

Introduzione al Data Warehousing

Molte di queste slide sono state realizzate dal

Prof. Stefano Rizzi
(<http://www-db.deis.unibo.it/~srizzi/>)

e sono state tratte dal suo libro

Data Warehouse - teoria e pratica della Progettazione
Autori: Matteo Golfarelli, Stefano Rizzi
Editore: McGraw-Hill

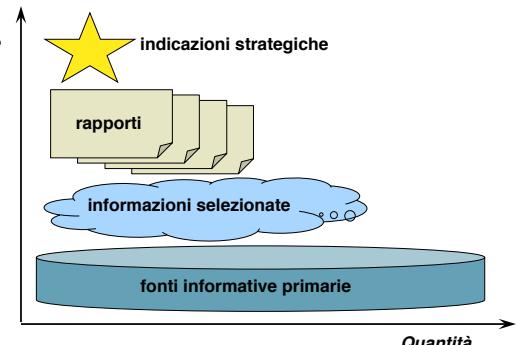
Business Intelligence

- Insieme di attività orientate a estrarre informazioni dai dati di business, solitamente quelli generati dai processi aziendali

- **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.



Data Warehousing

■ Decision Support System

insieme di tecniche e strumenti informatici atti a estrapolare informazioni da un insieme di dati memorizzati su supporti elettronici

- Tra i sistemi di supporto alle decisioni, l'approccio più usato prevede il progetto e la costruzione di un **data warehouse** a partire dal quale si procede con opportune tecniche di analisi dei dati o di estrazione di utili correlazioni fra i dati

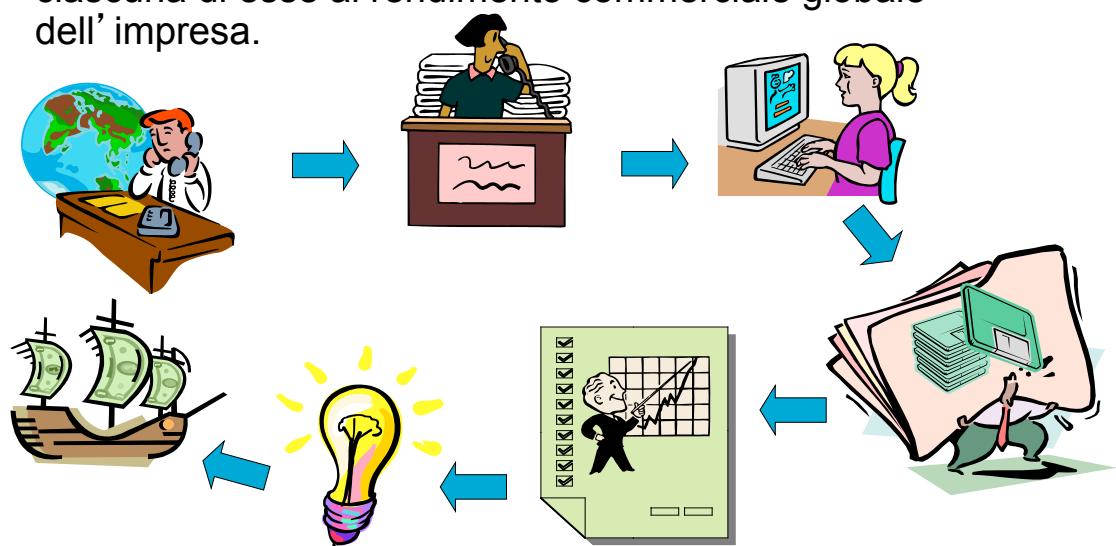
■ Data Warehousing:

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.

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...



5

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)
-

6

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 (www.kimballgroup.com)

7

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.

8

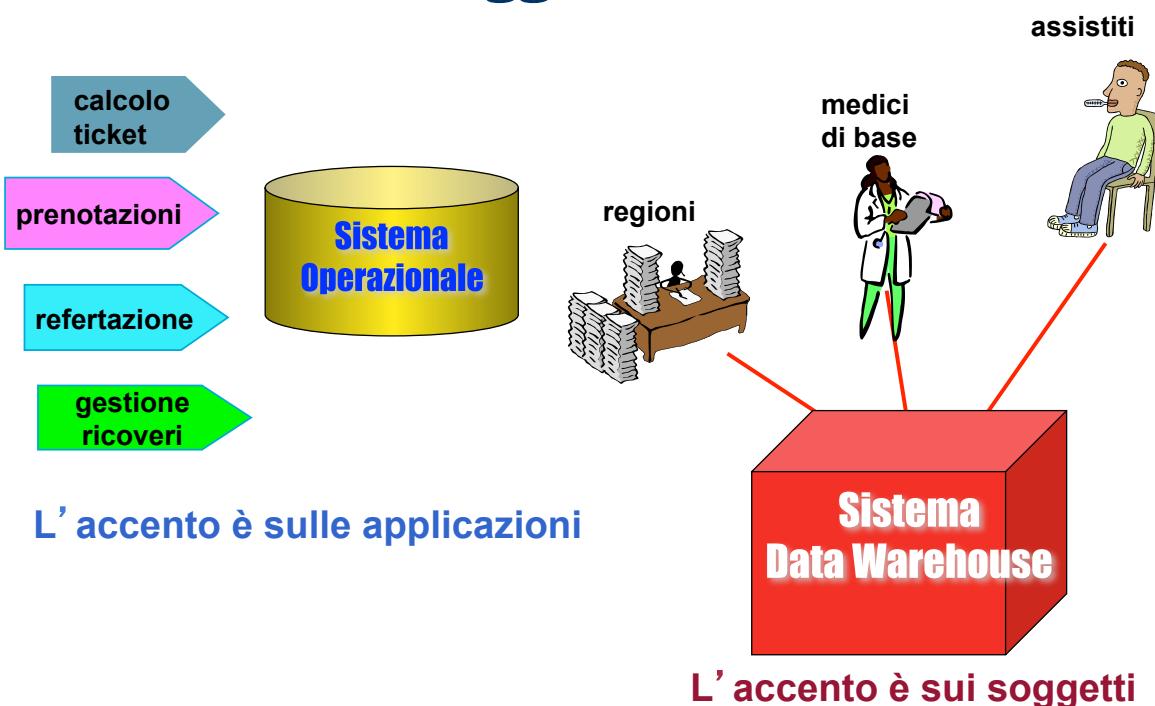
Il Data Warehouse

- 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.

9

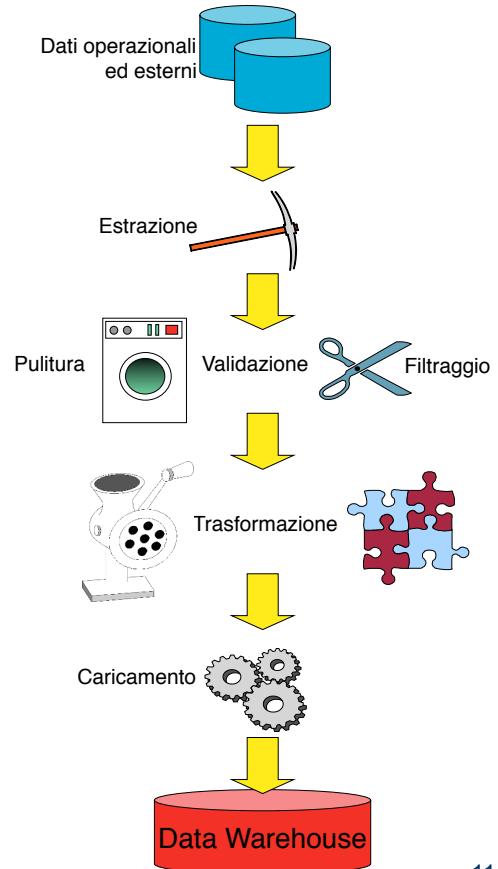
...orientato ai soggetti



10

...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.



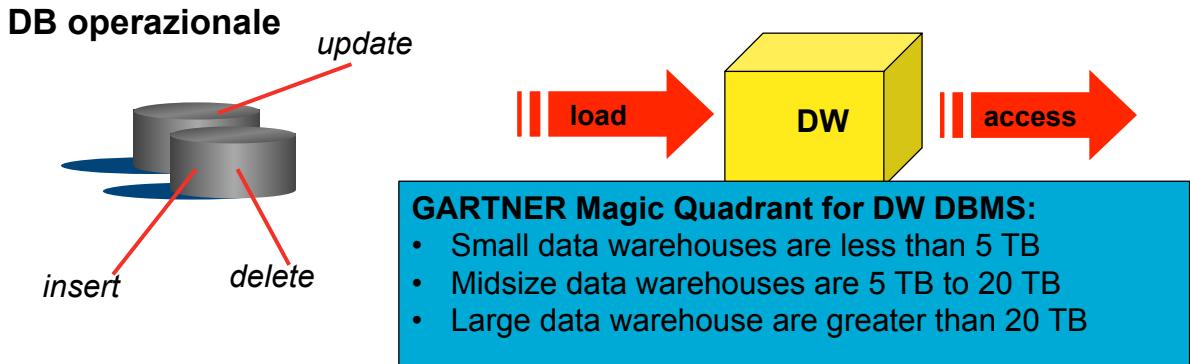
11

...rappresentativo dell'evoluzione temporale



12

...non volatile



- ✓ in un DW non si fanno generalmente update e insert quindi non sono necessarie tecniche sofisticate di gestione delle transazioni - come invece richiesto da un sistema operazionale.

13

Le interrogazioni: OLTP e OLAP

■ OLTP (On-Line Transactional Processing)

- ✓ Le interrogazioni eseguono transazioni che leggono e scrivono un ridotto numero di record da diverse tabelle
 - Transazioni, recovery, consistenza (scrittura e/o lettura)
 - Molte operazioni, elevato livello di concorrenza
- ✓ 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
 - Operazioni di sola lettura
 - Poche operazioni, basso livello di concorrenza
- ✓ L’ interattività è una caratteristica irrinunciabile delle sessioni di analisi e fa sì che il carico di lavoro effettivo vari continuamente nel tempo.

14

Separazione tra OLTP e OLAP

- Mescolare interrogazioni OLAP (analitiche) e OLTP (transazionali) porta a inevitabili rallentamenti che rendono insoddisfatti gli utenti di entrambe le categorie.
- L'idea di base del Data Warehouse è quella di separare l'elaborazione OLAP di tipo analitico da quella OLTP legata alle transazioni
- Separazione tra DataBase Operazionale e Data Warehouse

15

Architetture di un DW : tipologie

- ✓ Le principali tipologie di architetture per un DW
 - **Architettura ad un livello**
c'è solo il DB operazionale, il DW è *virtuale*
 - **Architettura a due livelli**
DB operazionale
DW alimentato dal DB operazionale
 - **Architettura a tre livelli**
DB operazionale
DB *riconciliato* alimentato dal DB operazionale DW
alimentato dal DB riconciliato

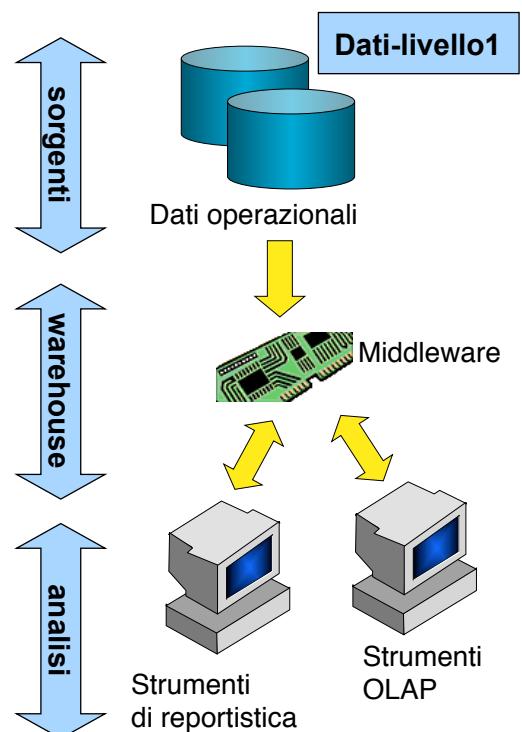
16

Data Mart (DM)

- Collezione di dati rilevante per una particolare area del business, una particolare divisione dell' azienda, una particolare categoria di soggetti.
- Possibili architetture
 - 1) Il DM è un sottoinsieme e/o aggregazione di dati presenti nel DW
⇒ **DM dipendente**, estratto dal DW
 - 2) Il DM è un sottoinsieme e/o aggregazione di dati presenti nel **DB operazionale**
⇒ **DM indipendente**, parte del DW,
ovvero il DW è l' unione dei DM
 - 3) Soluzione ibrida, combinando le precedenti

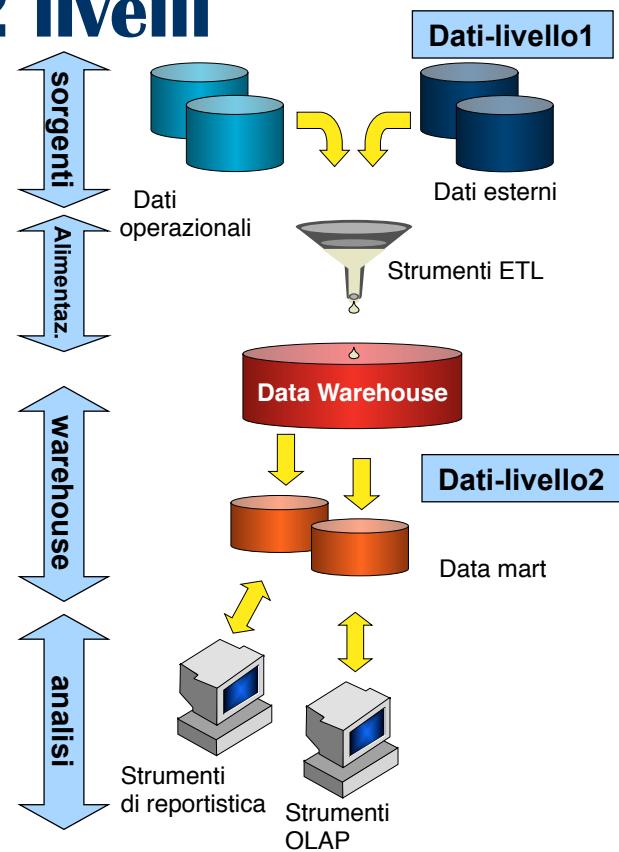
Architetture a 1 livello

- Prevede solo i DB operazionali, quindi i dati coincidono con quelli del DB operazionale
- *DW virtuale*
(non rispetta la separazione OLTP-OLAP)



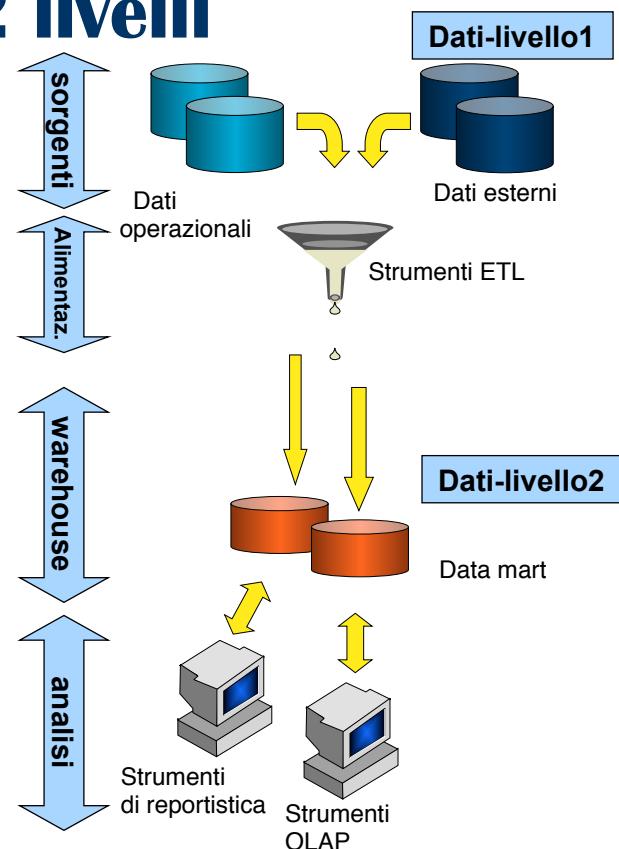
Architetture a 2 livelli

- **DM dipendenti**
alimentati dal DW
- I DM dipendenti 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



Architetture a 2 livelli

- **DM indipendenti**
alimentati dalle sorgenti
- ✓ 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



Architetture a 2 livelli

■ Vantaggi:

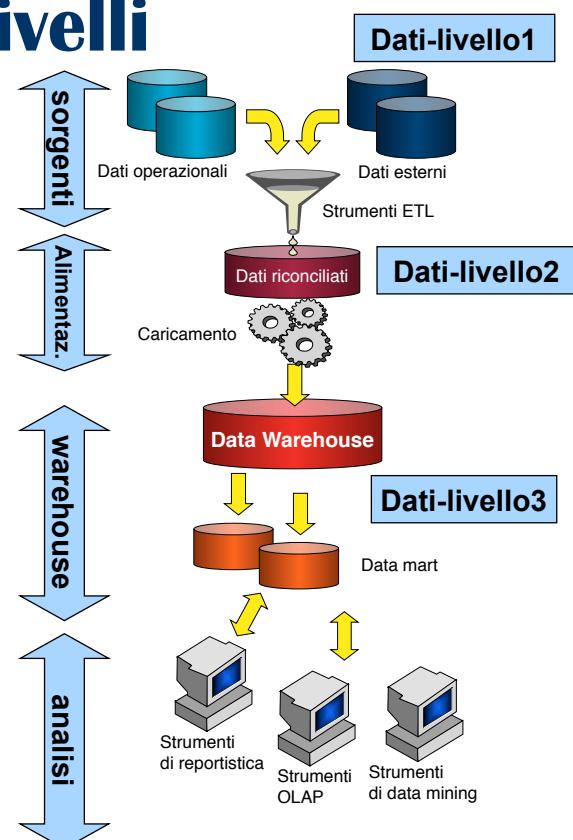
- ✓ A livello del warehouse è continuamente disponibile informazione di buona qualità anche quando, per motivi tecnici oppure organizzativi, è temporaneamente precluso l' accesso alle sorgenti
- ✓ L' interrogazione analitica effettuata sul DW non interferisce con la gestione delle transazioni a livello 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

Architetture a 3 livelli

■ DATI RICONCILIATI

dati operazionali ottenuti a valle del processo di integrazione e ripulitura dei dati sorgente: quindi dati integrati, consistenti, corretti, volatili, correnti

- ✓ Il vantaggio 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



ETL

- Il ruolo degli strumenti di *Extraction, Transformation and Loading* è quello di alimentare una sorgente dati singola, dettagliata, esauriente 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.

Data Warehousing : argomenti

- Progettazione concettuale
 - Modello multidimensionale concettuale DFM (Dimensional Fact Model)
 - Progettazione concettuale da schemi relazionali e schemi E/R
- Progettazione logica
 - Star schema e snowflake schema
- Progettazione dell'alimentazione
 - Metodi ed interrogazioni (SQL) per alimentare lo schema logico
- SQL per l'aggregazione e l'analisi dei dati
 - Estensioni del linguaggio SQL per interrogazioni OLAP
- Introduzione ai sistemi OLAP
 - Panoramica sulle tecniche e sui sistemi OLAP

Tesina

- In gruppi di 2/3 persone che avranno il compito di presentare i seguenti deliverable facenti parte della valutazione:
 - **Schema Concettuale e Schema Logico:**
rappresentazione concettuale e logica del dominio applicativo
 - **ETL e DM:**
script SQL di alimentazione del DM e DM implementato.
 - **Reporting:** schema multidimensionale implementato in **Zucchetti InfoBusiness** e **Report significativi**.
 - **Presentazione del progetto (PPT)**
- L'orale consiste in due parti, entrambe obbligatorie:
 - **Discussione comune sul progetto realizzato**
 - **Discussione ed orale singolo (di ciascun componente)**

Verso il modello multidimensionale

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

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

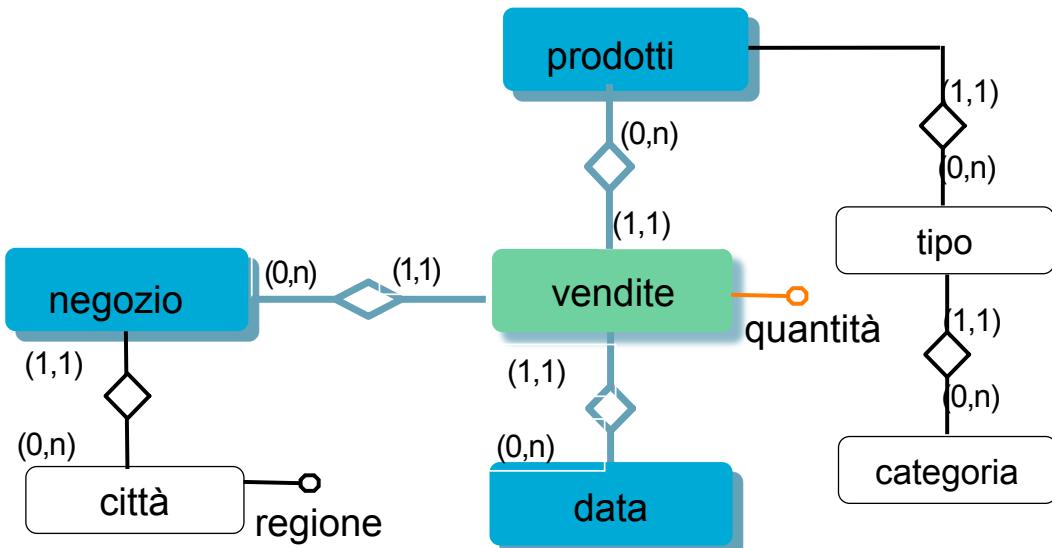
“Quali sono gli ordini che massimizzano gli incassi?”

“Quale di due nuove terapie risulterà in una 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?”

Perché un nuovo modello ?

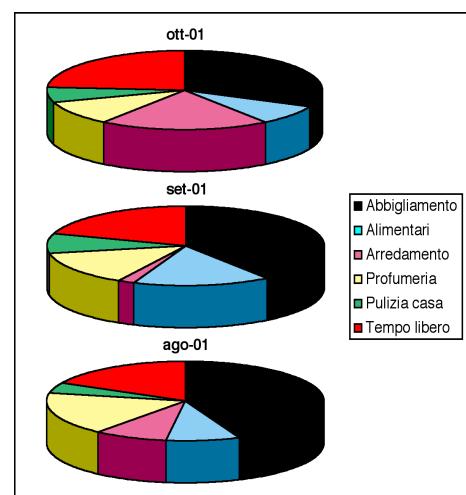
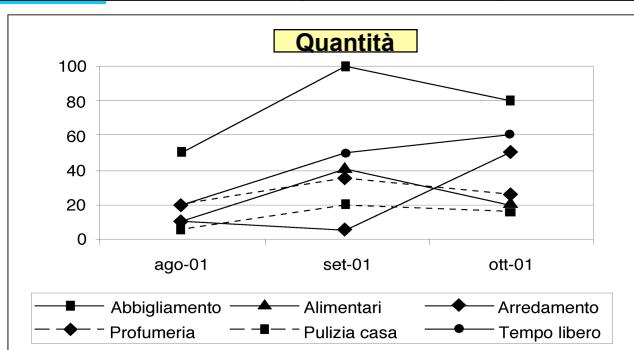
- Il classico modello relazionale e i classici strumenti di analisi (quali i report) spesso non sono sufficienti per un' analisi complessa e completa dei dati.
- Esempio - Schema E/R del DB operazionale delle *vendite*



29

Analisi delle vendite tramite reportistica

Quantità	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



- Un report consente di accedere a informazioni strutturate in modo pressoché invariabile
- Un' analisi differente (ad esempio rispetto alla settimana, ai negozi, ...) è possibile solo se il report lo prevede, ovvero la relativa interrogazione è stata predefinita

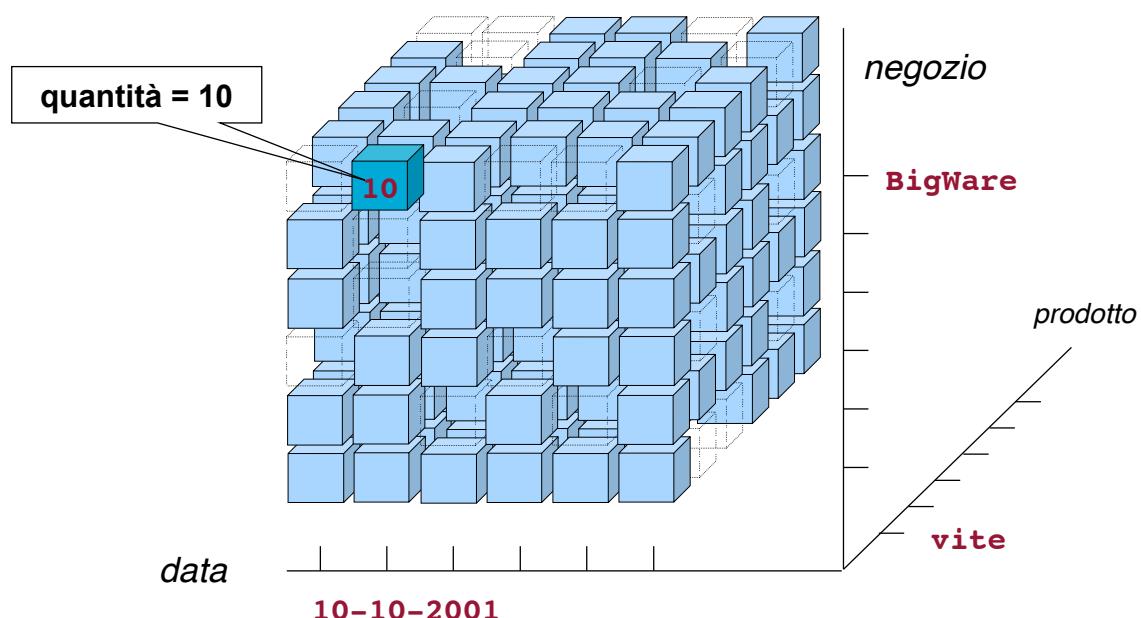
30

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 nel cubo.

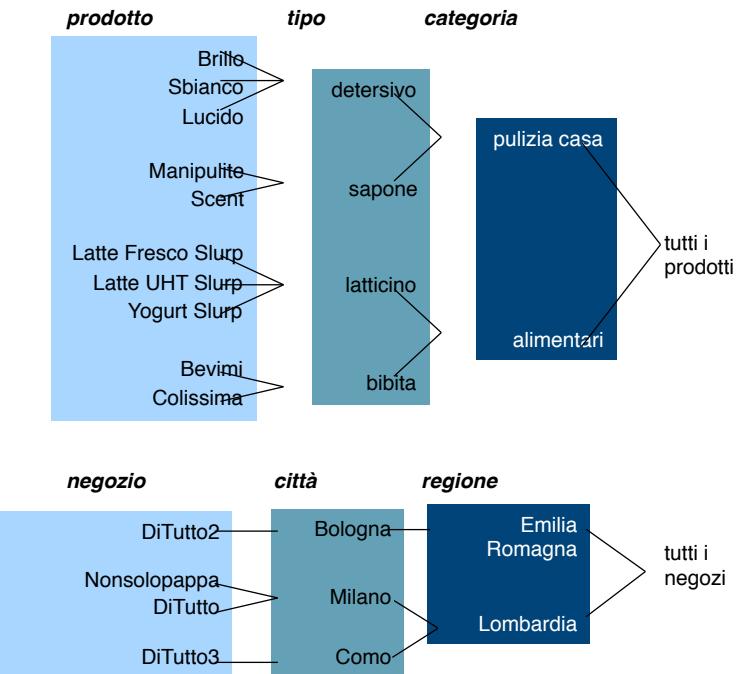
31

Il cubo delle vendite



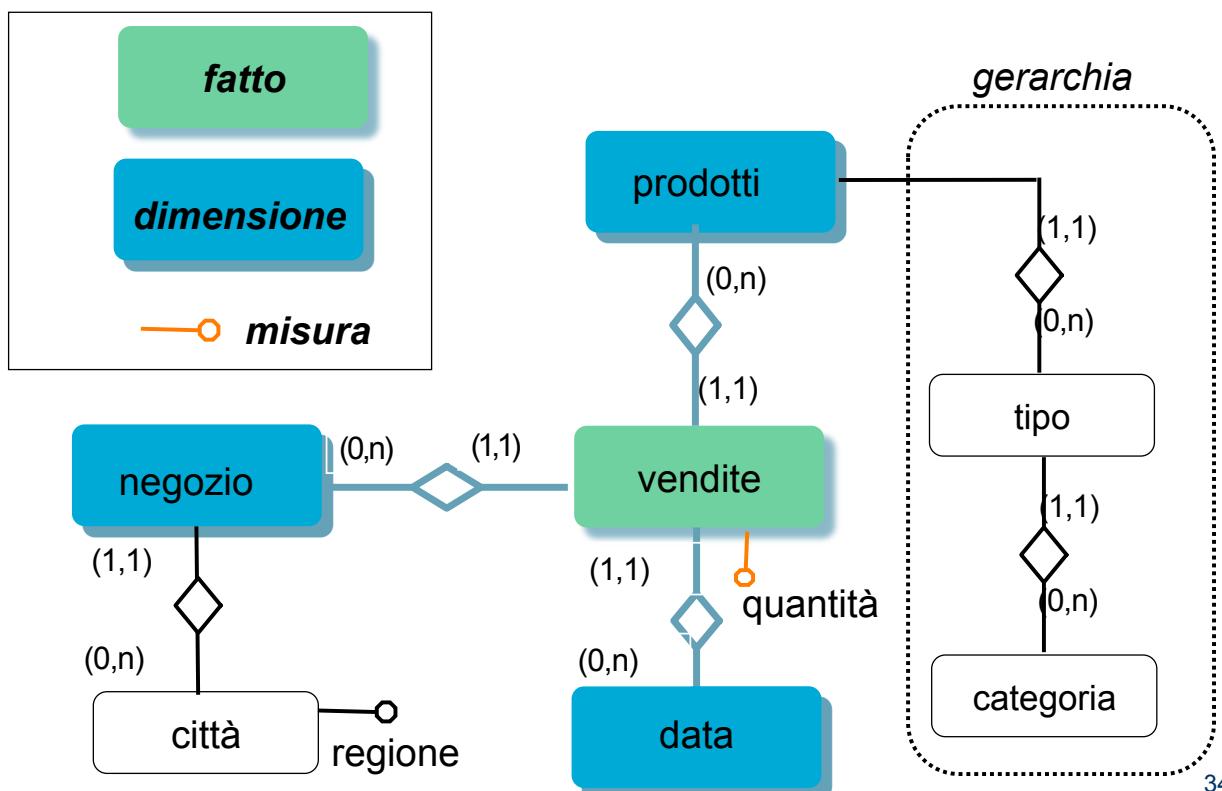
32

Le gerarchie



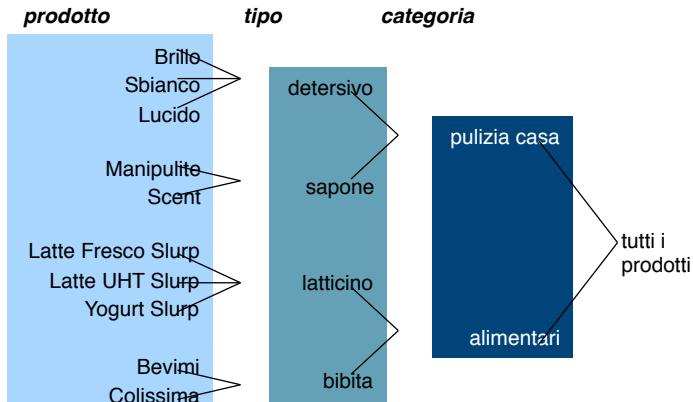
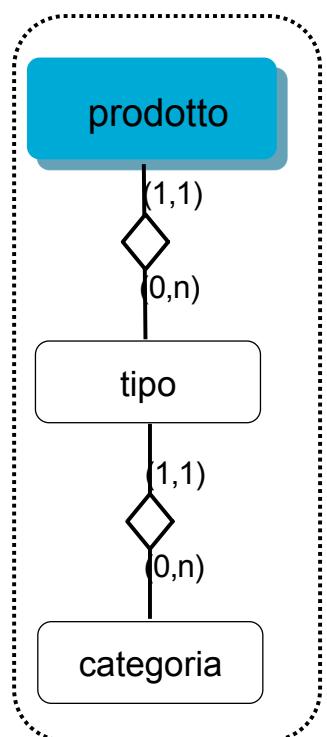
33

Fatti, dimensioni e gerarchie in E/R



34

Gerarchie: associazioni uno-a-molti e dipendenze funzionali

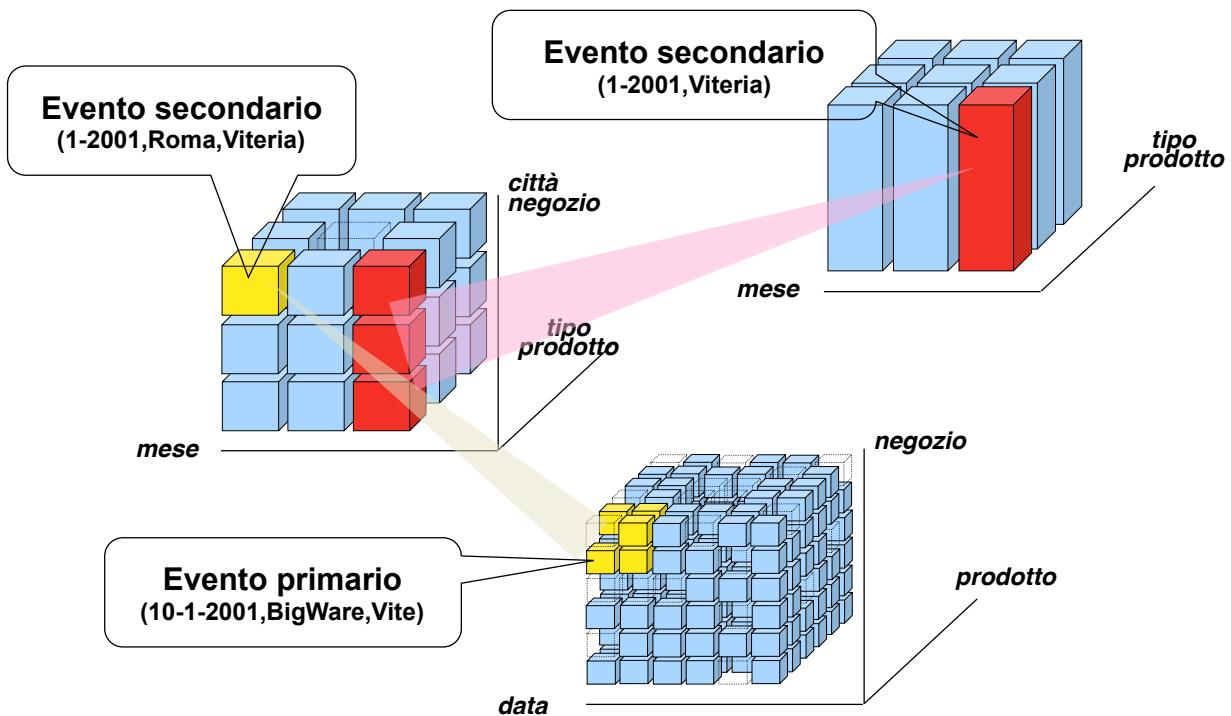


- Una gerarchia corrisponde ad **associazioni uno-a-molti**
- Un' associazione uno-a-molti corrisponde ad una **Dipendenza Funzionale**
 - prodotto → tipo
 - tipo → categoria

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
 - ✓ Nelle vendite, un evento primario registra che, il 10/10/2001, nel negozio BigWare è stata venduta una quantità 10 del prodotto Vite
- Dato un insieme di attributi dimensionali, 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.

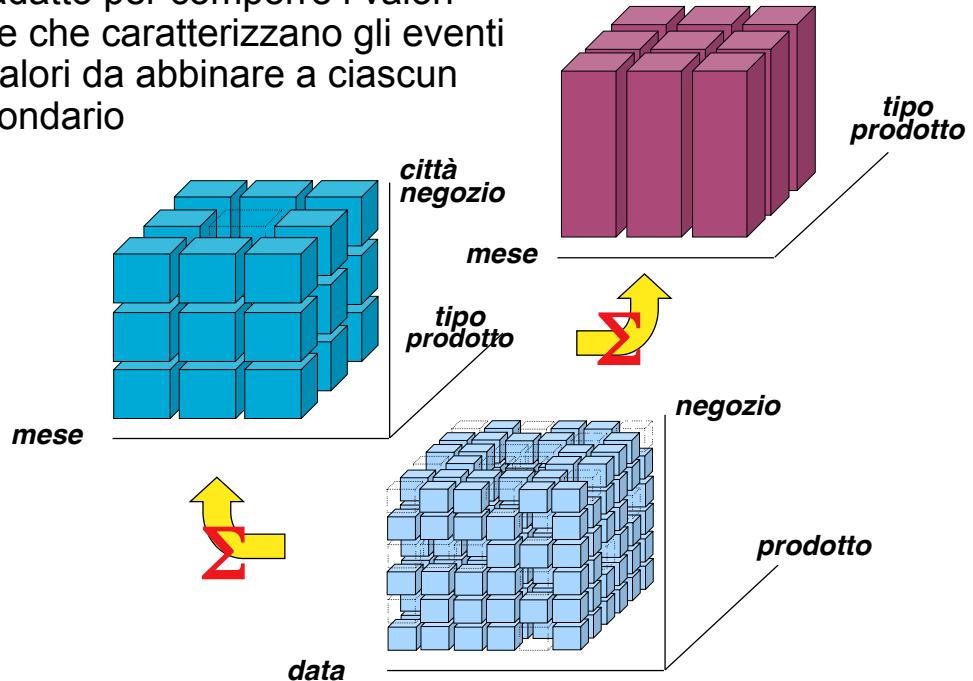
Eventi e aggregazione



37

Operatori di Aggregazione

- 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



38

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

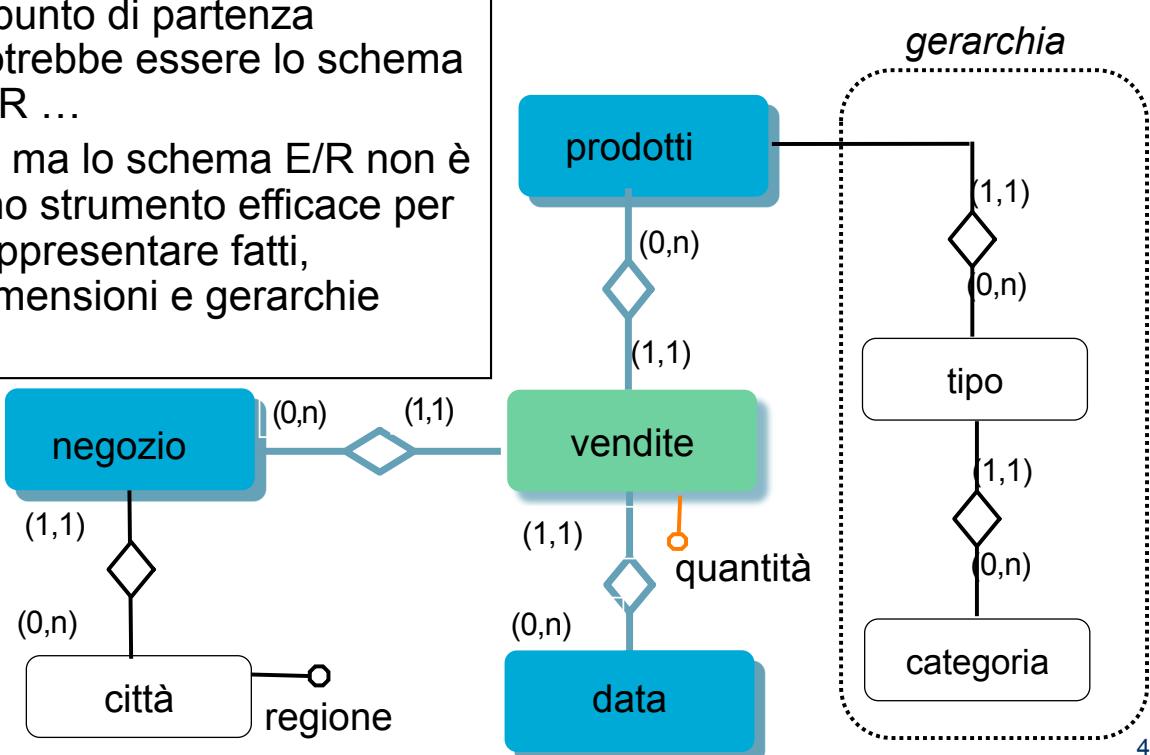


	DiTutto	DiTutto2	Nonsolopappa
<i>Totale:</i>	9000	6500	7800

39

Progettazione di un DW

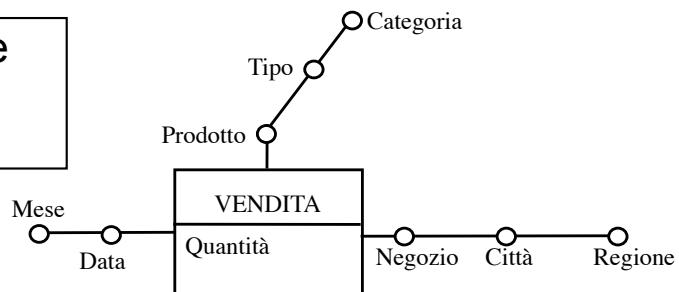
- Il punto di partenza potrebbe essere lo schema E/R ...
- ... ma lo schema E/R non è uno strumento efficace per rappresentare fatti, dimensioni e gerarchie



40

II Dimensional Fact Model

- E' un modello concettuale grafico per DW



- la sua efficacia si apprezza in situazioni complesse:



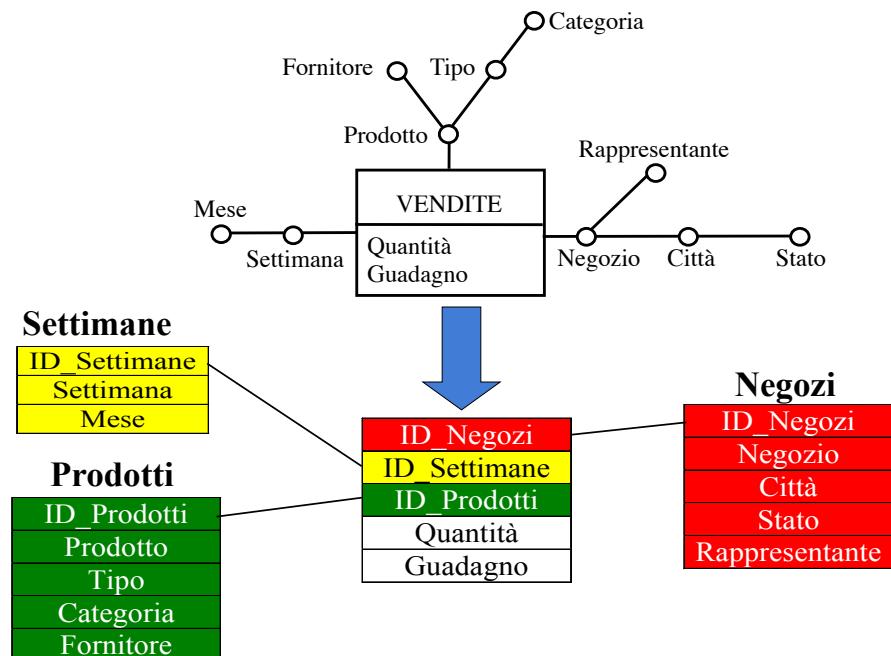
41

Progettazione logica: schema a stella

- Come riportare un cubo e le gerarchie in un DBMS relazionale?
- 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:
 - ✓ Una relazione *DMi* detta *dimension table* per ogni una dimensione.
La chiave della relazione è la dimensione
Gli altri attributi della relazione sono gli altri attributi dimensionali
 - ✓ Una relazione *FT*, chiamata *fact table*, che importa le chiavi di tutte le dimension table.
La chiave 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.

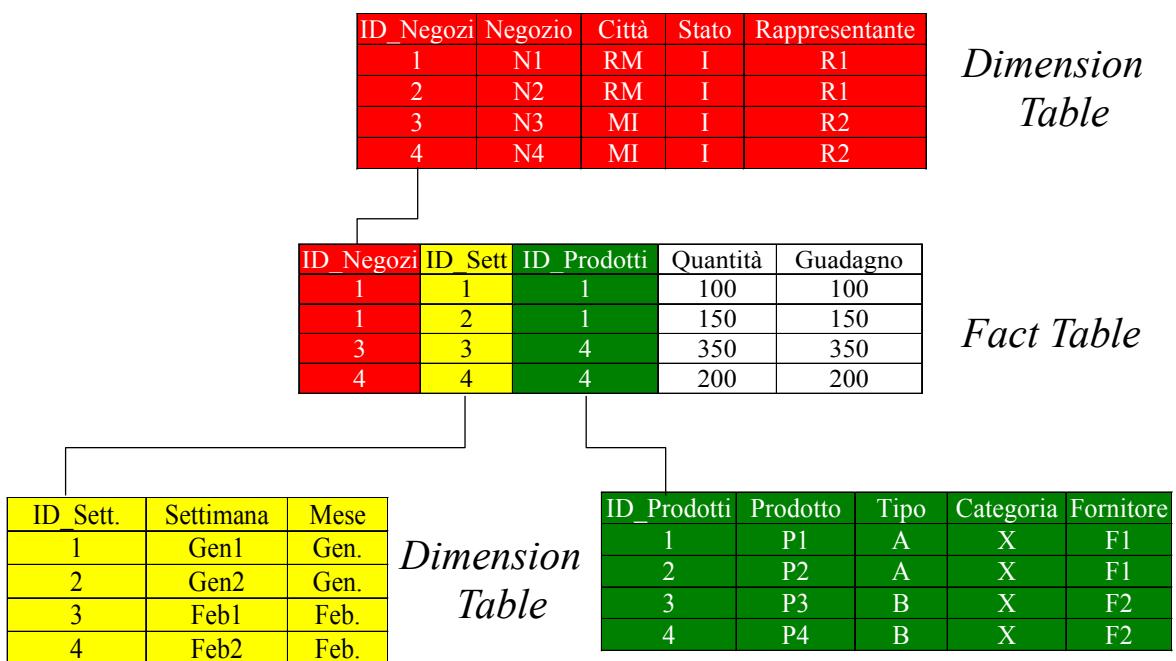
42

Lo schema a stella



43

Lo schema a stella



44

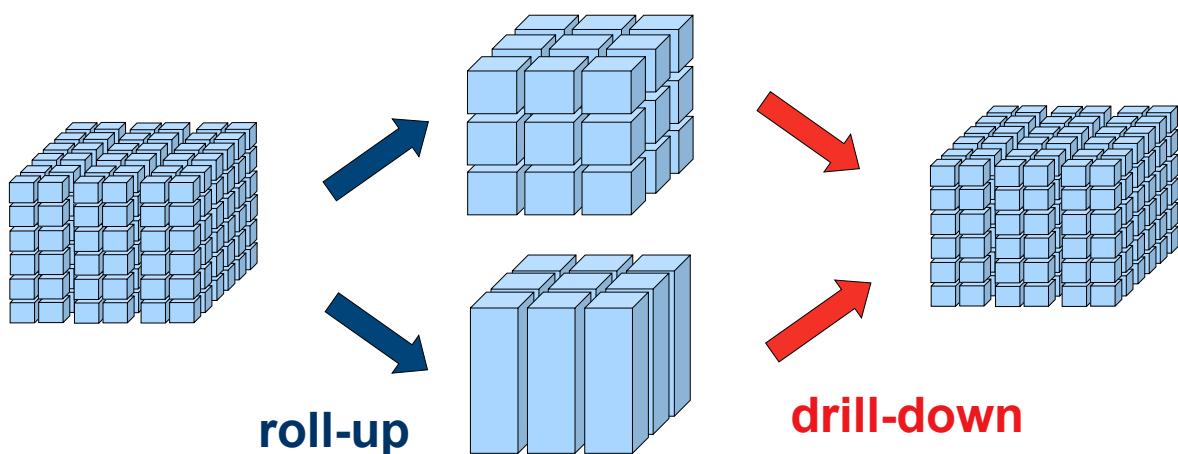
Interrogazione di un cubo : OLAP

- È la principale modalità di fruizione delle informazioni di 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 chi usa strumenti di reportistica svolge un *ruolo 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*
- 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

45

Principali operatori OLAP

- Ogni passo della sessione di analisi è scandito dall'applicazione di un **operatore OLAP** che trasforma l'ultima interrogazione formulata in una nuova interrogazione



46

Principali operatori OLAP

	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
.....

roll-up   drill-down

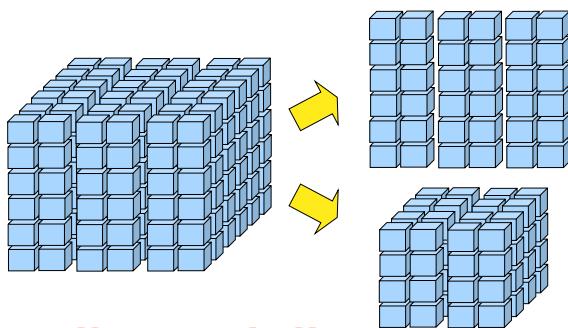
	DiTutto	DiTutto2	Nonsolopappa
2000	2400	2000	1600
2001	3200	2300	3000
2002	3400	2200	3200

roll-up   drill-down

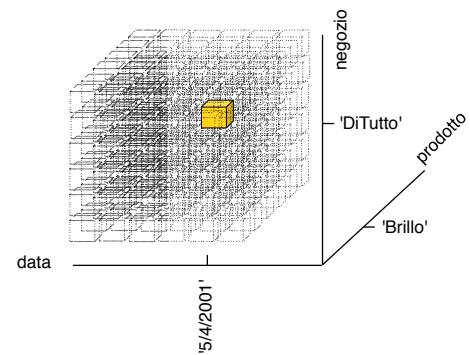
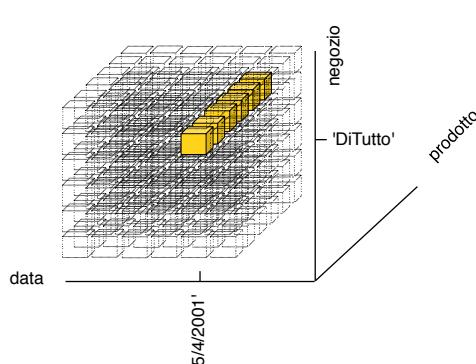
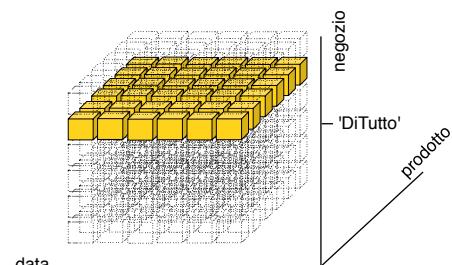
	DiTutto	DiTutto2	Nonsolopappa
<i>Total:</i>	9000	6500	7800

47

Altri operatori OLAP



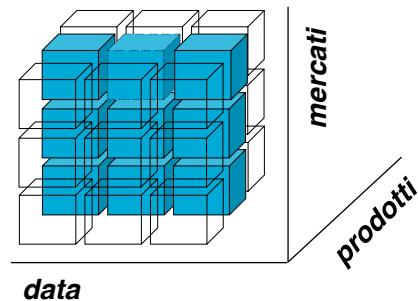
slice-and-dice



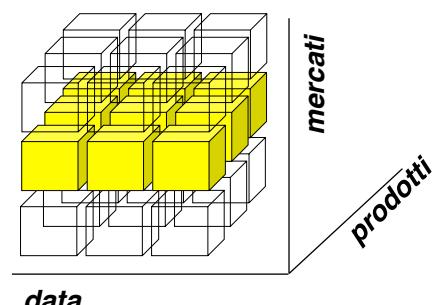
48

Esempi di Slicing

Al manager di prodotto interessa la vendita **di un prodotto** in tutti i periodi e in tutti i mercati



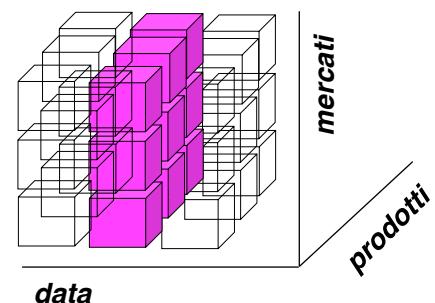
Al manager regionale interessa la vendita dei prodotti in tutti i periodi **nei propri mercati**



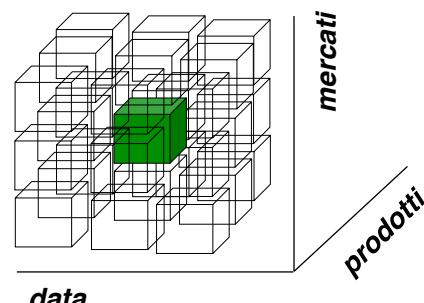
49

Esempi di Slicing

Al manager finanziario interessa la vendita dei prodotti in tutti i mercati relativamente al periodo corrente e quello precedente



Il manager strategico si concentra su una categoria di prodotti, una area regionale e un orizzonte temporale medio



50

Rappresentazioni di interrogazioni OLAP

- Il risultato delle interrogazioni è di tipo multidimensionale; gli strumenti OLAP rappresentano tipicamente i dati in modo tabellare (bidimensionale) evidenziando le diverse dimensioni mediante intestazioni multiple, colori ecc.

Category	Metrics	Dollar Sales		
		Year	1997	1998
Electronics		\$ 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	

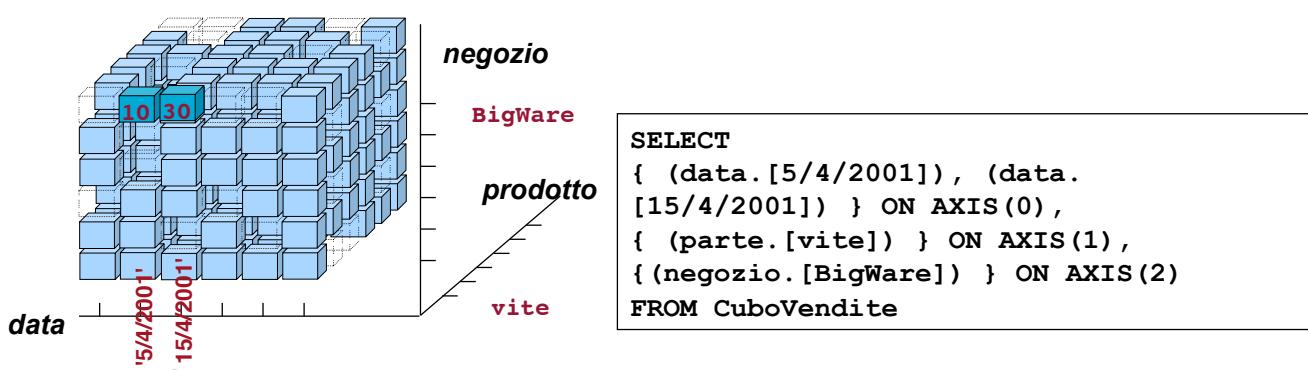
drill-down

Category	Metrics	Dollar Sales											
		Customer Region		North-East		Mid-Atlantic		South-East		Central		South	
		Year	1997	1998	1997	1998	1997	1998	1997	1998	1997	1998	1997
Electronics		\$ 138	\$ 1.184	\$ 1.774	\$ 4.529	\$ 384	\$ 1.892	\$ 138	\$ 7.232	\$ 2.346	\$ 651	\$ 2.554	\$ 9.488
Food		\$ 759	\$ 538	\$ 682	\$ 925	\$ 729	\$ 959	\$ 262	\$ 677	\$ 588	\$ 213	\$ 469	\$ 1.503
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
Health & Beauty		\$ 624	\$ 611	\$ 640	\$ 887	\$ 1.317	\$ 566	\$ 647	\$ 382	\$ 588	\$ 499	\$ 754	\$ 1.162
Household		\$ 5.354	\$ 5.787	\$ 4.112	\$ 5.320	\$ 5.410	\$ 5.416	\$ 4.446	\$ 6.812	\$ 3.058	\$ 4.334	\$ 3.974	\$ 5.008
Kid's Korner		\$ 201	\$ 247	\$ 398	\$ 422	\$ 485	\$ 441	\$ 186	\$ 380	\$ 409	\$ 221	\$ 323	\$ 592
Travel		\$ 624	\$ 608	\$ 505	\$ 559	\$ 564	\$ 1.096	\$ 386	\$ 611	\$ 300	\$ 464	\$ 978	\$ 316

51

Linguaggi di interrogazione OLAP : MDX

- MDX (MultiDimensional eXpression): per recuperare e manipolare dati multidimensionali nel sistema OLAP di SQL Server



- Per visualizzarlo in modo bi-dimensionale

```

SELECT
{ (negozio.[BigWare]) } ON COLUMNS,
{ (parte.[vite]),data.[5/4/2001]}, { (parte.[vite]),data.[15/4/2001]) } ON ROWS
FROM CuboVendite
  
```

VITE	BIGWARE	
	5/4/2001	10
	15/4/2001	30

52

Altre tecniche di analisi : Data Mining

- Attività orientata a scoprire informazioni nascoste nei dati
 - ✓ In presenza di moli di dati molto elevate, l'utente non è sempre in grado di individuare tutti i pattern (modelli) significativi presenti
 - ✓ Il *data mining* raccoglie tecniche di *intelligenza artificiale* e *pattern recognition* per aiutare l'utente nella ricerca di pattern: è sufficiente indicare cosa e dove si vuole ricercare
- Esistono in sostanza tre 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
 - ✓ *data mining*: richiede all'utente la conoscenza dei principi che stanno alla base degli strumenti utilizzati

53

Implementazione dei DW: ROLAP

- Relational OLAP: implementazione del DW su **RDBMS**
- Gartner Magic Quadrant for Data Warehouse DBMS:



54

Business Intelligence Platforms

■ Gartner Magic Quadrant for Business Intelligence Platforms

