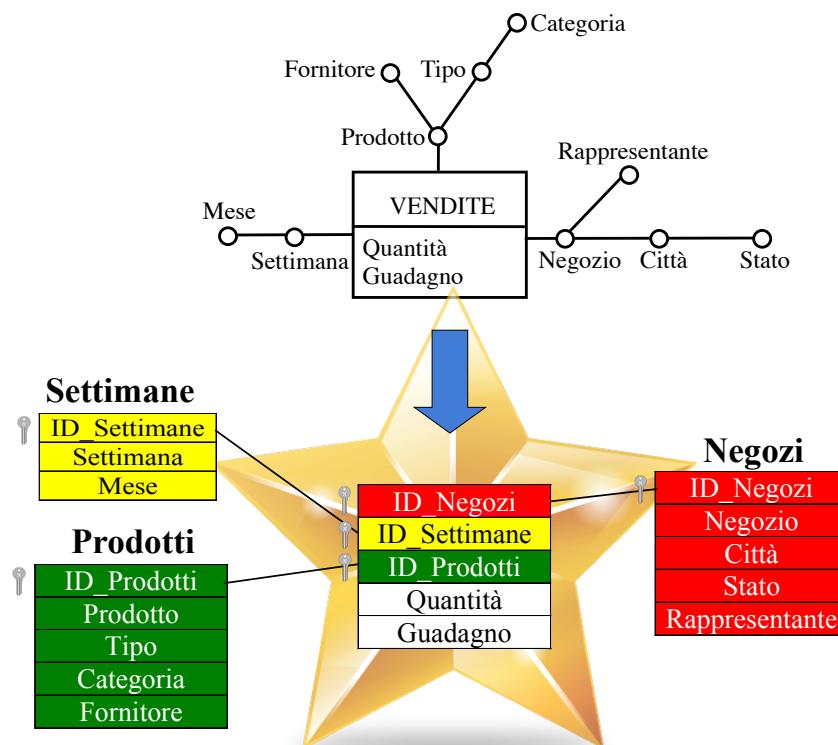
7

ROLAP: lo schema a stella

- La modellazione multidimensionale su sistemi relazionali è basata sul cosiddetto *schema a stella* (*star schema*) e sulle sue varianti.
- Uno schema a stella è composto da:
 - ✓ Un insieme di relazioni DT_1, \dots, DT_n , chiamate *dimension table*, ciascuna corrispondente a una dimensione. Ogni DT_i è caratterizzata da una chiave primaria (tipicamente surrogata) d_i e da un insieme di attributi che descrivono le dimensioni di analisi a diversi livelli di aggregazione.
 - ✓ Una relazione FT , chiamata *fact table*, che importa le chiavi di tutte le dimension table. La chiave primaria di FT è data dall’insieme delle chiavi esterne dalle dimension table, d_1, \dots, d_n ; FT contiene inoltre un attributo per ogni misura.

8

Lo schema a stella



9

Lo schema a stella

ID_Negozi	Negozio	Città	Stato	Rappresentante
1	DiTutto	Roma	I	Rossi
2	DiPiù	Roma	I	Rossi
3	NonSolo	Milano	I	Verdi
4	MaAnche	Milano	I	Verdi

Dimension Table

ID_Negozi	ID_Sett	ID_Prodotti	Quantità	Guadagno
1	1	1	100	100
1	2	1	150	150
3	3	4	350	350
4	4	4	200	200

Fact Table

ID_Sett.	Settimana	Mese
1	1-2019	Gen19
2	2-2019	Gen19
3	3-2019	Gen19
4	4-2019	Gen19

Dimension Table

ID_Prodotti	Prodotto	Tipo	Categoria	Fornitore
1	Pecorino	Latticini	Alimentari	Bianchi
2	Emmenthal	Latticini	Alimentari	Bianchi
3	Cola	Bibite	Alimentari	Carli
4	Aranciata	Bibite	Alimentari	Carli

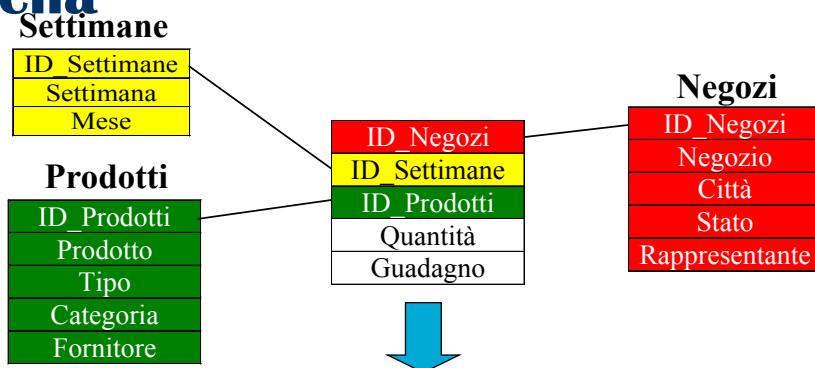
10

Lo schema a stella: considerazioni

- Le Dimension Table sono completamente denormalizzate (es. Prodotto → Tipo)
 - ↳ È sufficiente un join per recuperare tutti i dati relativi a una dimensione
 - ↳ C'è una forte ridondanza nei dati
- Non si hanno problemi di sparsità in quanto vengono memorizzate soltanto le tuple corrispondenti a punti dello spazio multi-dimensionale per cui esistono eventi

11

Interrogazioni OLAP su schemi a stella



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

```
select      Città, Settimana, Tipo, sum(Quantità)
from        Settimane, Negozi, Prodotti, Vendite
where       Settimane.ID_Settimane=Vendite.ID_Settimane and
           Negozi.ID_Negozi =Vendite.ID_Negozi and
           Prodotti.ID_Prodotti =Vendite.ID_Prodotti and
           Prodotti.Categoria = 'Alimentari'
group by    Città, Settimana, Tipo;
```

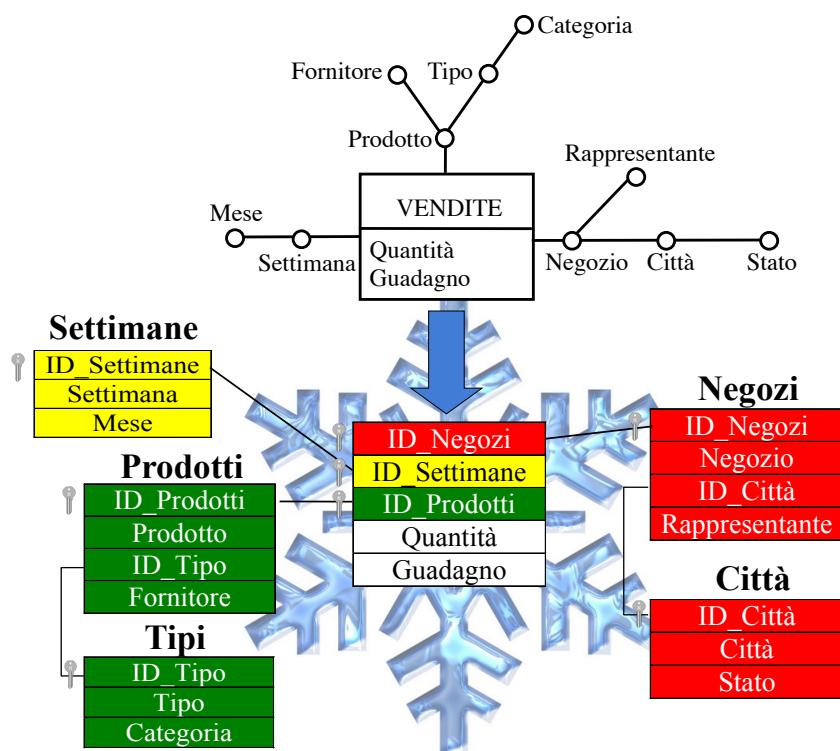
12

Lo snowflake schema

- Lo schema a fiocco di neve (*snowflake schema*) riduce la denormalizzazione delle dimension table DT_i degli schemi a stella eliminando alcune delle dipendenze transitive che le caratterizzano.
- Le dimension table $DT_{i,j}$ di questo schema sono caratterizzate da:
 - ✓ una chiave primaria (tipicamente surrogata) $d_{i,j}$
 - ✓ il sottoinsieme degli attributi di DT_i che dipendono funzionalmente da $d_{i,j}$.
 - ✓ zero o più chiavi esterne a importate da altre $DT_{i,k}$ necessarie a garantire la ricostruibilità del contenuto informativo di DT_i .
- Denominiamo **primarie** le dimension table le cui chiavi sono importate nella fact table, **secondarie** le rimanenti.

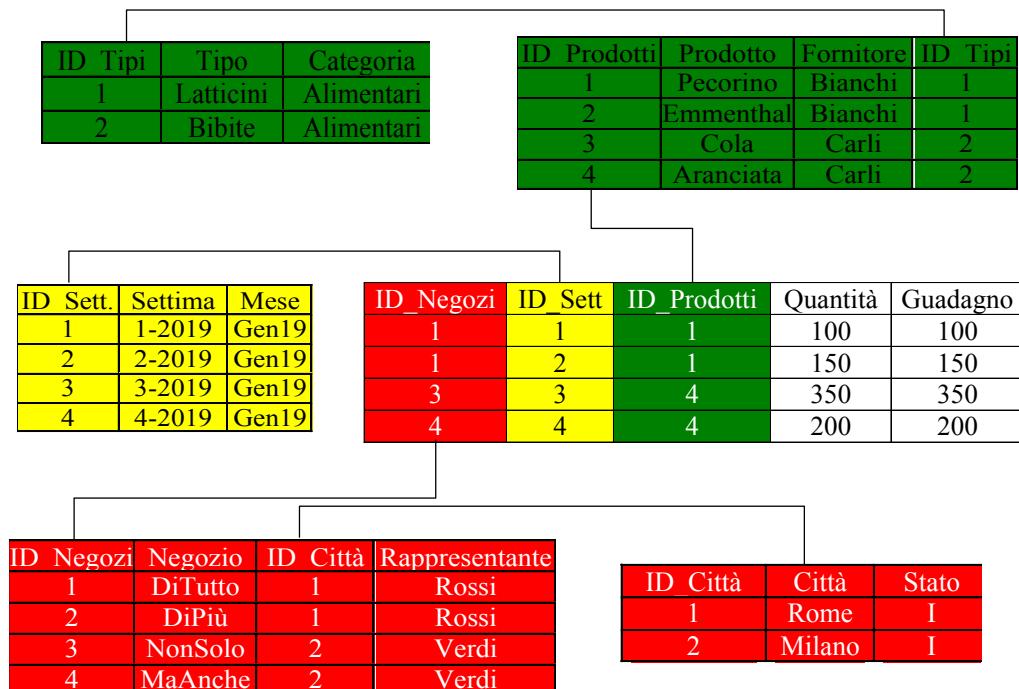
13

Lo snowflake schema



14

Lo snowflake schema



15

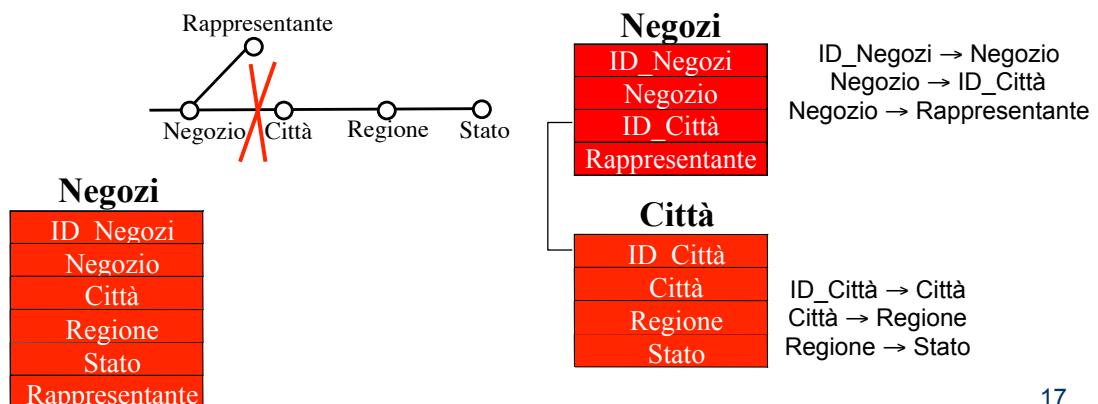
Lo snowflake schema: considerazioni

- Lo spazio richiesto per la memorizzazione dei dati si riduce grazie alla normalizzazione
- È necessario inserire nuove chiavi surrogate che permettano di determinare le corrispondenze tra dimension table primarie e secondarie
- L' esecuzione di interrogazioni che coinvolgono solo gli attributi contenuti nella fact table e nelle dimension table primarie è avvantaggiata
- Il tempo di esecuzione delle interrogazioni che coinvolgono attributi delle dimension table secondarie aumenta

16

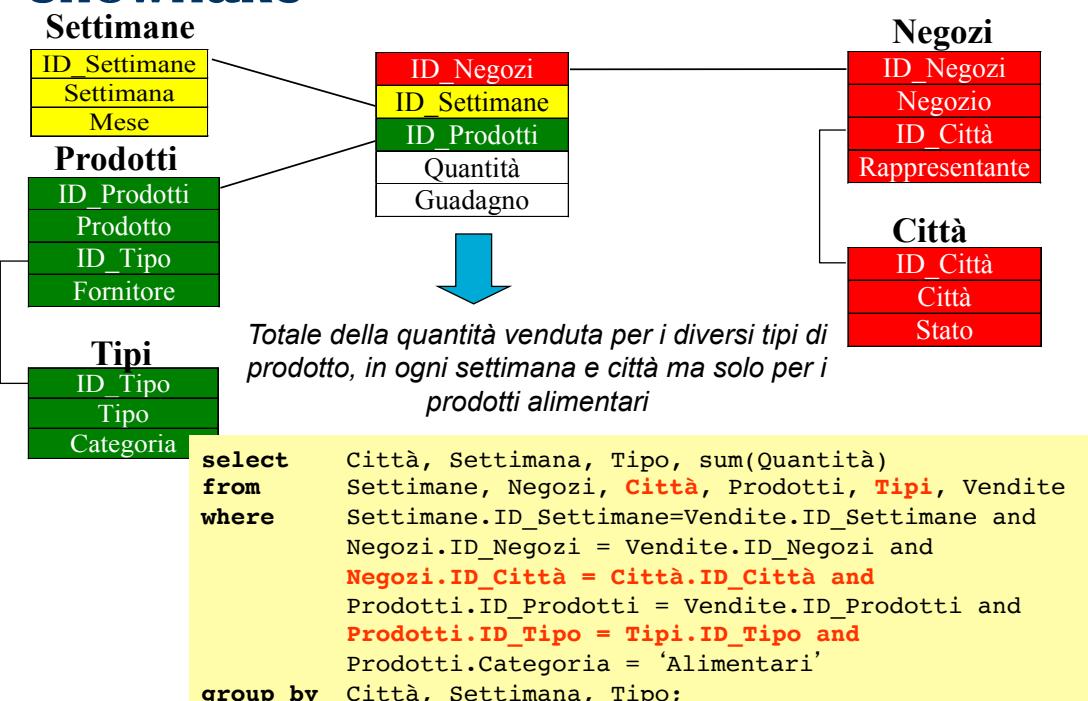
Normalizzazione con lo snowflake schema

- Le specifiche caratteristiche degli schemi a stella richiedono particolare attenzione affinché nella nuova relazione sia spostato il corretto insieme di attributi
- La presenza di più dipendenze funzionali transitive in cascata fa sì che, affinché la decomposizione sia efficace, tutti gli attributi che dipendono (transitivamente e non) dall'attributo che ha determinato lo snowflaking siano posti nella nuova relazione



17

Interrogazioni OLAP su schemi snowflake



18

Le viste

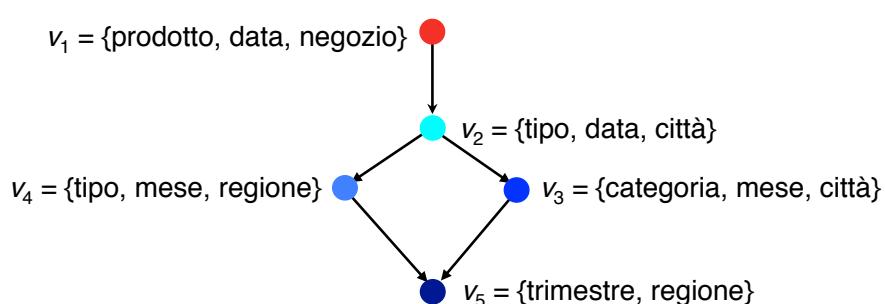
- L' analisi dei dati al massimo livello di dettaglio è spesso troppo complessa e non interessante per gli utenti che richiedono dati di sintesi
- L' aggregazione rappresenta il principale strumento per ottenere informazioni di sintesi
- L' elevato costo computazionale connesso con l' aggregazione induce a precalcolare i dati di sintesi maggiormente utilizzati

Con il termine *vista* si denotano le fact table contenenti dati aggregati

19

Le viste

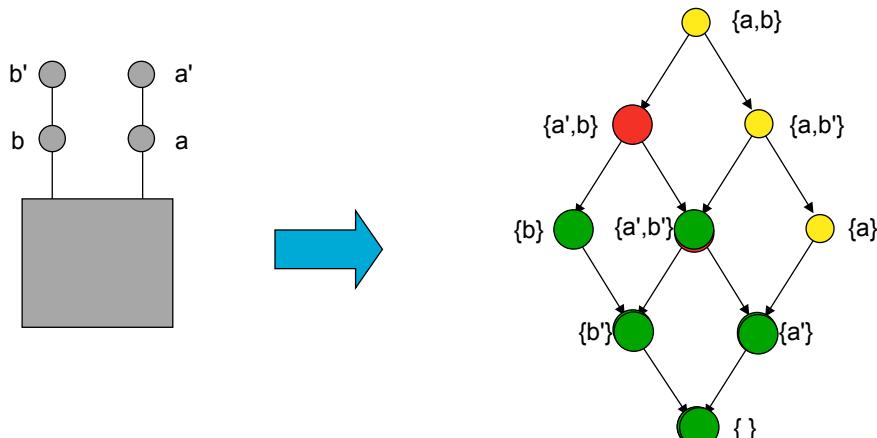
- Le viste possono essere identificate in base al livello (*group-by set*) di aggregazione che le caratterizza



20

Risolvibilità delle interrogazioni

- Una vista v sul group-by set p non serve solo per le interrogazioni con group-by set p ma anche per tutte quelle che richiedono i dati a group-by set p' più aggregati di p ($p \leq p'$)



Reticolo multidimensionale

21

Aggregazioni parziali

- Per la corretta gestione dei dati aggregati può essere necessario introdurre nuove misure
 - Misure derivate:** ottenute applicando operatori matematici a due o più valori appartenenti alla stessa tupla

Tipo	Prodotto	Quantità	Prezzo	Incasso
T1	P1	5	1,00	5,00
T1	P2	7	1,50	10,50
T2	P3	9	0,80	7,20

Sum AVG

Tipo	Quantità	Prezzo	Incasso
T1	12	1,25	15,00
T2	9	0,80	7,20

22,70 ? 22,20

La soluzione corretta è sempre quella che si ottiene aggregando i dati direttamente dalla vista primaria

22

Aggregazioni parziali

- ✓ **Misure di supporto:** sono necessarie in presenza di operatori di aggregazione non distributivi

Data	Livello di inventario
1/1/1999	100
10/2/1999	200
31/4/1999	60
5/6/1999	85
18/7/1999	125
31/12/1999	110

1999

113,33

Trimestre	Livello di inventario	Count	Livello di inventario
4/1999	120	3	360
8/1999	105	2	210
12/1999	110	1	110
1999			111,66
			113,33



La soluzione corretta è sempre quella che si ottiene aggregando i dati direttamente dalla vista primaria

23

Classificazione degli operatori di aggregazione

- I problemi visti in precedenza derivano dalla natura degli operatori di aggregazione che possono essere così classificati:
 - ✓ **Distributivi:** permettono di calcolare dati aggregati a partire direttamente da dati parzialmente aggregati (es. somma, massimo, minimo)
 - ✓ **Algebrici:** richiedono un numero finito di informazioni aggiuntive (*misure di supporto*) per calcolare dati aggregati a partire da dati parzialmente aggregati (es. media – richiede il numero dei dati elementari che hanno contribuito a formare un singolo dato parzialmente aggregato)
 - ✓ **Olistici:** non permettono di calcolare dati aggregati a partire da dati parzialmente aggregati utilizzando un numero finito di informazioni aggiuntive (es. mediana, moda)

24

Schemi relazionali e viste

- La soluzione più semplice consiste nell' utilizzare lo schema a stella memorizzando tutti i dati in una sola fact table
 - ✓ La dimensione dell' unica fact table cresce considerevolmente a discapito delle prestazioni
 - ✓ Le dimension table contengono tuple relative a diversi livelli di aggregazione. Il valore NULL viene utilizzato per identificare l' origine delle tuple

25

Schemi relazionali e viste

Sono relative al group-by set:
{Negozi, Settimane, **Prodotti**}

ID_Negozi	ID_Sett	ID_Prodotti	Quantità	Guadagno
1	1	1	100	100
1	2	1	150	150
3	3	4	350	350
4	4	4	200	200
2	1	5	3600	3600
1	3	6	2400	2400
1	1	7	1000	1000

ID_Prodotti	Prodotto	Tipo	Categoria	Fornitori
1	Pecorino	Latticini	Alimentari	Bianchi
2	Emmenthal	Latticini	Alimentari	Bianchi
3	Cola	Bibite	Alimentari	Carli
4	Aranciata	Bibite	Alimentari	Carli
5	-	Latticini	Alimentari	Bianchi
6	-	Bibite	Alimentari	Carli
7	-	-	-	Bianchi
8	-	-	-	Carli

26

Schemi relazionali e viste

Sono relative al group-by set:
{Negozi, Settimane, **Tipo**}

ID_Negozi	ID_Sett	ID_Prodotti	Quantità	Guadagno
1	1	1	100	100
1	2	1	150	150
3	3	4	350	350
4	4	4	200	200
2	1	5	3600	3600
1	3	6	2400	2400
1	1	7	1000	1000



ID_Prodotti	Prodotto	Tipo	Categoria	Fornitore
1	Pecorino	Latticini	Alimentari	Bianchi
2	Emmenthal	Latticini	Alimentari	Bianchi
3	Cola	Bibite	Alimentari	Carli
4	Aranciata	Bibite	Alimentari	Carli
5	-	Latticini	Alimentari	Bianchi
6	-	Bibite	Alimentari	Carli
7	-	-	-	Bianchi
8	-	-	-	Carli

27

Schemi relazionali e viste

È relativa al group-by set:
{Negozi, Settimane, **Fornitore**}

ID_Negozi	ID_Sett	ID_Prodotti	Quantità	Guadagno
1	1	1	100	100
1	2	1	150	150
3	3	4	350	350
4	4	4	200	200
2	1	5	3600	3600
1	3	6	2400	2400
1	1	7	1000	1000

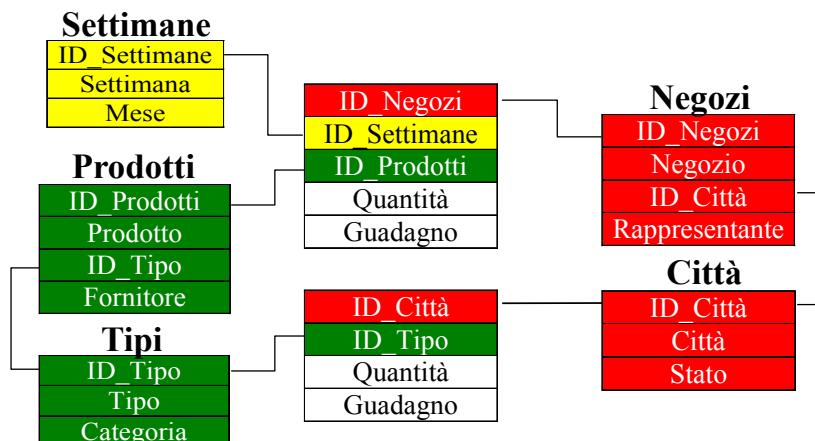


ID_Prodotti	Prodotto	Tipo	Categoria	Fornitore
1	Pecorino	Latticini	Alimentari	Bianchi
2	Emmenthal	Latticini	Alimentari	Bianchi
3	Cola	Bibite	Alimentari	Carli
4	Aranciata	Bibite	Alimentari	Carli
5	-	Latticini	Alimentari	Bianchi
6	-	Bibite	Alimentari	Carli
7	-	-	-	Bianchi
8	-	-	-	Carli

28

Schemi relazionali e viste

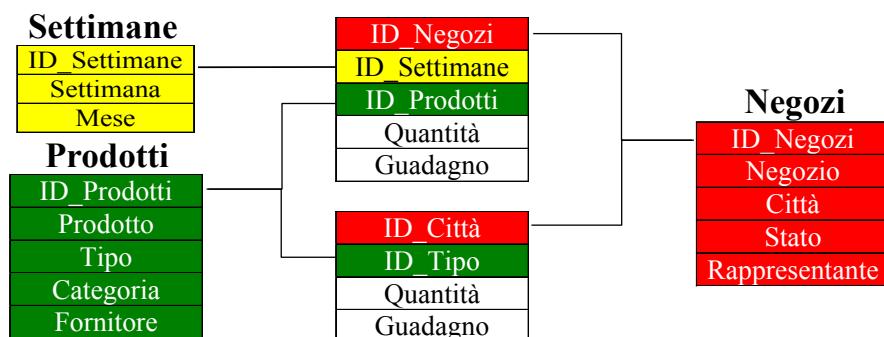
- Adottando lo snowflake schema è possibile memorizzare in fact table separate dati appartenenti a diversi group-by set
 - ✓ Lo snowflaking deve essere applicato in corrispondenza dei livelli di aggregazione a cui sono presenti viste



29

Schemi relazionali e viste

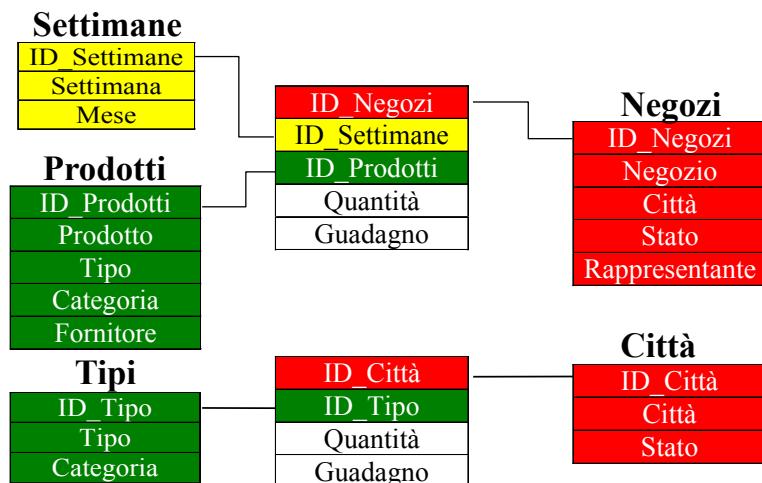
- Una soluzione intermedia rispetto alle due presentate prevede di memorizzare in fact table separate dati relativi a group-by set diversi senza però ricorrere alla normalizzazione delle dimension table (*constellation schema*)
 - ✓ L'accesso alle fact table è ottimizzato, quello alle dimension table no
 - ✓ La dimensione delle fact table è di molto superiore a quella delle dimension table e conseguentemente la loro ottimizzazione gioca un ruolo fondamentale



30

Schemi relazionali e viste

- Il massimo livello delle prestazioni si ottiene memorizzando in fact table separate dati a diversi livelli di aggregazione e replicando completamente anche le dimension table

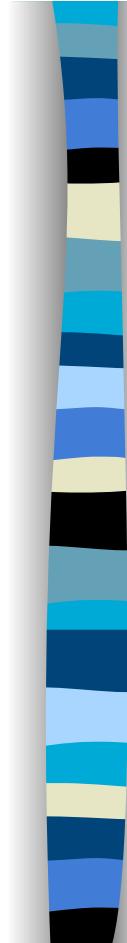


31

Aggregate navigator

- La presenza di più fact table contenenti i dati necessari a risolvere una data interrogazione pone il problema di determinare la vista che determinerà il minimo costo di esecuzione
- Questo ruolo è svolto dagli *aggregate navigator*, ossia i moduli preposti a riformulare le interrogazioni OLAP sulla “migliore” vista a disposizione
- Gli aggregate navigator dei sistemi commerciali gestiscono attualmente solo gli operatori distributivi riducendo così l’ utilità delle misure di supporto

32



Scenari temporali

- Il modello multidimensionale assume che gli eventi che istanziano un fatto siano **dinamici**, e che i valori degli attributi che popolano le gerarchie siano **statici**
- Questa visione non è realistica poiché anche i valori presenti nelle gerarchie variano nel tempo dando vita alle gerarchie dinamiche (**slowly-changing dimension**)
- L'adozione di gerarchie dinamiche implica un sovraccosto in termini di spazio e può comportare una forte riduzione delle prestazioni

33

Scenari temporali

.... Sono possibili diverse soluzioni

- Oggi per ieri (*attualizzazione*)
 - ✓ I dati vengono interpretati in base all'attuale configurazione della gerarchia
 - ✓ Implementabile sullo schema a stella
- Oggi o ieri (*verità storica*)
 - ✓ I dati vengono interpretati in base alla configurazione valida al momento in cui sono stati registrati
 - ✓ Implementabile sullo schema a stella
- Ieri per oggi (*retrodatazione*)
 - ✓ I dati vengono interpretati in base alla configurazione della gerarchia valida in un particolare istante
 - ✓ Richiede la storizziazione dei dati

34

Un esempio

Situazione al 1/1/2011

negozio	responsabile
DiTutto	Rossi
NonSoloPile	Bianchi
NonSoloPane	Bianchi

Situazione al 1/11/2011

negozio	responsabile
DiTutto	Rossi
NonSoloPile	Bianchi
PaneEPizza	Rossi

Situazione al 1/7/2011

negozio	responsabile
DiTutto	Rossi
NonSoloPile	Bianchi
NonSoloPane	Rossi

Situazione al 1/1/2012

negozio	responsabile
DiTutto	Bianchi
NonSoloPile	Bianchi
PaneEPizza	Rossi
DiTuttoDiPiù	Rossi

35

Un esempio

negozio	responsabile
DiTutto	Rossi
NonSoloPile	Bianchi
NonSoloPane	Bianchi

negozio	data	incasso
DiTutto	20/6/2011	10
NonSoloPile	20/6/2011	20
NonSoloPile	30/6/2011	15
NonSoloPile	2/7/2011	10
DiTutto	2/7/2011	30
NonSoloPile	10/7/2011	15
NonSoloPile	12/7/2011	10
NonSoloPile	15/7/2011	20

1/7/2011

tempo

36

Un esempio

negozi	responsabile
DiTutto	Rossi
NonSoloPile	Bianchi
NonSoloPane	Bianchi

negozi	data	incasso
DiTutto	20/6/2011	10
NonSoloPane	20/6/2011	20
NonSoloPane	30/6/2011	15
NonSoloPile	2/7/2011	10
DiTutto	2/7/2011	30
NonSoloPile	10/7/2011	15
NonSoloPile	12/7/2011	10
NonSoloPile	15/7/2011	20

- Incassi totali per responsabile (16/7/2011)

✓ attualizzazione

responsabile	incasso
Rossi	100
Bianchi	30

Un esempio

negozi	responsabile
DiTutto	Rossi
NonSoloPile	Bianchi
NonSoloPane	Bianchi

negozi	data	incasso
DiTutto	20/6/2011	10
NonSoloPane	20/6/2011	20
NonSoloPane	30/6/2011	15
NonSoloPile	2/7/2011	10
DiTutto	2/7/2011	30
NonSoloPile	10/7/2011	15
NonSoloPile	12/7/2011	10
NonSoloPile	15/7/2011	20

- Incassi totali per responsabile (16/7/2011)

✓ attualizzazione

✓ verità storica

responsabile	incasso
Rossi	100
Bianchi	30

responsabile	incasso
Rossi	65
Bianchi	65

Un esempio

negozi	responsabile
DiTutto	Rossi
NonSoloPile	Bianchi
NonSoloPane	Bianchi

negozi	responsabile
DiTutto	Rossi
NonSoloPile	Bianchi
NonSoloPane	Rossi

negozi	data	incasso
DiTutto	20/6/2011	10
NonSoloPane	20/6/2011	20
NonSoloPane	30/6/2011	15
NonSoloPane	2/7/2011	10
DiTutto	2/7/2011	30
NonSoloPane	10/7/2011	15
NonSoloPile	12/7/2011	10
NonSoloPile	15/7/2011	20

- Incassi totali per responsabile (16/7/2011)

✓ attualizzazione

✓ verità storica

✓ retrodatazione al 25/6/2016

responsabile	incasso
Rossi	100
Bianchi	30

responsabile	incasso
Rossi	65
Bianchi	65

responsabile	incasso
Rossi	40
Bianchi	90

39

Gerarchie dinamiche: tipo I

- Supportano solo lo scenario oggi per ieri, pertanto tutti gli eventi, anche quelli passati, vengono interpretati in base all'attuale configurazione delle gerarchie senza tenere traccia del passato
- Questa soluzione è realizzabile sullo schema a stella sovrascrivendo il vecchio valore con quello nuovo ogni volta che si verifica un cambiamento

40

Gerarchie dinamiche: tipo I

Situazione al 1/1/2011

chiaveN	negozi	responsabile	...
1	DiTutto	Rossi	...
2	NonSoloPile	Bianchi	...
3	NonSoloPane	Bianchi	...

Situazione al 1/7/2011

chiaveN	negozi	responsabile	...
1	DiTutto	Rossi	...
2	NonSoloPile	Bianchi	...
3	NonSoloPane	Rossi	...

Tutte le vendite di NonSoloPane vengono attribuite a Rossi anche se erano state effettuate durante la gestione di Bianchi

41

Gerarchie dinamiche: tipo II

- Supportano solo lo scenario oggi o ieri, e consentono di registrare la verità storica
- Gli eventi memorizzati nella fact table vengono associati ai dati dimensionali che erano validi quando si è verificato l' evento
- Questa soluzione è realizzabile sullo schema a stella: ogni modifica a una gerarchia comporta l' inserimento di un nuovo record che codifichi le nuove caratteristiche nella dimension table corrispondente
- È possibile adottare strategie diverse per attributi appartenenti alla stessa gerarchia

42

Gerarchie dinamiche: tipo II

Situazione al 1/1/2011

chiaveN	negozi	responsabile	...
1	DiTutto	Rossi	...
2	NonSoloPile	Bianchi	...
3	NonSoloPane	Bianchi	...

Situazione al 1/7/2011

chiaveN	negozi	responsabile	...
1	DiTutto	Rossi	...
2	NonSoloPile	Bianchi	...
3	NonSoloPane	Bianchi	...
4	NonSoloPane	Rossi	...

Dopo l' 1/7 i record della fact table relativi a NonSoloPane importeranno il valore di chiaveN = 4

N.B. Solo le selezioni su campi che hanno subito modifiche sono sensibili alle modifiche stesse!!

43

Gerarchie dinamiche: tipo III

- Supportano tutti gli scenari temporali. La loro adozione richiede la storicizzazione dell' attributo e non può pertanto essere basata sul classico schema a stella
- Gli elementi necessari per la gestione di una gerarchia di tipo 3 sono:
 - ✓ Una coppia di marche temporali (*time-stamp*) che indichino l' intervallo di validità di una tupla
 - ✓ Un meccanismo per individuare le tuple coinvolte in una serie di modifiche (tramite per esempio un attributo *master*)
- In uno schema così modificato la dinamicità viene gestita aggiungendo, per ogni modifica, un nuovo record nella dimension table e aggiornando di conseguenza i valori dei time-stamp e dell' attributo master

44

Gerarchie dinamiche: tipo III

Situazione al 1/1/2011

chiaveN	negozi	responsabile	...	da	a	Master
1	DiTutto	Rossi	...	1/1/2011	—	1
2	NonSoloPile	Bianchi	...	1/1/2011	—	2
3	NonSoloPane	Bianchi	...	1/1/2011	—	3

Situazione al 1/1/2012

chiaveN	negozi	responsabile	...	da	a	Master
1	DiTutto	Rossi	...	1/1/2011	31/12/2011	1
2	NonSoloPile	Bianchi	...	1/1/2011	—	2
3	NonSoloPane	Bianchi	...	1/1/2011	30/6/2011	3
4	NonSoloPane	Rossi	...	1/7/2011	31/10/2011	3
5	PaneEPizza	Rossi	...	1/11/2011	—	3
6	DiTuttoDiPiù	Rossi	...	1/1/2012	—	6
7	DiTutto	Bianchi	...	1/1/2012	—	1

45

Gerarchie dinamiche: tipo III

- Avendo a disposizione lo schema descritto in precedenza è facile realizzare i differenti scenari temporali:
 - ✓ **Oggi per ieri:** si identificano dapprima le tuple della dimension table attualmente valide (in base ai time-stamp) e per ciascuna si individuano eventuali altre tuple da cui esse hanno avuto origine
 - ✓ **Ieri per oggi:** fissata una particolare data si individuano le tuple valide in quel particolare momento, quindi si procede come nel caso precedente
 - ✓ **Oggi o ieri:** non richiede l' analisi delle marche temporali poiché l' aggiornamento delle tuple nelle dimension table avviene come per le gerarchie di tipo 2

46

Gerarchie dinamiche: tipo III

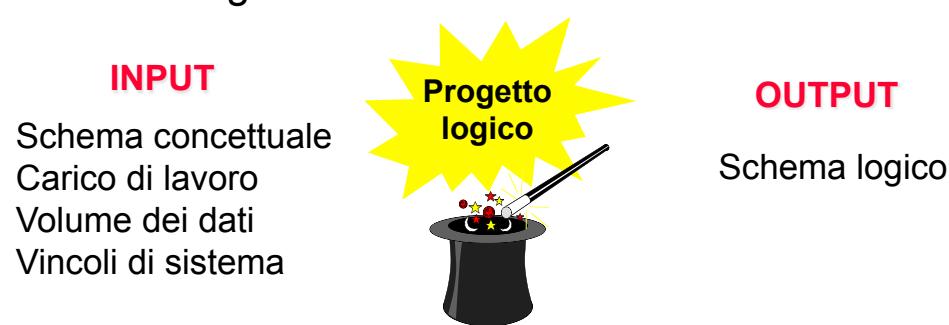
- Utilizzando la soluzione ieri per oggi, l' interrogazione SQL che richiede: “*La quantità totale venduta dai diversi responsabili se si considera l'assegnamento ai negozi vero il 1/10/2011*” è la seguente

```
select      N1.Responsabile, sum(Quantità)
from        Negozi N1, Negozi N2, Vendite
where       N1.Da <= 1/10/11
           AND N1.A > 1/10/11
           AND N1.Master=N2.Master
           AND Vendite.ChiaveN=N2.ChiaveN
group by    N1.Responsabile;
```

47

Progettazione logica

- Include l' insieme dei passi che, a partire dallo schema concettuale, permettono di determinare lo schema logico del data mart



- È basata su principi diversi e spesso in contrasto con quelli utilizzati nei sistemi operazionali
 - ✓ Ridondanza dei dati
 - ✓ Denormalizzazione delle relazioni

48

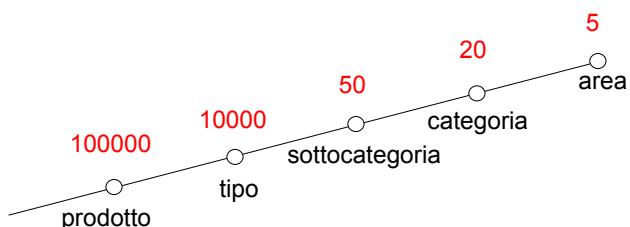
Progettazione logica

- Le principali operazioni da svolgere durante la progettazione logica sono:
 1. Scelta dello schema logico da utilizzare (es. star/snowflake schema)
 2. Traduzione degli schemi concettuali
 3. Scelta delle viste da materializzare
 4. Applicazione di altre forme di ottimizzazione (es. frammentazione verticale/orizzontale)

49

Star VS Snowflake

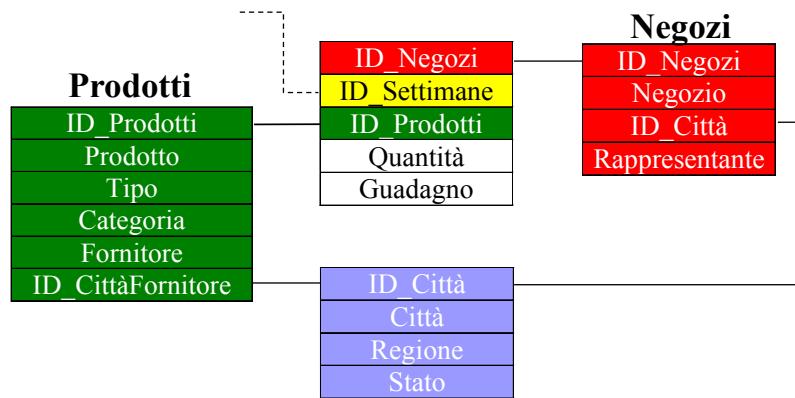
- Esistono pareri contrastanti sull' utilità dello snowflaking:
 - ✓ Contrasta con la filosofia del data warehousing
 - ✓ Rappresenta un inutile “abbellimento” dello schema
- Può essere utile
 1. Quando il rapporto tra le cardinalità della dimension table primaria e secondaria è elevato, poiché determina un forte risparmio di spazio



50

Star VS Snowflake

- Può essere utile
 - 2. Quando una porzione di una gerarchia è comune a più dimensioni

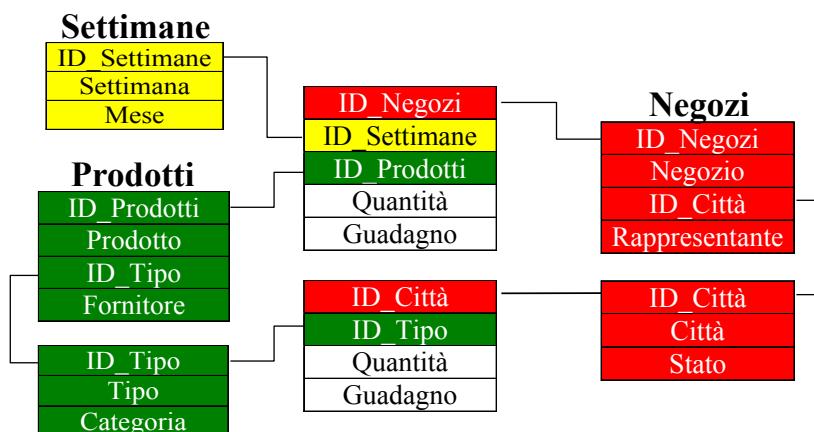


La dimension table secondaria è riutilizzata per più gerarchie

51

Star VS Snowflake

- Può essere utile
 - 3. In presenza di viste aggregate

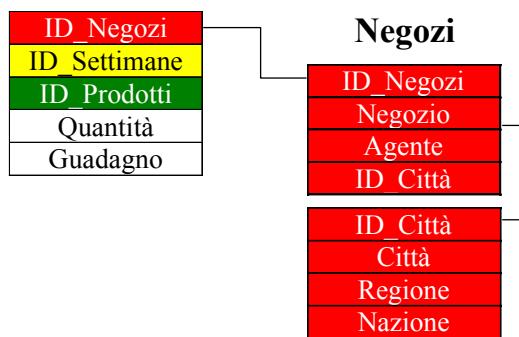


La dimension table secondaria della vista primaria coincide con la dimension table primaria della vista secondaria

52

Star VS Snowflake

- Può essere utile
 - 4. Quando una parte della gerarchia è soggetta a frequenti aggiornamenti



L'agente del negozio varia frequentemente, mentre la regione e nazione della città del negozio sono statici

53

Dagli schemi di fatto agli schemi a stella

- La regola di base per la traduzione di uno schema di fatto in schema a stella prevede di:

Creare una fact table contenente tutte le misure e gli attributi descrittivi direttamente collegati con il fatto e, per ogni gerarchia, creare una dimension table che ne contiene tutti gli attributi.

- In aggiunta a questa semplice regola, la corretta traduzione di uno schema di fatto richiede una trattazione approfondita dei costrutti avanzati del DFM

54

Attributi descrittivi

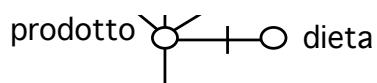
- Contiene informazioni non utilizzabili per effettuare aggregazioni ma che si ritiene comunque utile mantenere
 - ✓ Se collegato a un attributo dimensionale, va incluso nella dimension table che contiene l' attributo
 - ✓ Se collegato direttamente al fatto deve essere incluso nella fact table



55

Archi opzionali

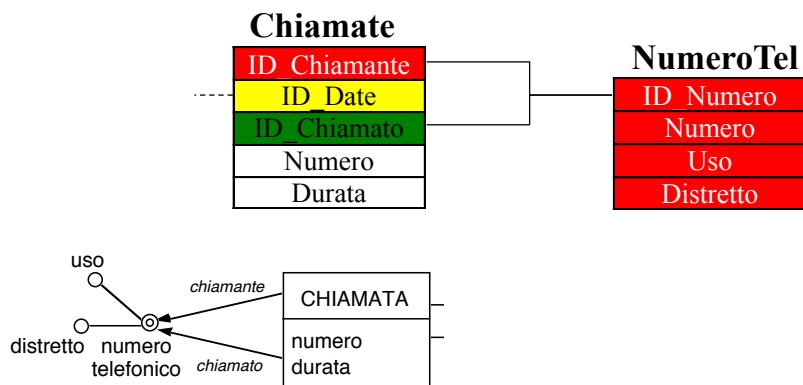
- Alcune porzioni delle gerarchie possono essere opzionali
 - ✓ Nella dimension table, nelle righe per cui non è definito un valore viene inserito un valore fittizio (NULL oppure NON APPLICABILE)
- A causa dei vincoli di integrità, l' opzionalità di un' intera gerarchia NON può essere gestita introducendo un valore nullo nella chiave esterna della fact table, occorre invece inserire un' intera tupla fittizia nella dimension table



56

Gerarchie condivise

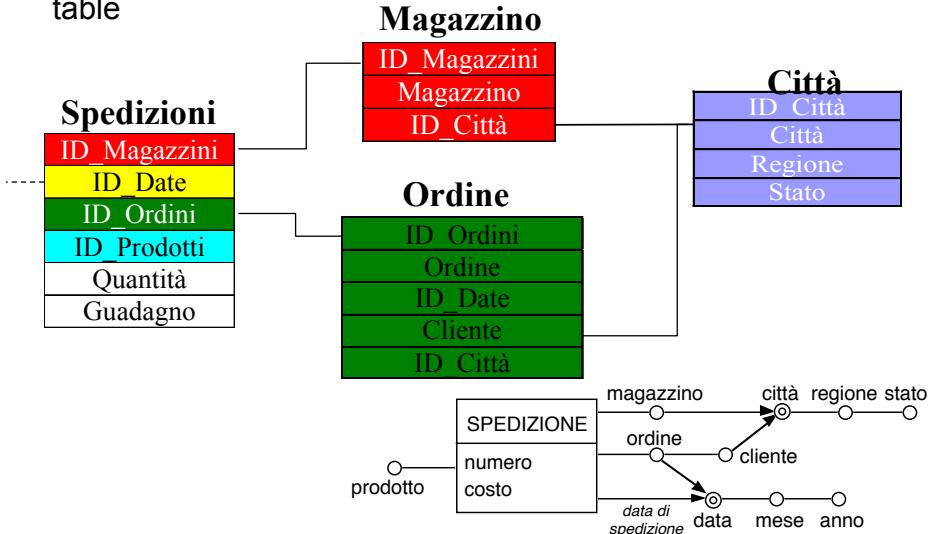
- Se una gerarchia si presenta più volte nello stesso fatto (o in due fatti diversi) non conviene introdurre copie ridondanti delle relative dimension table
- Se le due gerarchie contengono esattamente gli stessi attributi sarà sufficiente importare due volte la chiave della medesima dimension table



57

Gerarchie condivise

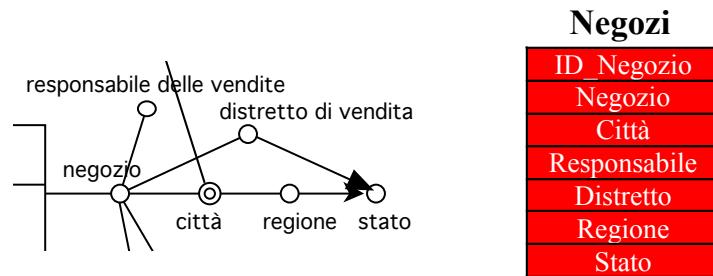
- Se le due gerarchie condividono solo una parte degli attributi è necessario decidere se:
 - Introdurre ulteriore ridondanza nello schema duplicando le gerarchie e replicando i campi comuni
 - Eseguire uno snowflake sul primo attributo condiviso introducendo una terza tabella comune a entrambe le dimension table



58

Convergenza

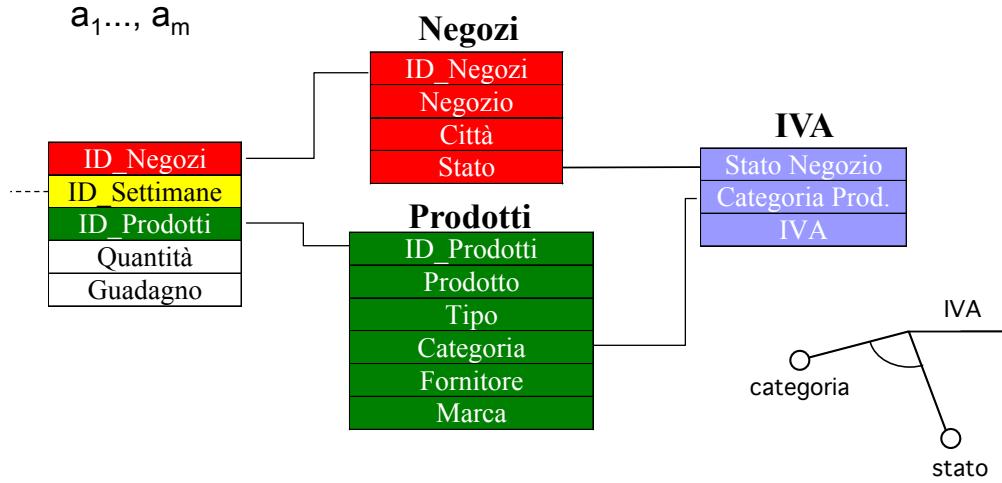
- Gli attributi di convergenza si includono nella stessa dimension table dei loro attributi padri, senza particolari accorgimenti



59

Attributi cross-dimensional

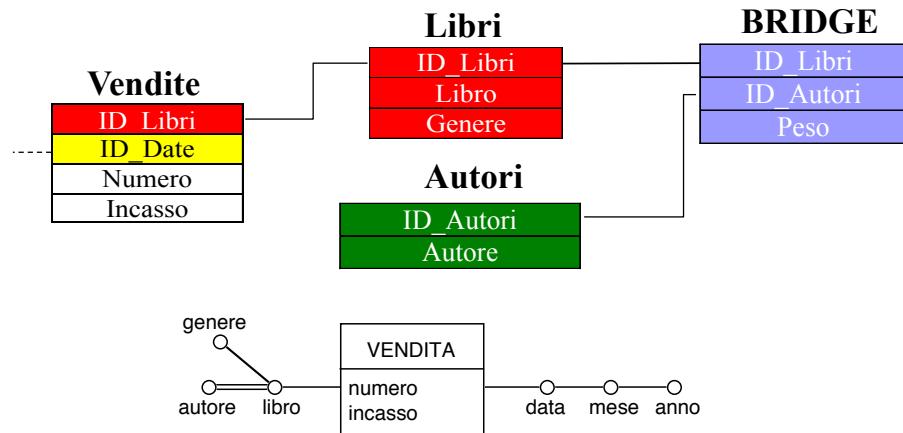
- Dal punto di vista concettuale, un attributo cross-dimensional b definisce un' associazione molti-a-molti tra due o più attributi dimensionali a_1, \dots, a_m
- La sua traduzione a livello logico richiede l' inserimento di una nuova tabella che includa b e abbia come chiave gli attributi a_1, \dots, a_m



60

Archi multipli

- La soluzione progettuale più ovvia è quella di inserire una tabella aggiuntiva (*bridge table*) che modelli l' arco multiplo:
 - ✓ La chiave della bridge table è composta dalla combinazione degli attributi collegati all' arco multiplo
 - ✓ Un eventuale attributo *peso* può permettere di attribuire importanza diversa alle tuple partecipanti



61

Archi multipli

LIBRO

<u>chiaveL</u>	libro	genere
1	Il DFM	tecnico
2	Mi Sembra Logico	tecnico
3	La Giusta Misura	attualità
4	Un Fatto Come e Perchè	attualità
5	La Quarta Dimensione	fantascienza

AUTORE

<u>chiaveA</u>	autore
1	Matteo Golfarelli
2	Stefano Rizzi

BRIDGE_AUTORE

<u>chiavel</u>	<u>chiaveA</u>	<u>peso</u>
1	1	0,5
1	2	0,5
2	1	1,0
3	2	1,0
4	1	0,5
4	2	0,5
5	1	1,0

VENDITE

<u>chiavel</u>	<u>chiaveD</u>	numero	incasso
1	1	3	150
2	1	5	250
3	1	10	300
4	1	4	80
5	1	8	400

62

Archi multipli

- Possono essere necessari sino a 3 join per recuperare tutte le informazioni contenute nella gerarchia
- La soluzione con bridge table rende possibili due tipi di interrogazioni:
 - ✓ **Interrogazioni pesate:** considerano il peso dell' arco multiplo e forniscono pertanto l' effettivo totale

Incasso di ciascun autore

```
SELECT AUTORI.Autore, sum(VENDITE.Incasso * BRIDGE.Peso)
FROM AUTORI, BRIDGE, LIBRI, VENDITE
WHERE AUTORI.ID_Autori = BRIDGE.ID_Autori
AND BRIDGE.ID_Libri = LIBRI.ID_Libri
AND LIBRI.ID_Libri = VENDITE.ID_Libri
GROUP BY AUTORI.Autore
```

63

Archi multipli

- Possono essere necessari sino a 3 join per recuperare tutte le informazioni contenute nella gerarchia
- La soluzione con bridge table rende possibili due tipi di interrogazioni:
 - ✓ **Interrogazioni pesate:** considerano il peso dell' arco multiplo e forniscono pertanto l' effettivo totale
 - ✓ **Interrogazioni di impatto:** non considerano il peso e perciò forniscono valori più elevati

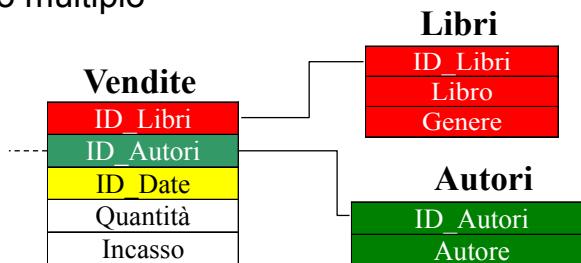
Copie vendute per ogni autore

```
SELECT AUTORI.Autore, sum(VENDITE.Quantità)
FROM AUTORI, BRIDGE, LIBRI, VENDITE
WHERE AUTORI.ID_Autori = BRIDGE.ID_Autori
AND BRIDGE.ID_Libri = LIBRI.ID_Libri
AND LIBRI.ID_Libri = VENDITE.ID_Libri
GROUP BY AUTORI.Autore
```

64

Archi multipli

- Nel caso si voglia continuare a utilizzare lo schema a stella è necessario rendere più fine la granularità del fatto modellando così l' arco multiplo direttamente nella fact table (*push-down*)
- Questa soluzione richiede l' aggiunta alla fact table di una nuova dimensione corrispondente all' attributo terminale dell' arco multiplo



65

Archi multipli: comparazione

- Il potere informativo delle due soluzioni è identico
- Con la **soluzione con push-down**:
 - ✓ Si introduce una forte ridondanza nella fact-table le cui righe devono essere replicate tante volte quante sono le corrispondenze dell' arco multiplo
 - ✓ Il peso è codificato permanentemente all' interno della fact table e il suo aggiornamento può risultare molto complesso
 - ✓ Le interrogazioni di impatto risultano molto complesse
 - ✓ Il costo di esecuzione delle interrogazioni si riduce grazie al minor numero di join necessari
 - ✓ Il calcolo degli eventi pesati avviene durante l' alimentazione
- Con la **soluzione con bridge-table**:
 - ✓ Il costo di esecuzione delle interrogazioni si riduce a causa del minor numero di tuple coinvolte
 - ✓ Il calcolo degli eventi pesati avviene durante l' interrogazione

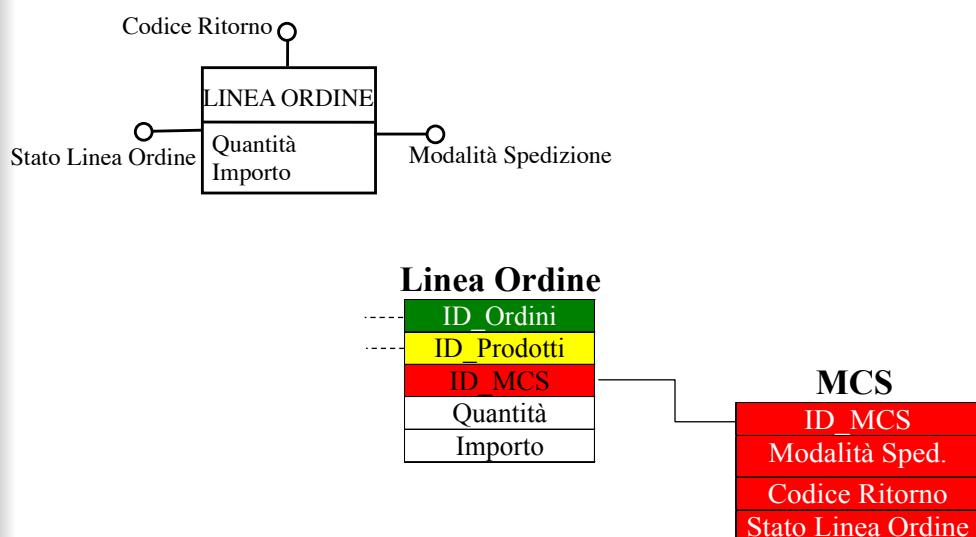
66

Dimensioni degeneri

- Questo termine indica una dimensione la cui gerarchia contiene un solo attributo
- Se la lunghezza dell' attributo non è eccessiva può convenire evitare la creazione di una specifica dimension table importando direttamente i valori dell' attributo nella fact table
- Una soluzione alternativa è quella di utilizzare un' unica dimension table per modellare più dimensioni degeneri (**junk dimension**)
 - ✓ In una junk dimension non esiste alcuna dipendenza funzionale tra gli attributi per cui risultano valide tutte le possibili combinazioni di valori
 - ✓ Questa soluzione risulta attuabile solo quando il numero di valori distinti per gli attributi coinvolti è limitato

67

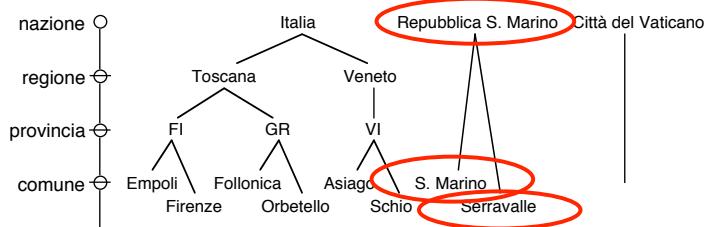
Dimensioni degeneri



68

Gerarchie incomplete

- Questo termine indica una gerarchia in cui per alcune istanze risultano assenti uno o più livelli di aggregazione
- Vengono gestite a livello estensionale inserendo opportuni valori fittizi
- Il problema è più complesso rispetto al caso degli attributi opzionali poiché la mancanza di un valore di un attributo non implica la mancanza dei successivi nella gerarchia di aggregazione



- È necessario mantenere la consistenza rispetto all' operatore di roll-up
- Sono possibili più soluzioni che si differenziano per il tipo di segnaposto inserito

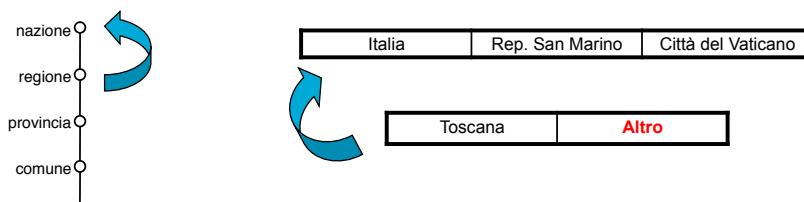
69

Gerarchie incomplete

- Bilanciamento per esclusione:** in tutte le tuple viene inserito un segnaposto generico (es. "altro")

nazione	Italia	Rep. San Marino	Rep. San Marino	Città del Vaticano
regione	Toscana	Altro	Altro	Altro
provincia	Firenze	Altro	Altro	Altro
comune	Empoli	San Marino	Serravalle	Altro

- Preferibile quando il numero di dati mancanti è elevato
- Questa soluzione viola la semantica del roll-up poiché aggregando i dati si avrà un maggior livello di dettaglio delle informazioni



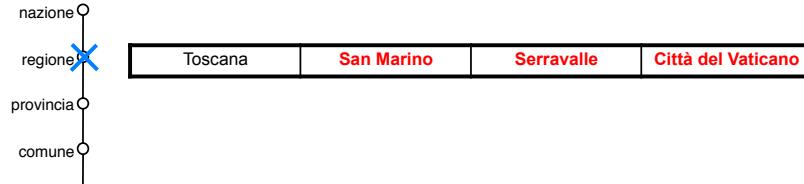
70

Gerarchie incomplete

- **Bilanciamento verso il basso:** i valori mancanti vengono rimpiazzati con il valore dell' attributo che lo precede nella gerarchia



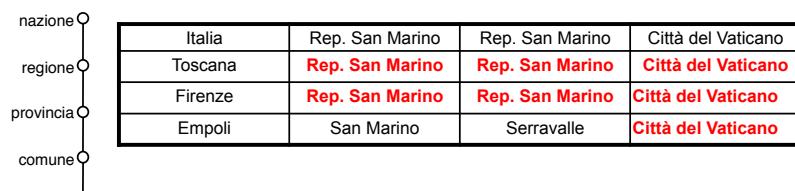
- Preferibile quando il numero di dati mancanti è limitato
- L' interpretazione dei report è complicata dal fatto che appariranno valori non corrispondenti al livello di aggregazione prescelto



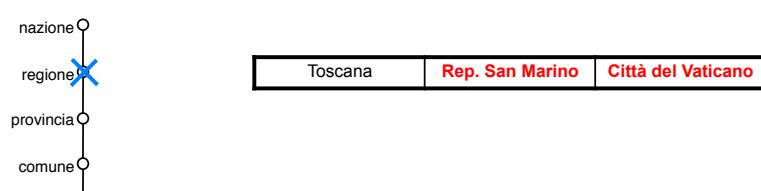
71

Gerarchie incomplete

- **Bilanciamento verso l' alto:** i valori mancanti vengono rimpiazzati con i valori dell' attributo che lo segue nella gerarchia



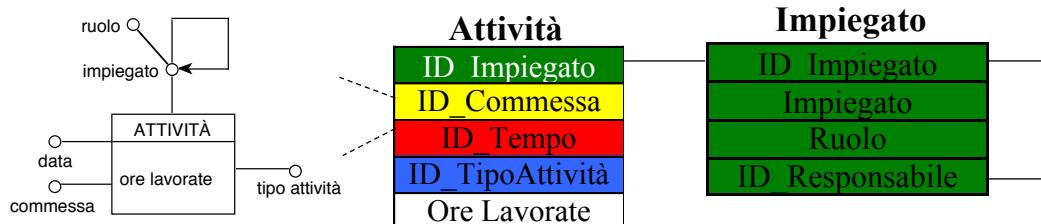
- Preferibile quando il numero di dati mancanti è elevato
- Rispetto alla soluzione precedente i report risultano più leggibili perché presentano un numero inferiore di valori



72

Gerarchie ricorsive

- Questo termine indica una gerarchia in cui il numero dei livelli di aggregazione non è codificabile nello schema e può variare da istanza a istanza
- Non può essere modellata tramite schema a stella
- Una possibile soluzione prevede l'utilizzo di un autoanello

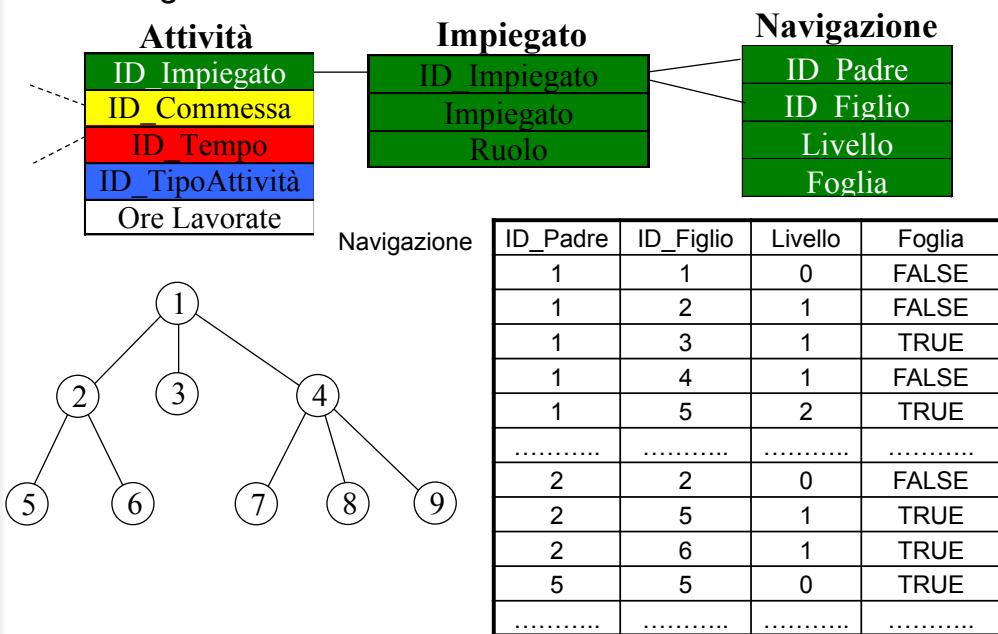


- Non sempre è gestibile in modo ottimale con DBMS commerciali
- SQL non è un linguaggio ricorsivo

73

Gerarchie ricorsive

- Una soluzione più potente prevede di appiattire la gerarchia esplicitando tutti i legami da essa indotti in una tabella di navigazione



74

Gerarchie ricorsive

- La dimensione della tabella di navigazione cresce in modo esponenziale con la profondità della gerarchia
- Se la dimensione della tabella è trattabile questa soluzione garantisce un maggiore potere espressivo
- Per **descendere** la gerarchia:

Il totale delle ore lavorate dal gruppo di cui è responsabile il sig. Rossi

```
SELECT sum(ore lavorate)
FROM ATTIVITA A, IMPIEGATO I, NAVIGAZIONE N
WHERE I.Nome= 'Rossi' AND I.ID_Impiegato=N.ID_Padre
AND N.ID_Figlio = A.ID_Impiegato;
```

Il totale delle ore lavorate dai subordinati diretti dal sig. Rossi

```
SELECT sum(ore lavorate)
FROM ATTIVITA A, IMPIEGATO I, NAVIGAZIONE N
WHERE I.Nome= 'Rossi' AND I.ID_Impiegato=N.ID_Padre
AND N.ID_Figlio = A.ID_Impiegato AND N.Livello=1;
```

75

Gerarchie ricorsive

- La dimensione della tabella di navigazione cresce in modo esponenziale con la profondità della gerarchia
- Se la dimensione della tabella è trattabile questa soluzione garantisce un maggiore potere espressivo.
- Per **risalire** la gerarchia:

Il totale delle ore lavorate dai responsabili del sig. Rossi

```
SELECT sum(ore lavorate)
FROM ATTIVITA A, IMPIEGATO I, NAVIGAZIONE N
WHERE I.Nome= 'Rossi' AND I.ID_Impiegato=N.ID_Figlio
AND N.ID_Padre = A.ID_Impiegato;
```

- Escludendo dai join la tabella di navigazione si continua ad avere uno schema a stella

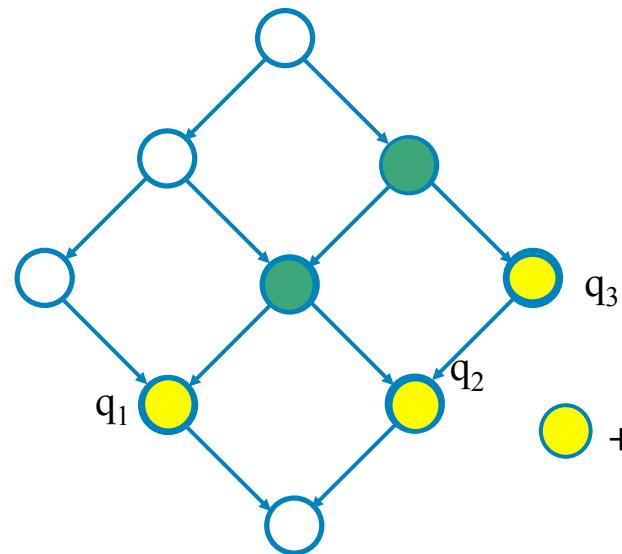
76

Scelta delle viste

- La scelta delle viste da materializzare è un compito complesso, la soluzione rappresenta un trade-off tra numerosi requisiti in contrasto:
 1. Minimizzazione di funzioni di costo
 2. Vincoli di sistema
 - ✓ Spazio su disco
 - ✓ Tempo a disposizione per l' aggiornamento dei dati
 3. Vincoli utente
 - ✓ Tempo massimo di risposta
 - ✓ Freschezza dei dati

77

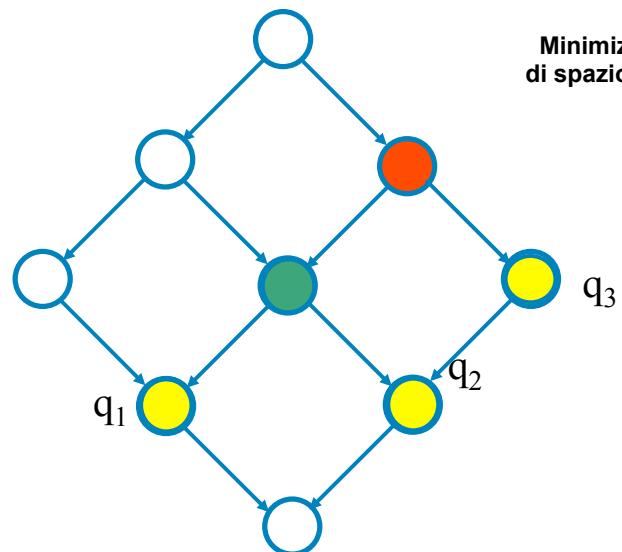
Scelta delle viste



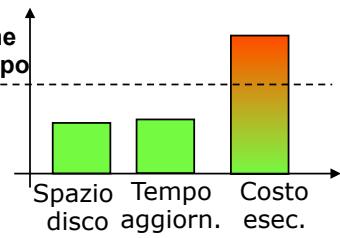
viste candidate,
ossia potenzialmente
utili a ridurre il costo
di esecuzione del
carico di lavoro

78

Scelta delle viste

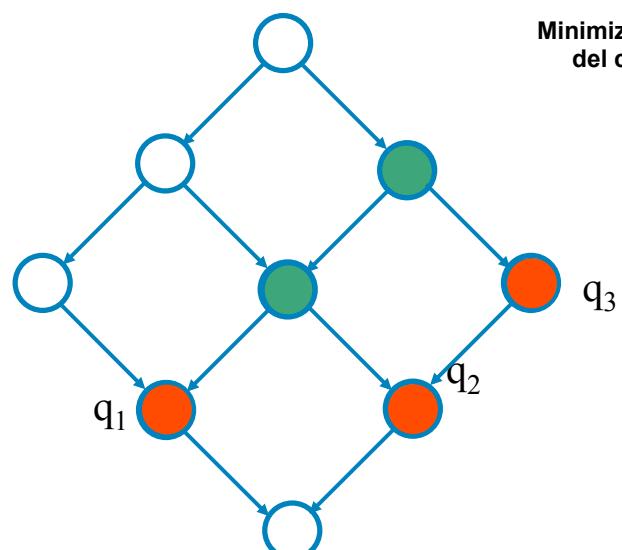


Minimizzazione
di spazio e tempo

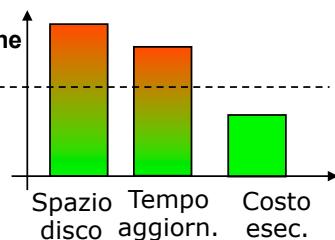


79

Scelta delle viste

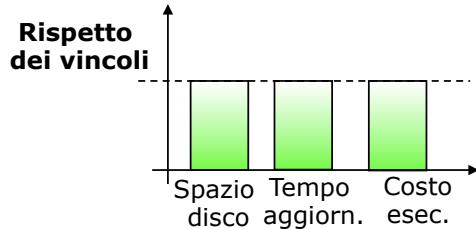
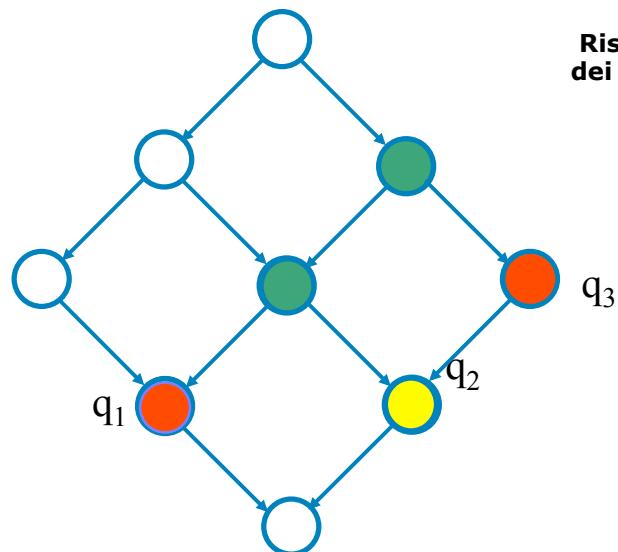


Minimizzazione
del costo



80

Scelta delle viste



81

Scelta delle viste

- È utile materializzare una vista quando:
 - ✓ Risolve direttamente una interrogazione frequente
 - ✓ Permette di ridurre il costo di esecuzione di molte interrogazioni

- Non è consigliabile materializzare una vista quando:
 - ✓ Il suo group-by set è molto simile a quello di una vista già materializzata
 - ✓ Il suo group-by set è molto fine
 - ✓ La materializzazione non riduce di almeno un ordine di grandezza il costo delle interrogazioni

82



Frammentazione delle viste

- Con il termine frammentazione si intende la suddivisione delle fact table (primarie e secondarie) in più frammenti al fine di aumentare le prestazioni del sistema.
- Le specifiche caratteristiche dei DW (ridondanza dei dati, cubi correlati, ecc.) rendono particolarmente utile questa forma di ottimizzazione.
 - ✓ **Frammentazione orizzontale:** la relazione viene suddivisa in più parti, ognuna delle quali contiene tutti gli attributi ma solo una parte delle tuple di quella di origine.
 - ✓ **Frammentazione verticale:** la relazione viene suddivisa in più parti, ognuna delle quali contiene tutte le tuple ma solo una parte degli attributi di quella di origine.

83

Frammentazione orizzontale

- È la forma di frammentazione maggiormente utilizzata.
- I criteri di selezione delle tuple da inserire nei frammenti sono determinati in base alle condizioni di selezione maggiormente utilizzate a uno specifico livello di aggregazione.
- L'attributo maggiormente utilizzato a tal fine è il tempo che, oltre a essere largamente coinvolto nelle interrogazioni, permette una facile gestione degli aggiornamenti.
- La riduzione dei tempi di esecuzione delle interrogazioni è dovuta alla possibilità di operare su fact table più piccole e su cui è già stata operata una (parziale) selezione.
- A differenza della frammentazione verticale quella orizzontale non comporta alcun costo aggiuntivo in termini di spazio richiesto per la memorizzazione dei dati.

84

Frammentazione verticale

- La frammentazione verticale costituisce una soluzione più specializzata al problema della materializzazione delle viste
- Per ogni cubo e per ogni livello di aggregazione è possibile materializzare solo le misure utili per uno specifico carico di lavoro

Per esempio, sarà molto utile conoscere il valore dell' IVA da versare aggregandola in base al periodo di pagamento (mese o trimestre), mentre ne sarà richiesto raramente il valore per altri periodi

- La frammentazione verticale:
 - ✓ Può richiedere spazio aggiuntivo per la memorizzazione dei dati a causa delle replicazioni dei campi chiave della fact table
 - ✓ Determina un risparmio di spazio rispetto alla materializzazione di viste ognqualvolta si evita di materializzare una misura

85

Progettazione dell'ETL

Progettazione dell'ETL

- Durante la fase di progettazione dell'ETL vengono definite le procedure necessarie a caricare all'interno del data mart i dati provenienti dalle sorgenti operazionali.
 - ✓ **Dalle sorgenti operazionali al livello riconciliato:** realizzano a livello estensionale le trasformazioni definite nella fase di integrazione
 - ✓ **Dal livello riconciliato al livello del data mart:** si definiscono le procedure che permettono di conformare la struttura dei dati del livello riconciliato agli schemi a stella utilizzati in ambito multidimensionale

87

Alimentazione dello schema riconciliato

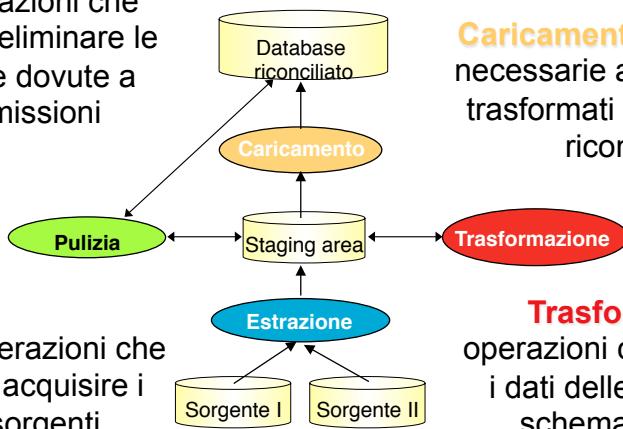
Staging area: spazio utilizzato per memorizzare in via transitoria le informazioni necessarie all'esecuzione delle procedure

Pulizia: operazioni che permettono di eliminare le incongruenze dovute a errori e omissioni

Caricamento: operazioni necessarie a inserire i dati trasformati nel database riconciliato

Estrazione: operazioni che permettono di acquisire i dati dalle sorgenti

Trasformazione: operazioni che conformano i dati delle sorgenti allo schema riconciliato



88

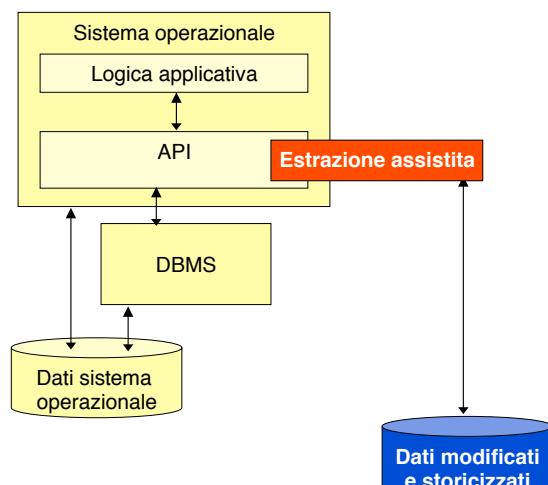
Estrazione dei dati

- Le operazioni di estrazione dipendono dalla natura dei dati presenti nelle sorgenti operazionali
 - ✓ **Transitoria:** il sistema mantiene solo l' immagine corrente sovrascrivendo i dati che non sono più validi (es. dati di inventario, scorte di magazzino)
 - ✓ **Semi-storicizzata:** il sistema mantiene un limitato numero degli stati precedenti e non è possibile determinare per quanto tempo ciascun dato verrà conservato nel sistema
 - ✓ **Storicizzata:** tutte le modifiche intervenute nei dati vengono mantenute in un intervallo di tempo ben definito (es. dati bancari e assicurativi)
- L' estrazione può essere
 - ✓ **Statica:** il livello riconciliato viene ricreato ex-novo
 - ✓ **Incrementale:** vengono aggiunti solo i dati prodotti dal sistema operazionale nell' intervallo di tempo intercorso dall' ultimo caricamento
 - Immediata
 - Ritardata

89

Estrazione dei dati

- Estrazione assistita dall' applicazione
 - ✓ Tecnica di estrazione immediata
 - ✓ Le modifiche vengono rilevate da specifiche funzioni implementate direttamente all' interno delle applicazioni
 - ✓ Richiedono la modifica delle applicazioni OLTP
 - ✓ È utile quando si lavora con sistemi legacy che non forniscono sistemi di triggering, log, ecc.
 - ✓ Trova applicazione anche nei sistemi moderni quando è disponibile un livello di API comuni a tutte le applicazioni: una sola modifica per tutti gli accessi di uno stesso tipo

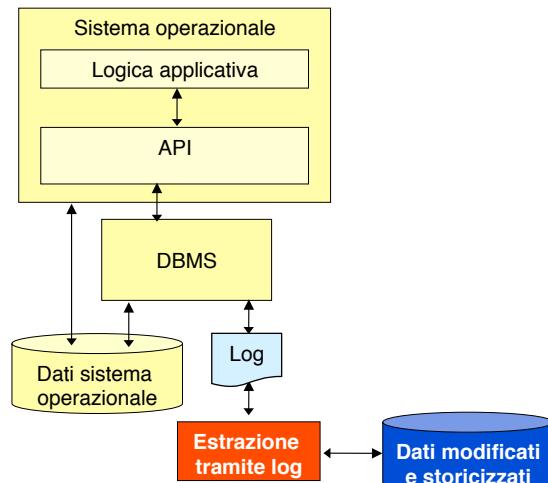


90

Estrazione dei dati

■ Estrazione basata su log

- ✓ Tecnica di estrazione ritardata
- ✓ Le modifiche vengono memorizzate in appositi file prodotti dal DBMS
- ✓ Può risultare molto complesso interpretare il contenuto dei file il cui formato è normalmente proprietario dello specifico DBMS
- ✓ È consigliabile solo quando il modulo di estrazione è fornito direttamente dal produttore del DBMS

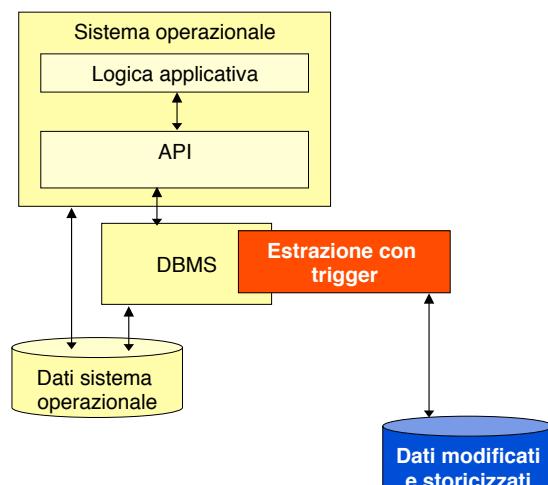


91

Estrazione dei dati

■ Estrazione basata su trigger

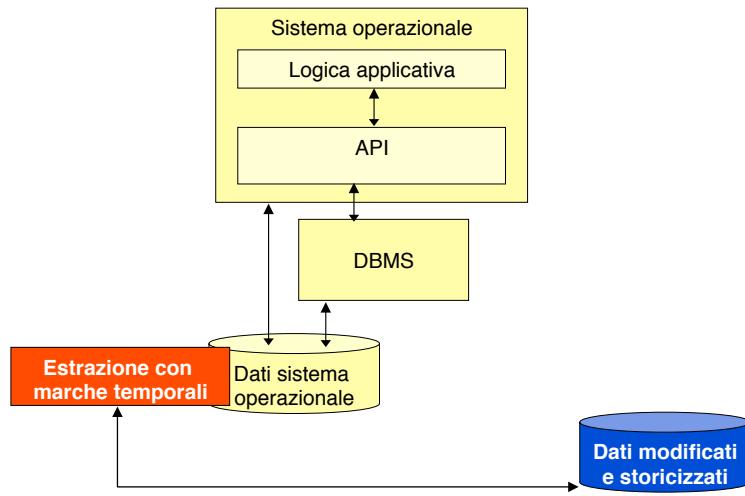
- ✓ Tecnica di estrazione immediata
- ✓ Le modifiche vengono individuate mediante funzioni basate su eventi implementate e controllate direttamente nel DBMS
- ✓ Per motivi prestazionali non è possibile adottare in modo estensivo questa tecnica che richiederebbe al DBMS di monitorare continuamente tutte le transazioni potenzialmente in grado di innescare un trigger



92

Estrazione dei dati

- Estrazione basata su marche temporali
 - ✓ Tecnica di estrazione ritardata
 - ✓ Prevede la modifica dello schema del database che dovrà contenere uno o più campi necessari a contrassegnare i record modificati
 - ✓ Il modulo di estrazione opera a posteriori individuando il tipo di modifica subita dai dati



93

Estrazione dei dati

- L'efficacia della tecnica basata su marche temporali dipende dalla struttura stessa del sistema operazionale:

se i dati sono transitori o semi-storici l'estrazione basata su marche temporali non può identificare gli stati intermedi di quei record modificati più volte durante l'intervalllo di aggiornamento

Situazione al 1/4/2002

Cod	prodotto	cliente	qtà	Data
1	Greco di tufo	Malavasi	50	15/3/2002
2	Barolo	Maio	100	1/4/2002
...



Estratto 1/4/2002

Situazione al 2/4/2002

Cod	prodotto	cliente	qtà	Data
1	Greco di tufo	Malavasi	50	15/3/2002
2	Barolo	Maio	200	2/4/2002
...

94

Estrazione dei dati

- L'efficacia della tecnica basata su marche temporali dipende dalla struttura stessa del sistema operazionale:

se i dati sono transitori o semi-storicizzati l'estrazione basata su marche temporali non può identificare gli stati intermedi di quei record modificati più volte durante l'intervallo di aggiornamento

Situazione al 3/4/2002

Cod	prodotto	cliente	qtà	Data
1	Greco di tufo	Malavasi	50	15/3/2002
2	Barolo	Maio	150	3/4/2002
...

→ Estratto 3/4/2002

Situazione al 2/4/2002

Cod	prodotto	cliente	qtà	Data
1	Greco di tufo	Malavasi	50	15/3/2002
2	Barolo	Maio	200	2/4/2002
...

← Modifica persa

95

Il risultato dell'estrazione

- Qualunque tecnica incrementale si utilizzi, il risultato della fase di estrazione consiste nell'insieme di record della sorgente modificati, aggiunti o cancellati rispetto alla precedente esecuzione della procedura di estrazione
- I dati risiedono nella staging area
- Per facilitare le fasi successive è opportuno associare a ogni record estratto il tipo di operazione (Inserimento, Modifica, Cancellazione) che ne ha generato la variazione

96

Il risultato dell' estrazione

Situazione al 4/4/2002

cod	prodotto	cliente	qtà
1	Greco di tufo	Malavasi	50
2	Barolo	Maio	150
3	Barbera	Lumini	75
4	Sangiovese	Cappelli	45

Situazione al 6/4/2002

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	oper
3	Barbera	Lumini	75	C
4	Sangiovese	Cappelli	145	M
5	Vermentino	Maltoni	25	I
6	Trebbiano	Maltoni	150	I

97

Caricamento dei dati

- La modalità di caricamento dei dati dalla staging area al database riconciliato dipende dalla tecnica utilizzata in fase di estrazione e dal livello di storizziazione del livello riconciliato
 - ✓ Estrazione statica → Riscrittura completa
 - ✓ Estrazione incrementale
 - Livello riconciliato non storizzato: memorizzo solo il tipo di operazione che ha determinato la variazione
 - Livello riconciliato storizzato: memorizzo anche una coppia di marche temporali che indicano l' intervallo di validità della tupla

Il livello di storizziazione dello schema riconciliato dipende da quello delle sorgenti operazionali e dai requisiti utente relativi alla reportistica operativa

98

Trasformazione e pulizia

- L'insieme delle operazioni atte a garantire la correttezza e la consistenza dei dati presenti nel livello riconciliato rispetto a:
 - ✓ Errori di battitura
 - ✓ Differenza di formato dei dati nello stesso campo
 - ✓ Inconsistenza tra valori e descrizione dei campi
 - Evoluzione del modo di operare dell' azienda
 - Evoluzioni della società
 - Convenzioni interne ai reparti e diverse da quelle generali del sistema informativo
 - ✓ Inconsistenza tra valori di campi correlati
 - Città='Bologna' Regione='Lazio'

La maggior parte delle inconsistenti può essere prevenuta rendendo più rigorose le regole di inserimento dei dati nelle applicazioni del sistema operazionale

99

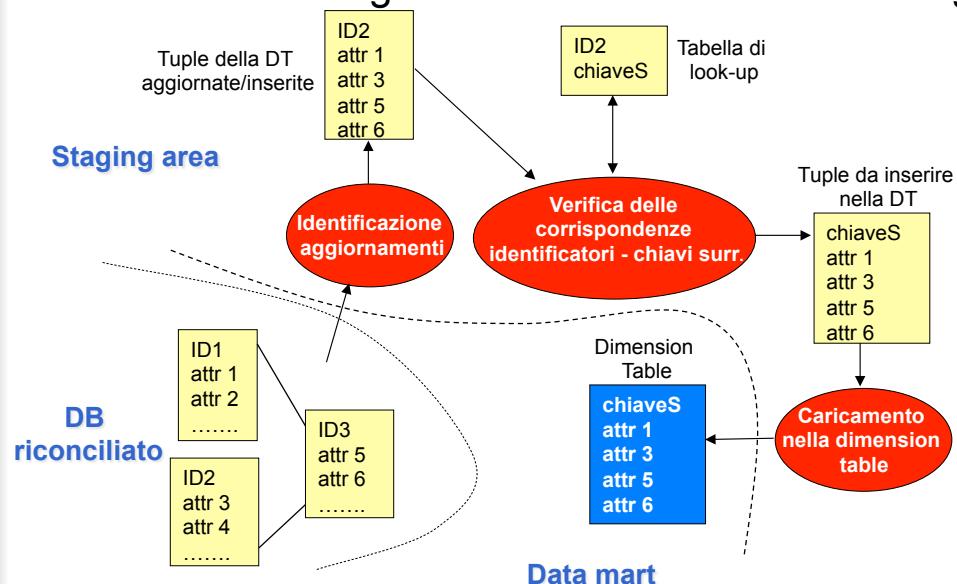
Trasformazione e pulizia

- Ogni problema richiede una tecnica specifica per la soluzione e molti sistemi commerciali propongono moduli specifici per la pulizia dei dati
 - ✓ Tecniche basate su dizionari: utilizzano tabelle di look-up per identificare ed eliminare sinonimi e abbreviazioni
 - Utilizzabili solo quando il dominio dell' attributo è conosciuto e limitato
 - Utili per errori di battitura e discrepanze di formato
 - ✓ Tecniche ad hoc: ogni dominio applicativo ha regole proprie, troppo specifiche per essere verificate tramite strumenti standard
 - Equazioni: *profitto = guadagno - spese*
 - Outliers: *variazione di prezzo di oltre il 20%*
 - ✓ Tecniche di fusione approssimata: permettono di identificare record corrispondenti in assenza di identificatori comuni
 - Join approssimati
 - Purge/merge problem

100

Alimentazione delle dimension table

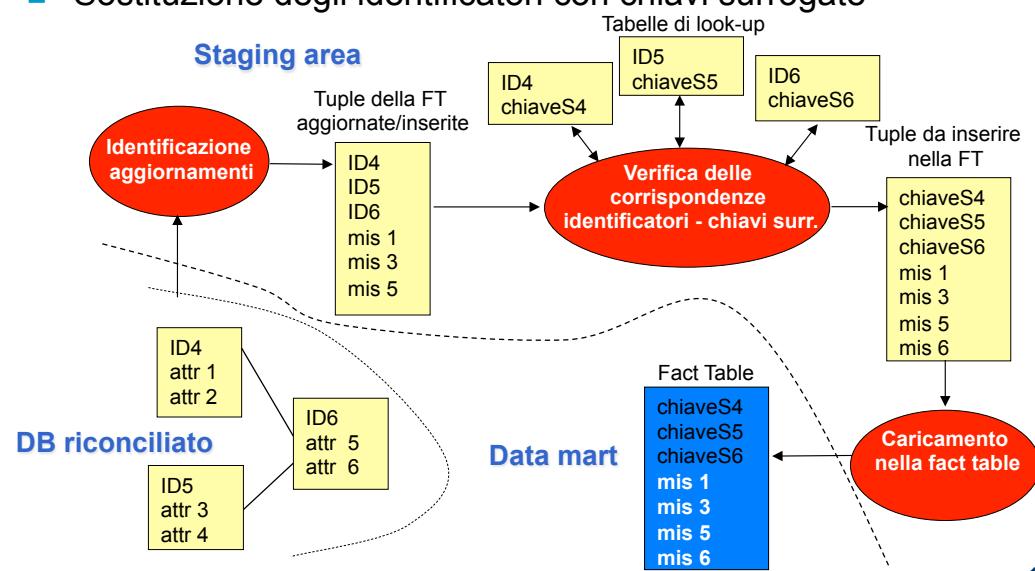
- Identificazione dei dati da caricare
- Sostituzione degli identificatori con chiavi surrogate



101

Alimentazione delle fact table

- Segue l' alimentazione delle dimension table per poter rispettare i vincoli di integrità referenziale
- Identificazione dei dati da caricare
- Sostituzione degli identificatori con chiavi surrogate

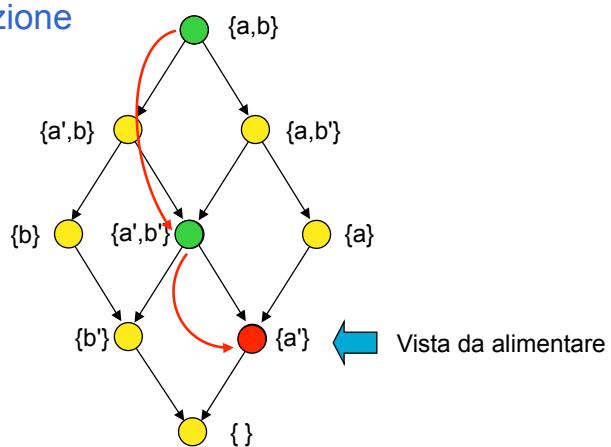


102

Alimentazione delle viste

- Scelta della vista aggiornata che minimizza il costo di aggiornamento

✓ È la più piccola vista che permette di risolvere l'interrogazione



- Attenzione alla corretta scelta dell'operatore di aggregazione

103

Progettazione fisica

Indici per il Data Mart

- Le specifiche caratteristiche dei data mart permettono di utilizzare classi di indici diverse dal ben noto B⁺-tree utilizzato nella maggior parte dei DBMS commerciali.
 - ✓ Accessi in sola lettura
 - ✓ Aggiornamento periodico dei dati con possibilità di riorganizzazione degli indici
 - ✓ Accessi ad ampie porzioni di dati
- Alcuni indici sono nati in conseguenza delle esigenze di data warehousing
- Altri erano preesistenti ma non venivano utilizzati

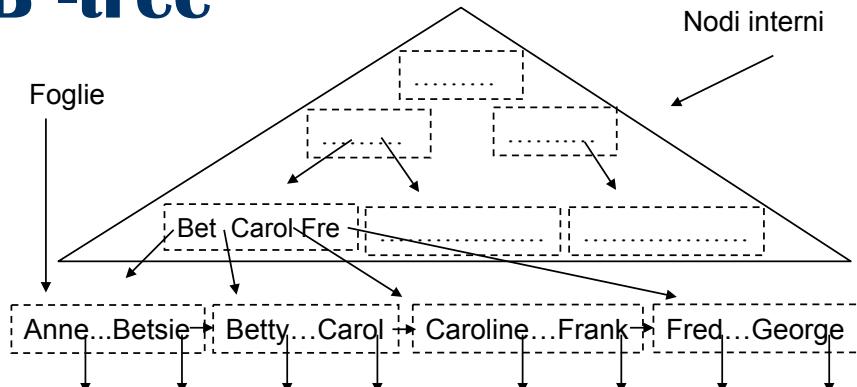
Indici Bitmap

Indici di join

Indici Star

105

I B⁺-tree



- Le **foglie** contengono tutti i valori di chiave
- I **nodi interni**, organizzati come un B-tree, costituiscono solo una mappa per consentire una rapida localizzazione delle chiavi, e memorizzano dei separatori

106

Perché i B⁺-tree non sono più sufficienti ?

- Forniscono buone prestazioni quando la selettività dei predicati è molto elevata.
 - ✓ Le interrogazioni OLAP utilizzano spesso predicati a bassa selettività (es. sesso)
- Sono più adatti a interrogazioni semplici
- Possono richiedere molto spazio per la loro memorizzazione

I B⁺-tree non sono più sufficienti ma rimangono ancora molto utili

107

Gli indici bitmap

- Un indice bitmap su un attributo è composto da una matrice di bit contenente:
 - ✓ Tante righe quante sono le tuple della relazione
 - ✓ Tante colonne quanti sono i valori distinti di chiave dell' attributo
- Il bitmap (i,j) è posto a TRUE se nella tupla i -esima è presente il valore j -esimo

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

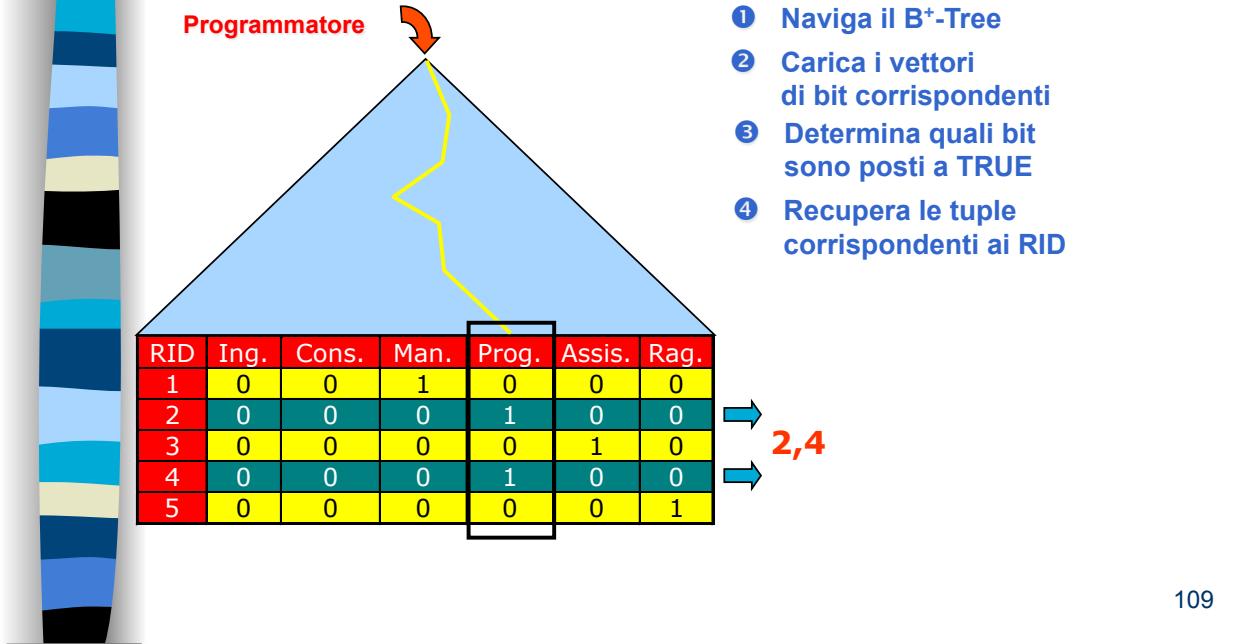
L' impiegato corrispondente al RID 1 è un Manager

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

108

Implementazione dei bitmap

- Normalmente i bitmap sono associati a B⁺-Tree le cui foglie contengono vettori di bit invece di RID



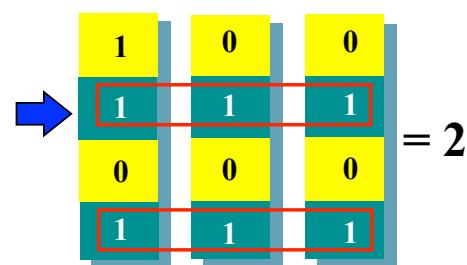
109

I vantaggi degli indici bitmap

- Lo spazio richiesto su disco può essere molto ridotto
- I/O è molto basso poiché vengono letti solo i vettori di bit necessari
- Ottimi per interrogazioni che non richiedono l'accesso ai dati
- Permettono l'utilizzo di operatori binari per l'elaborazione dei predicati

Esempio: “*Quanti maschi in Emilia-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



110

Occupazione su disco

- Gli indici bitmap sono adatti ad attributi con una ridotta cardinalità poiché ogni nuovo valore distinto di chiave richiede un ulteriore vettore di bit
- All'aumentare del numero di chiavi distinte aumenta la sparsità della matrice

Esempio:

$NR = 10.000.000$

$Len(Pointer) = 4 \times 8 \text{ bit}$

B-tree

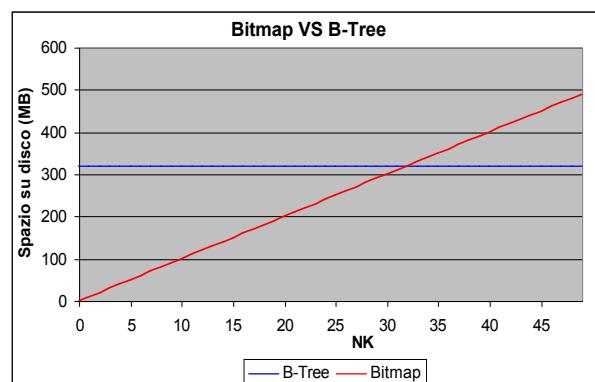
Bitmap

$NR \times Len(Pointer)$

$NR \times NK \times 1 \text{ bit}$

Si ha un risparmio di spazio se:

$$\text{Densità media} \geq \frac{1}{Len(RID)}$$

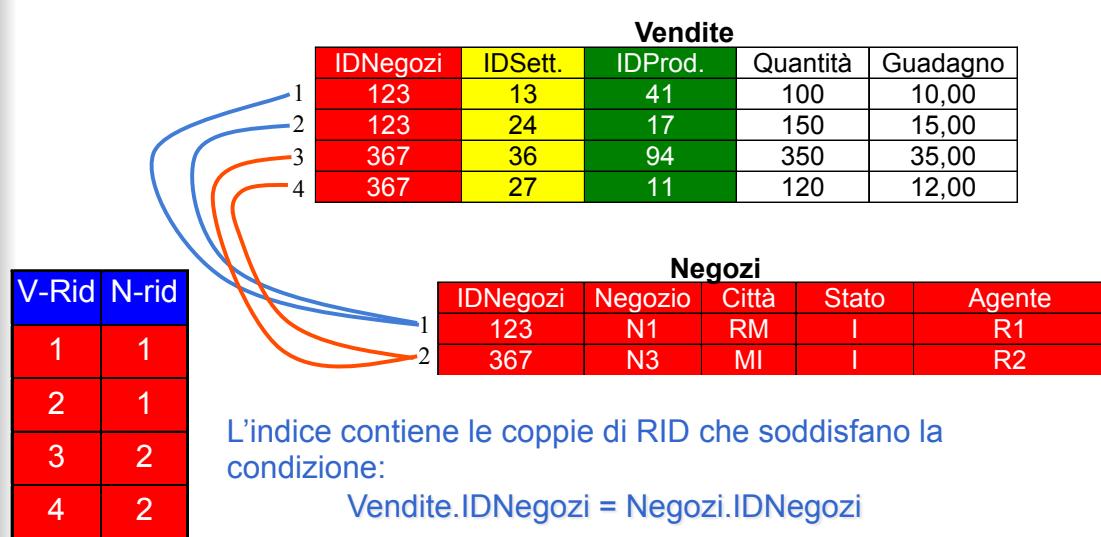


La compressione delle matrici riduce il fattore di crescita della dimensione

111

Gli indici di join

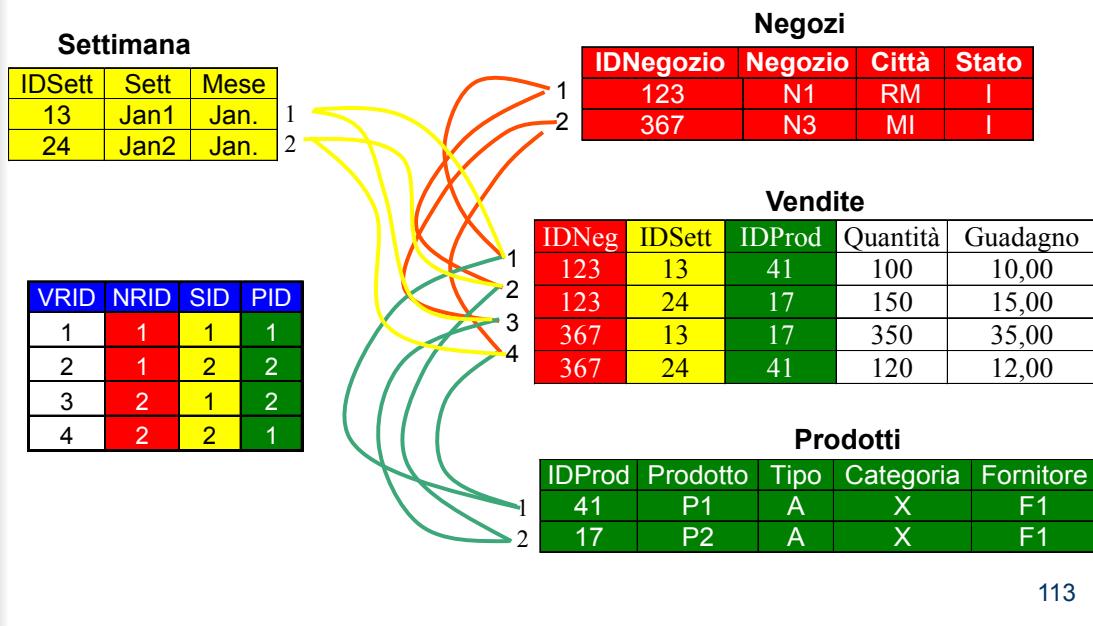
- Le interrogazioni su schemi a stella richiedono sempre uno o più join
- Un indice di join calcola in anticipo le tuple che soddisfano un particolare predicato di join



112

Gli indici a stella

- Estendono il concetto di indice di join a più tabelle concatenando i valori delle colonne della fact table e di più dimension table



Gli indici a stella

- Rappresentano esplicitamente l' aspetto multidimensionale dei dati e dipendono fortemente dall' ordinamento delle colonne.
- Sono molto efficienti quando utilizzati in interrogazioni che coinvolgono tutte o le sole colonne iniziali dell' indice.
- Forniscono prestazioni sub-ottime in caso contrario.

Il numero di indici a stella necessari a rispondere efficientemente a interrogazioni che coinvolgono un insieme arbitrario di dimensioni è funzione del numero di permutazioni dell' insieme di dimensioni

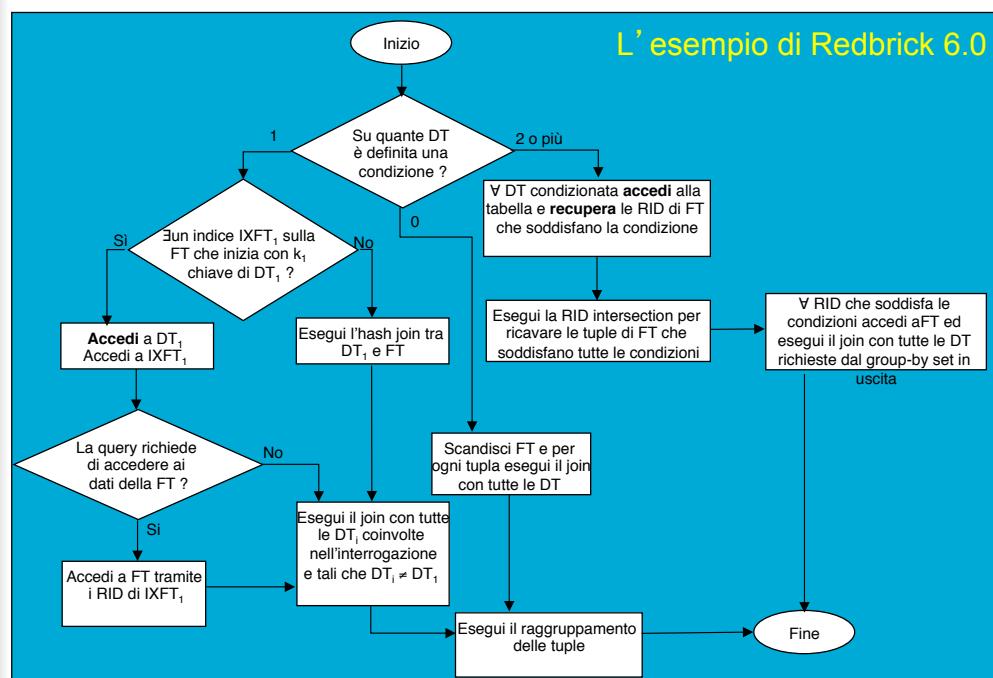
Si consiglia di creare indici su..

- Chiavi importate sulla fact table per aumentare la velocità di esecuzione dei join (B⁺-Tree o indici di join, indici star, bitmapped join index)
- Attributi dimensionali che sono spesso coinvolti nei criteri di selezione (B⁺-Tree o bitmap)
- Misure che sono spesso coinvolte in clausole di selezione (bitmap evoluti)

Se il DBMS non utilizza statistiche per definire il piano di accesso la creazione degli indici deve essere valutata con molta attenzione

115

Si consiglia di creare indici su..



116