Data Warehousing e Business Intelligence

Urbino – 15 maggio 2008

Prof. Matteo Golfarelli
Alma Mater Studiorum - Università di Bologna

Introduzione al Data Warehousing

L'evoluzione dei sistemi informativi

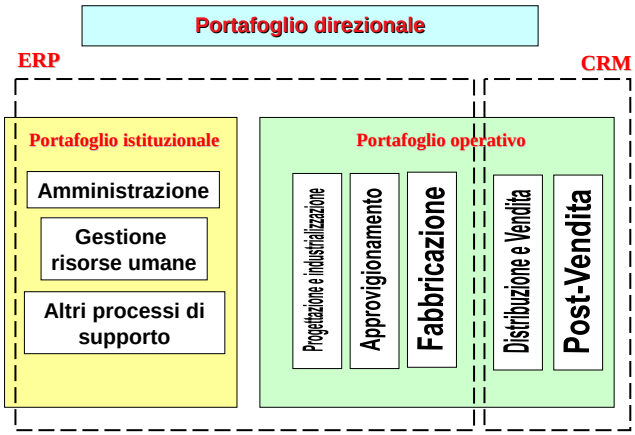
- Il ruolo dei Sistemi Informatici è radicalmente cambiato dai primi anni '70 a oggi. I sistemi informatici si sono trasformati da semplici strumenti per migliorare l'efficienza dei processi a elementi centrali dell'organizzazione aziendale in grado di rivoluzionare la struttura dei processi aziendali

Il duplice ruolo dell'informatica

Tecnologia di supporto alla gestione del Sistema Informativo

Disciplina organizzativa che influenza i processi, servizi e struttura aziendale

Il portafoglio applicativo



Il portafoglio direzionale

- È 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…

...disciplina che consente a chi deve decidere in azienda di capire, attraverso soluzioni software, i fattori chiave del business e conseguentemente di prendere le migliori decisioni in quel momento

5

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

6

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

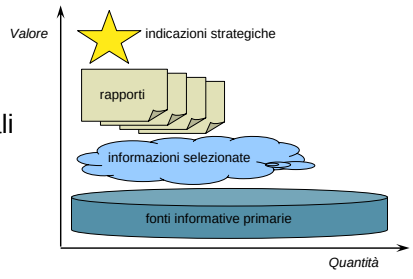
dati = ~~informazione~~

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

7

Dai dati alle informazioni

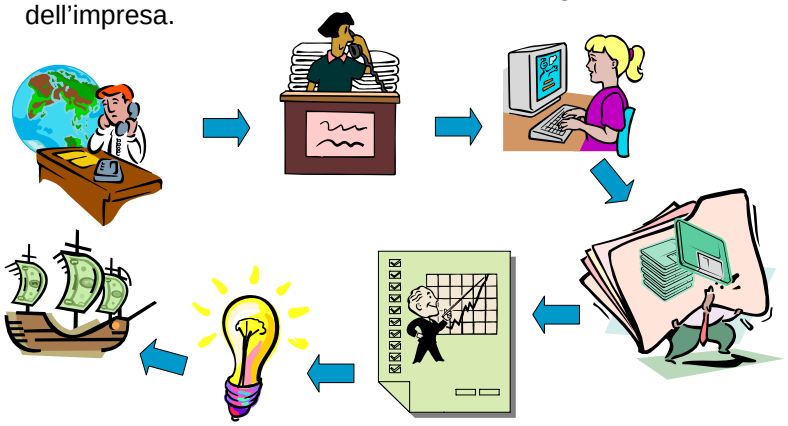
- Per ogni azienda è fondamentale poter disporre in maniera rapida e completa delle informazioni necessarie al processo decisionale: le indicazioni strategiche sono estrapolate principalmente dalla mole dei dati operazionali contenuti nei database aziendali, attraverso un procedimento di selezione e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



8

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.



9

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.

10

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, On-Line Analytical Processing) da quella legata alle transazioni (OLTP, On-Line Transactional Processing), 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

11

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

12

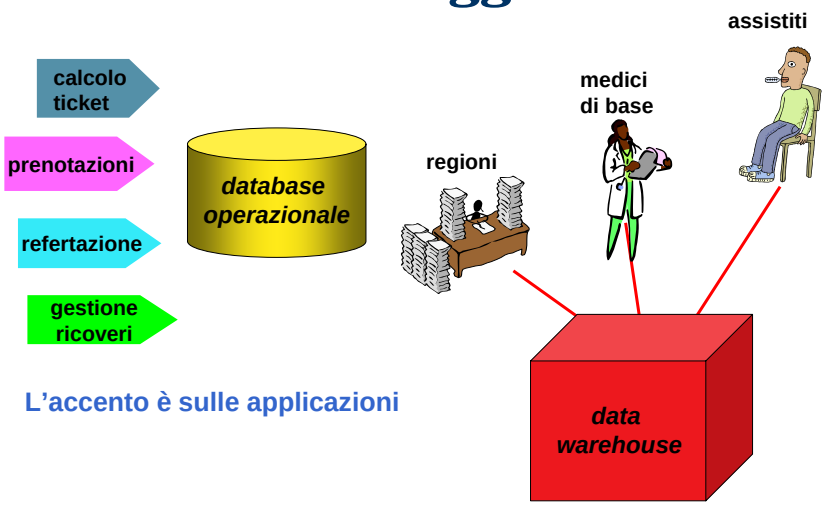
Il Data Warehouse

- 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.*

...orientato ai soggetti

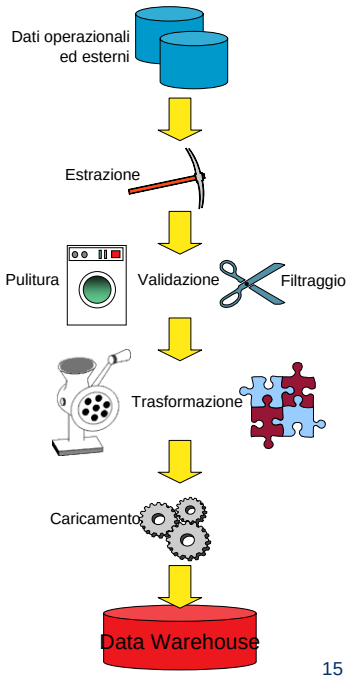


L'accento è sulle applicazioni

L'accento è sui soggetti

...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.



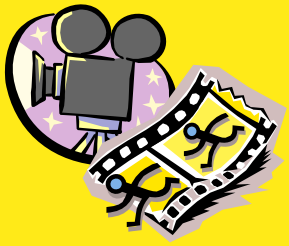
15

...rappresentativo dell'evoluzione temporale

DB operazionali

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

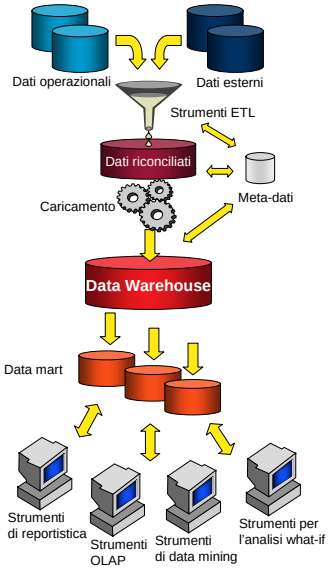
DW



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

16

Architetture a 3 livelli



Livello delle sorgenti

Livello di alimentazione

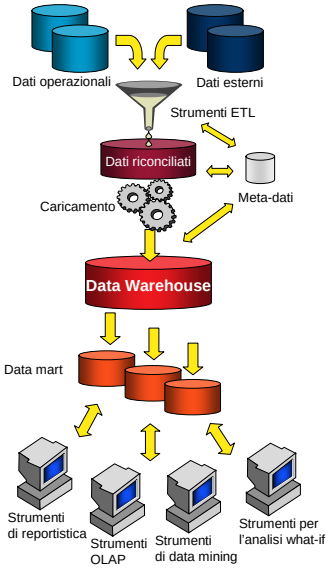
Livello del warehouse

Livello di analisi

DATI RICONCILIATI:

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

Architetture a 3 livelli



Livello delle sorgenti

Livello di alimentazione

Livello del warehouse

Livello di analisi

DATA MART:

un sottoinsieme o un'aggregazione dei dati presenti nel DW primario, contenente l'insieme delle informazioni rilevanti per una particolare area del business, una particolare divisione dell'azienda, una particolare categoria di soggetti.

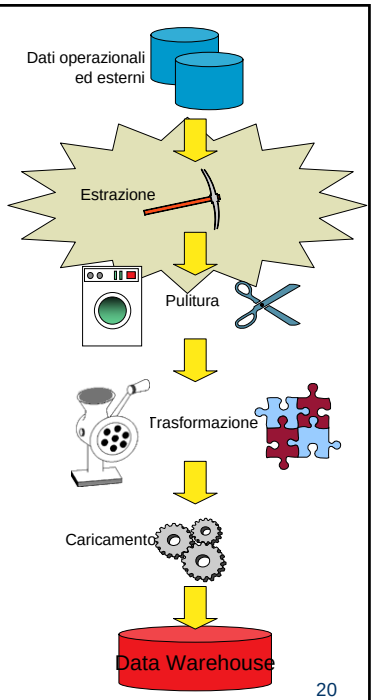
ETL

- 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.
  - ✓ estrazione
  - ✓ pulitura
  - ✓ trasformazione
  - ✓ caricamento

19

Estrazione

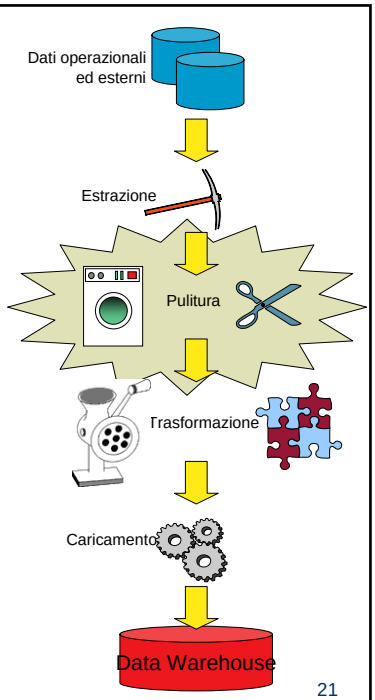
- I dati rilevanti vengono estratti dalle sorgenti tipicamente con modalità incrementale ossia catturando solamente i cambiamenti avvenuti nelle sorgenti dall'ultima estrazione.



20

Pulitura

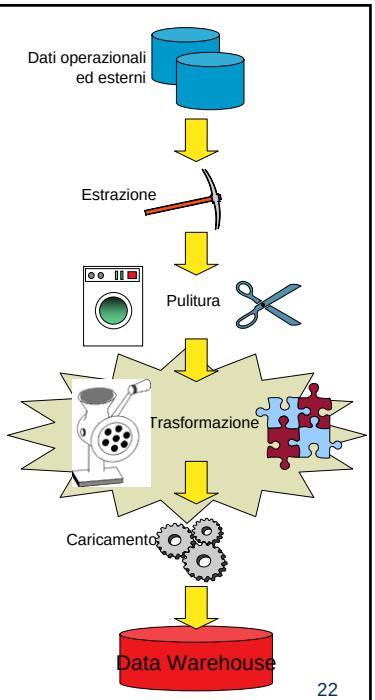
- Si incarica di migliorare la qualità dei dati delle sorgenti
  - ✓ dati duplicati
  - ✓ inconsistenza tra valori logicamente associati
  - ✓ dati mancanti
  - ✓ uso non previsto di un campo
  - ✓ valori impossibili o errati
  - ✓ valori inconsistenti per la stessa entità dovuti a differenti convenzioni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



21

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



22

Pulitura e trasformazione

Carlo Bianchi
P.zza Grande 12
50126 Bologna (I)

Normalizzazione

nome:	Carlo
cognome:	Bianchi
indirizzo:	P.zza Grande 12
CAP:	50126
città:	Bologna
nazione:	I

Standardizzazione

nome:	Carlo
cognome:	Bianchi
indirizzo:	Piazza Grande 12
CAP:	50126
città:	Bologna
nazione:	Italia

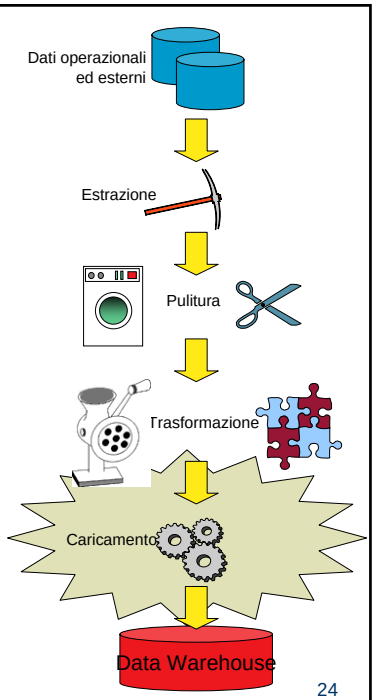
Correzione

nome:	Carlo
cognome:	Bianchi
indirizzo:	Piazza Grande 12
CAP:	40126
città:	Bologna
nazione:	Italia

23

Caricamento

- Vengono aggiunti nel DW i soli cambiamenti occorsi nei dati sorgente



24

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?"

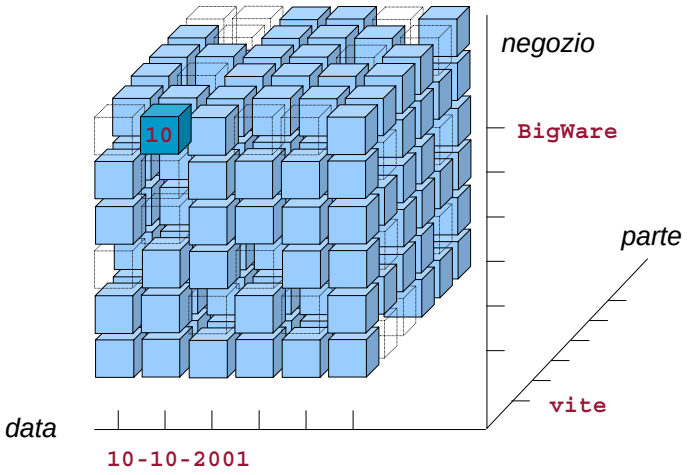
25

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.

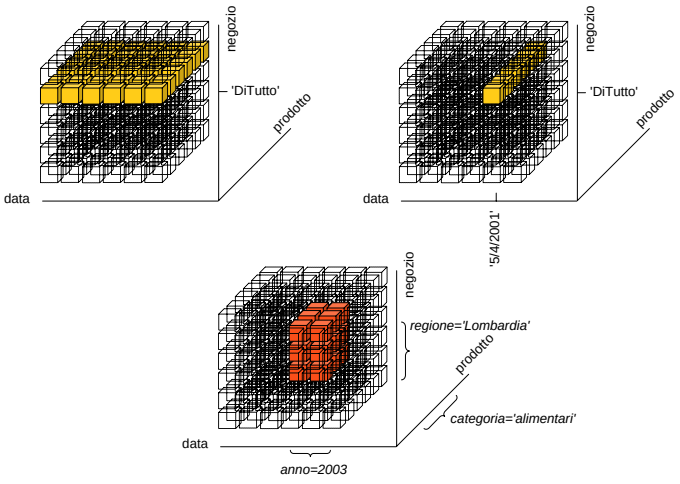
26

Il cubo delle vendite



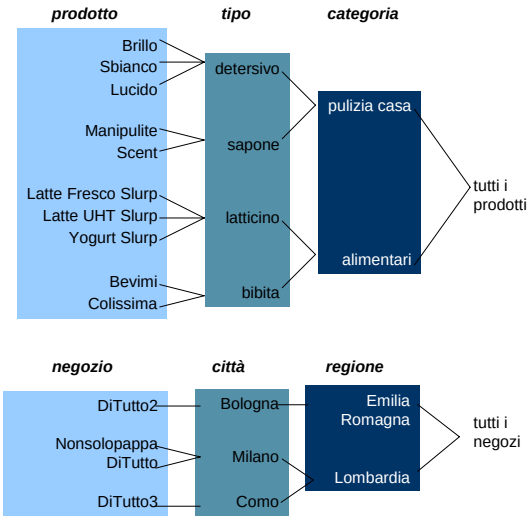
27

Slicing and dicing

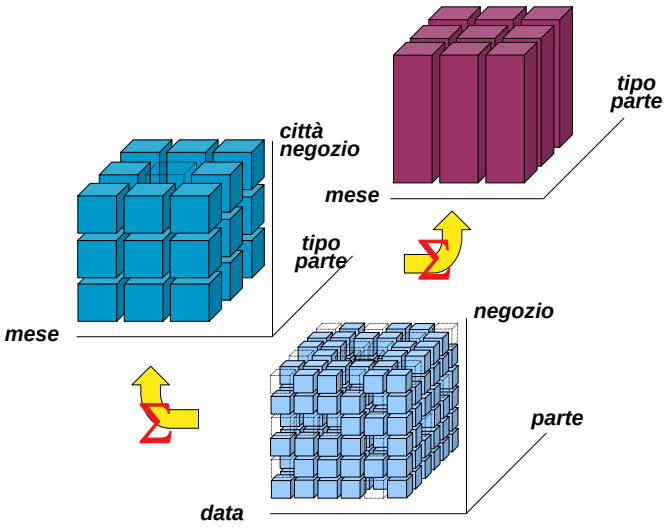


28

Le gerarchie



Aggregazione



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
Totale	9000	6500	7800

31

Tecniche di analisi dei dati

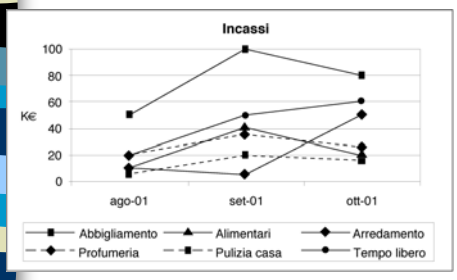
- Una volta che i dati sono stati ripuliti, integrati e trasformati, occorre capire come trarne il massimo vantaggio informativo
- Esistono in sostanza tre approcci differenti, supportati da altrettante categorie di strumenti, all'interrogazione di un DW da parte degli utenti finali:
  - ✓ *reportistica:* 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

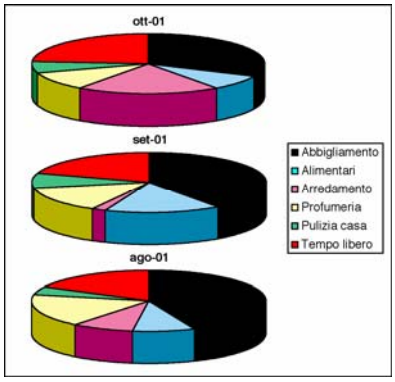
32

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





33

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

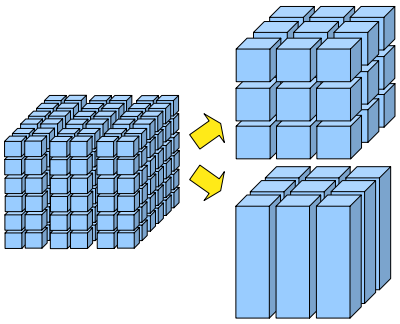
34

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.

35

OLAP: operatori



roll-up

36

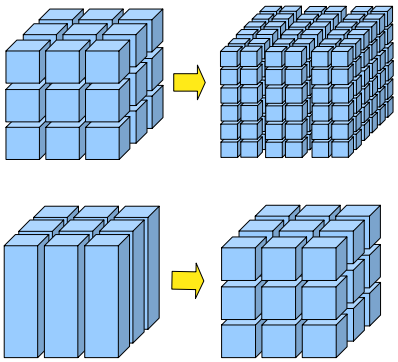
OLAP: operatori

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	$ 598	$ 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	$ 1.588	$ 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	$ 3.514	$ 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	$ 1.127	$ 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	$ 1.885	$ 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	$ 1.204
Q2 1998	$ 1.773	$ 3.658	$ 1.862	$ 1.213	$ 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

OLAP: operatori



drill-down

OLAP: operatori

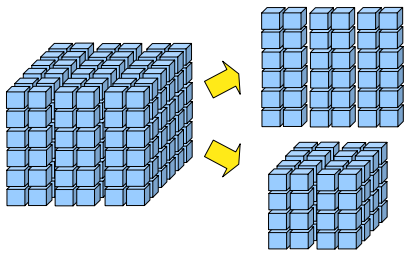
Metrics / Customer Region / Quarter	Dollar Sales: North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germany	Canada
Q1 1997	$ 1.526	$ 1.249	$ 978	$ 1.885	$ 3.650	$ 4.855	$ 1.834	$ 2.552	$ 1.920	$ 643	$ 663
Q2 1997	$ 2.979	$ 1.415	$ 2.664	$ 1.582	$ 1.130	$ 1.906	$ 884	$ 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	$ 1.204
Q2 1998	$ 1.773	$ 3.658	$ 1.862	$ 1.213	$ 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

drill-down

Metrics / Customer City / Quarter	Dollar Sales: Arlin	San Pedro	Springfield	Chappel Hill	Scranburg	Pebble Beach	Martinsville	Maddon	Peoria	Pecos	Lake Barkley	Alcameda	Fingers Lake	S
Q1 1997	$ 675										$ 39			
Q2 1997				$ 203					$ 53				$ 135	
Q3 1997				$ 276								$ 252	$ 63	
Q4 1997	$ 215	$ 124			$ 113	$ 45	$ 192	$ 348				$ 79	$ 98	
Q1 1998			$ 140	$ 174			$ 85				$ 237	$ 30	$ 119	
Q2 1998								$ 12	$ 17					
Q3 1998	$ 734					$ 25	$ 1.535							
Q4 1998						$ 219	$ 119	$ 142		$ 85	$ 1.533			

39

OLAP: operatori



slice-and-dice

40

OLAP: operatori

Category	Year	Metrics / Customer Region	Dollar Sales North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germa
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	$ 464	$ 316	$ 573	$ 257	$ 198	$

slice-and-dice

Filter Details:
Year = 1998

Category	Metrics / Customer Region	Dollar Sales North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germany	Ca
Electronics		$ 1.184	$ 4.529	$ 1.892	$ 7.232	$ 651	$ 9.488	$ 476	$ 2.683	$ 462	$ 702	
Food		$ 538	$ 925	$ 959	$ 677	$ 213	$ 1.503	$ 261	$ 165	$ 175	$ 100	
Gifts		$ 1.955	$ 2.785	$ 2.800	$ 2.695	$ 1.813	$ 2.844	$ 1.778	$ 1.158	$ 717	$ 686	
Health & Beauty		$ 611	$ 887	$ 566	$ 382	$ 499	$ 1.162	$ 1.044	$ 273	$ 72		
Household		$ 5.787	$ 5.320	$ 5.416	$ 6.812	$ 4.334	$ 5.008	$ 7.588	$ 2.139	$ 3.649	$ 2.791	$
Kid's Korner		$ 247	$ 422	$ 441	$ 380	$ 221	$ 592	$ 290	$ 198	$ 19	$ 69	
Travel		$ 608	$ 559	$ 1.096	$ 611	$ 464	$ 316	$ 573	$ 257	$ 198	$ 55	

41

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
    - Regole associative
    - Clustering
    - Alberi di decisione
    - Serie temporali

42

Data mining: regole associative

- Consentono di determinare le regole di implicazione logica presenti nella base di dati, quindi di individuare i gruppi di affinità tra oggetti
- Applicazioni:
  - ✓ studio delle abitudini di acquisto per la pubblicità mirata e l'organizzazione della merce sugli scaffali (*market-basket analysis*)
  - ✓ studio della variabilità delle vendite in assenza di un certo prodotto

Quali sono i prodotti che vengono acquistati assieme?

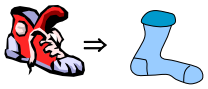
{Acquisto X} ⇒ {Acquisto Y}

Data mining: regole associative

- Consentono di determinare le regole di implicazione logica presenti nella base di dati, quindi di individuare i gruppi di affinità tra oggetti
- Applicazioni:
  - ✓ studio delle abitudini di acquisto per la pubblicità mirata e l'organizzazione della merce sugli scaffali (*market-basket analysis*)
  - ✓ studio della variabilità delle vendite in assenza di un certo prodotto

{scarpe} ⇒ {calze}

supporto=70%
confidenza=85%



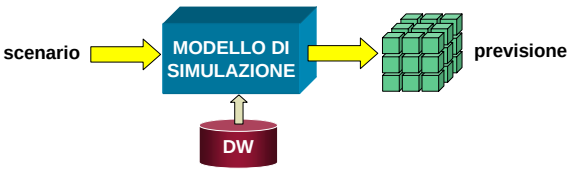
Analisi what-if

- Per valutare in anticipo le conseguenze di una mossa strategica o tattica, le aziende hanno bisogno di sistemi previsionali affidabili
- I DW supportano l'analisi dettagliata dei dati passati, ma non sono in grado di dare anticipazioni sui trend futuri

45

Analisi what-if

- L'analisi what-if è una simulazione data-intensive il cui obiettivo è studiare il comportamento di un sistema complesso (il sistema-azienda o una sua parte) alla luce di una data ipotesi (*scenario*)
- Più pragmaticamente, l'analisi what-if misura come le variazioni in un insieme di variabili indipendenti impattano sui valori di un insieme di variabili dipendenti con riferimento a un dato modello di simulazione
  - ✓ Esempio di quesito what-if nel dominio del PCT: "**Come** varierebbe la durata media di un processo **se** ogni tribunale avesse a disposizione 2 giudici in più? E **se** si riducessero del 20% i tempi burocratici di gestione di una pratica?"



46

Cenni sulla progettazione di 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

48

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

49

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

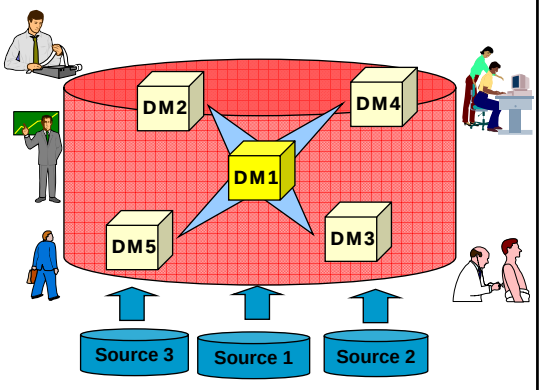
50

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

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



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.

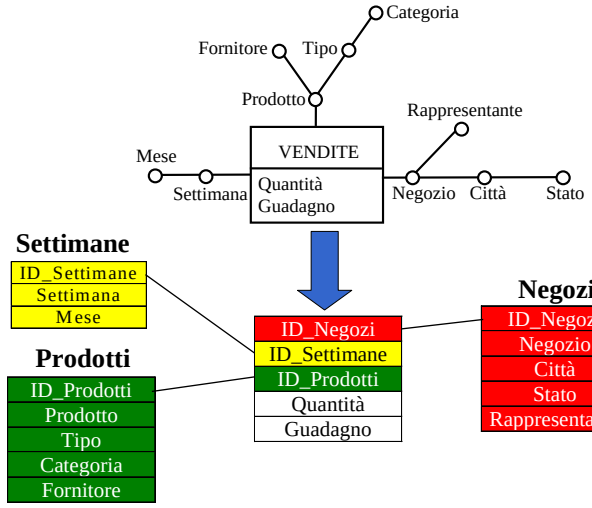
53

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…, 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…, d_n; *FT* contiene inoltre un attributo per ogni misura.

54

Lo schema a stella



Lo schema a stella

ID_Negozi	Negozio	Città	Stato	Rappresentante
1	N1	RM	I	R1
2	N2	RM	I	R1
3	N3	MI	I	R2
4	N4	MI	I	R2

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	Gen1	Gen.
2	Gen2	Gen.
3	Feb1	Feb.
4	Feb2	Feb.

Dimension Table

ID_Prodotti	Prodotto	Tipo	Categoria	Fornitore
1	P1	A	X	F1
2	P2	A	X	F1
3	P3	B	X	F2
4	P4	B	X	F2

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
  - La denormalizzazione introduce una forte ridondanza nei dati
- La Fact Table contiene tuple relative a diversi livelli di aggregazione
  - L'elevata dimensione incide sui tempi di accesso ai 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

57

Ricerca nel settore del Data Warehouse

… alcuni esempi

- Sistemi di BI distribuiti: la risposta a una query è ottenuta consultando DW distinti
  - Quali sono le architetture più idonee (web-service?)
  - Come faccio ad individuare una semantica comune (ontologie?)
- Query OLAP con preferenze
  - "Voglio vedere i dati delle vendite superiori a 1000, ma se non ne esistono mi vanno bene le più alte tra quelle sotto soglia"
- Data mining + OLAP = OLAM
  - Algoritmi di mining eseguibili direttamente su cubi multidimensionali e da utenti non esperti

59