Data Warehousing

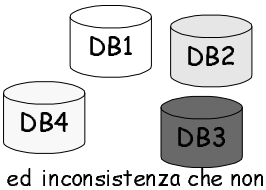
1

Introduzione al data warehousing

2

Il problema

In genere:



↑ abbondanza di dati

ma anche

↓ abbondanza di ridondanza ed inconsistenza che non permette di utilizzare i dati in modo utile a fini decisionali

↓la disponibilità di troppi dati rende difficile estrapolare le informazioni veramente importanti

3

Tipiche richieste

- Qual è il volume delle vendite per regione e categorie di prodotto durante l'ultimo anno?
- Come si correlano i prezzi delle azioni delle società produttrici di hardware con i profitti trimestrali degli ultimi 10 anni?
- Quali sono stati i volumi di vendita dello scorso anno per regione e categoria di prodotto?
- In che modo i dividendi di aziende di hardware sono correlati ai profitti trimestrali negli ultimi 10 anni?

4

Possibili applicazioni

contesti ⇒

- gestione dei rischi
- analisi finanziaria
- programmi di marketing
- analisi statistica
- integrazione DB clienti
- integrazione relazioni clienti
- analisi temporale

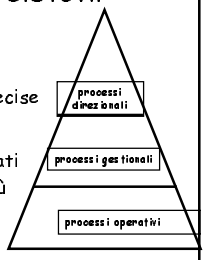
- telecomunicazioni
- banking
- università
- assicurazioni
- beni di consumo
- salute
- produzione

⇐ **problematiche**

5

Processi, dati e decisioni

- processi operativi
  - dati dipartimentali e dettagliati
  - decisioni strutturate, con regole precise
- processi gestionali
  - dati settoriali, parzialmente aggregati
  - decisioni semistrutturate: regole più intervento creativo/responsabile
- processi direzionali
  - dati integrati e fortemente aggregati
  - decisioni non strutturate



6

Sistemi informatici: una classificazione

- Transaction processing systems:
  - per i processi operativi
- Management information systems:
  - settoriali, per i processi gestionali
- Decision support systems:
  - fortemente integrati, di supporto ai processi direzionali

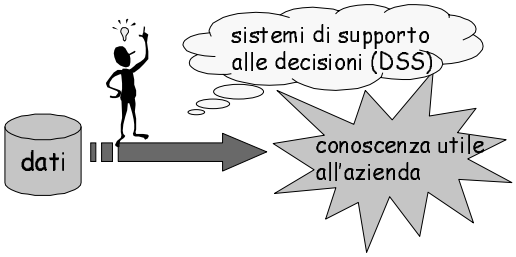
7

Sistemi di supporto alle decisioni

- Richiedono operazioni non previste a priori
- Coinvolgono spesso grandi quantità di dati, anche storici e aggregati
- Coinvolgono dati provenienti da varie fonti operative, anche esterne

8

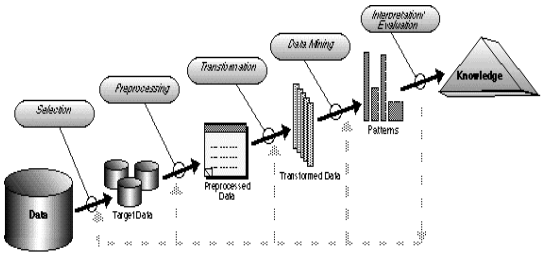
In sintesi …



DSS: Tecnologia che supporta la dirigenza aziendale nel prendere decisioni tattico-strategiche in modo migliore e più veloce

9

Il processo di scoperta: dai dati alla conoscenza



10

Sistemi di supporto alle decisioni

- Ruolo

Nel passato	Nel futuro
descrivere il passato	anticipare il futuro
descrivere i problemi	suggerire i cambiamenti
ridurre i costi	aumentare i profitti

- Problematiche
  - gestire grandi moli di dati
  - accedere a diverse fonti dati su piattaforme eterogenee
  - garantire l'accesso a più utenti per interrogazioni, analisi in tempo reale, simulazioni
  - gestire versioni storiche dei dati

11

Perché i sistemi tradizionali non sono sufficienti?

- no dati storici
- sistemi eterogenei
- basse prestazioni
- DBMS non adeguati al supporto decisionale
- problemi di sicurezza

12

Più formalmente ...

- Sistemi tradizionali
  - On-Line Transaction Processing (OLTP)
- Sistemi di data warehousing
  - On-Line Analytical Processing (OLAP)

  ⇒ Profondamente diversi

13

OLAP

- Elaborazione di operazioni per il supporto alle decisioni
  - Operazioni complesse e casuali
  - Ogni operazione può coinvolgere molti dati
  - Dati aggregati, storici, anche non attualissimi
  - Le proprietà "acide" non sono rilevanti, perché le operazioni sono di sola lettura
- OLAP e OLTP
  - I requisiti sono contrastanti
  - Le applicazioni dei due tipi possono danneggiarsi a vicenda

14

In dettaglio ...

	OLTP	OLAP
funzione	gestione giornaliera	supporto alle decisioni
progettazione	orientata alle applicazioni	orientata al soggetto
frequenza	gironaliera	sporadica
dati	recenti, dettagliati	storici, riassuntivi, multidimensionali
sorgente	singola DB	DB multiple
uso	ripetitivo	ad hoc
accesso	read/write	read
flessibilità accesso	uso di programmi precompilati	generatori di query
# record acceduti	decine	migliaia
tipo utenti	operatori	manager
# utenti	migliaia	centinaia
tipo DB	singola	multiple, eterogenee
performance	alta	bassa
dimensione DB	100 MB - GB	100 GB - TB

15

Evoluzione dei DSS

- Anni '60: rapporti batch
  - difficile trovare ed analizzare i dati
  - costo, ogni richiesta richiede un nuovo programma
- Anni '70: DSS basato su terminale
  - non integrato con strumenti di automazione d'ufficio
- Anni '80: strumento d'automazione d'ufficio
  - strumenti di interrogazione, fogli elettronici, interfacce grafiche
  - accesso ai dati operazionali
- Anni '90: data warehousing, con strumenti integrati OLAP

16

I sistemi di data warehousing

- Costruzione di un un nuovo raccoglitore di informazioni che integri i dati elementari 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

17

I sistemi di data warehousing

- Strumenti di archiviazione e interrogazione per ottenere facilmente e in tempi ridotti, dall'enorme quantità di dati disponibili nei database, informazioni di sintesi che permettano la valutazione di un fenomeno, la scoperta di correlazioni significative, e l'acquisizione di conoscenza utile

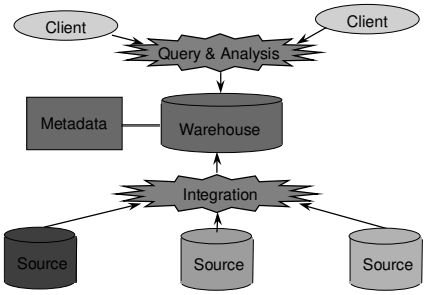
18

I sistemi di data warehousing

Il Data Warehousing si può definire come il processo di integrazione di basi di dati indipendenti in un singolo repository (il data warehouse) dal quale gli utenti finali possano facilmente ed efficientemente eseguire query, generare report ed effettuare analisi

19

I sistemi di data warehousing



20

Data warehousing

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

21

Il data warehouse

Collezione di dati che soddisfa le seguenti proprietà:

- usata per il supporto alle decisioni
- orientata ai soggetti
- integrata: livello aziendale e non dipartimentale
- correlata alla variabile tempo: ampio orizzonte temporale
- con dati tipicamente aggregati: per effettuare stime
- fuori linea: dati aggiornati periodicamente

22

Il data warehouse

- **Orientata ai soggetti**: considera i dati di interesse ai soggetti dell'organizzazione e non quelli rilevanti ai processi organizzativi
- Le basi di dati operazionali sono costruite a supporto dei singoli processi operativi o applicazioni
  - produzione
  - vendita
- Il data warehouse è costruito attorno alle principali entità del patrimonio informativo aziendale
  - prodotto
  - cliente

23

Il data warehouse

- **Integrata**:
  - i dati provengono da tutte le sorgenti informative
  - il data warehouse rappresenta i dati in modo univoco, riconciliando le eterogeneità delle diverse rappresentazioni:
    - nomi
    - struttura
    - codifica
    - rappresentazione multipla

24

Il data warehouse

- **Correlata alla variabile tempo**: presenza di dati storici per eseguire confronti, previsioni e per individuare tendenze
- Le basi di dati operazionali mantengono il valore corrente delle informazioni

  ⇒ L'orizzonte temporale di interesse è dell'ordine dei pochi mesi
- Nel data warehouse è di interesse l'evoluzione storica delle informazioni

  ⇒ L'orizzonte temporale di interesse è dell'ordine degli anni

25

Il data warehouse

- **Dati aggregati**: nell'attività di analisi dei dati per il supporto alle decisioni
  - non interessa "chi" ma "quanti"
  - non interessa un dato ma la somma, la media, il minimo, il massimo di un insieme di dati
- Le operazioni di aggregazione sono fondamentali

26

Il data warehouse

- **Fuori linea**:
- In una base di dati operazionale, i dati vengono
  - acceduti, inseriti, modificati, cancellati

  pochi record alla volta
- Nel data warehouse, abbiamo
  - operazioni di accesso e interrogazione — "diurne"
  - operazioni di caricamento e aggiornamento dei dati — "notturne"

  che riguardano milioni di record

27

... una base di dati separata ...

- Per tanti motivi
  - non esiste un'unica base di dati operazionale che contiene tutti i dati di interesse
  - la base di dati deve essere integrata
  - non è tecnicamente possibile fare l'integrazione in linea
  - i dati di interesse sarebbero comunque diversi
    - devono essere mantenuti dati storici
    - devono essere mantenuti dati aggregati
  - l'analisi dei dati richiede per i dati organizzazioni speciali e metodi di accesso specifici
  - degrado generale delle prestazioni senza la separazione

28

Data warehouse

- La costruzione di un sistema di data warehousing non comporta l'inserimento di nuove informazioni ma la riorganizzazione di quelle esistenti e implica l'esistenza di un sistema informativo
- Il fatto che non vengano mai eliminati dati da un data warehouse e che gli aggiornamenti siano tipicamente eseguiti a freddo (fuori linea) permettono di considerare il data warehouse come un database a sola lettura
  - non c'è necessità di gestione transazionale
  - denormalizzaziione delle tabelle
  - analisi dinamica e interattiva vs esecuzione di applicazioni predefinite

29

Architettura di riferimento

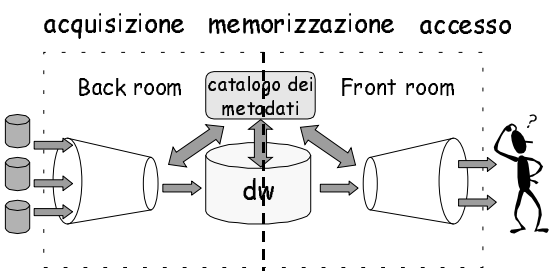
30

Caratteristiche architetturali irrinunciabili

- **Separazione:** l'elaborazione analitica e quella transazionale devono essere il più possibile separate
- **Scalabilità:** l'architettura hw e sw deve essere facilmente ridimensionabile
- **Estendibilità:** deve essere possibile accogliere nuove applicazioni e tecnologie
- **Sicurezza:** il controllo sugli accessi è essenziale (dati strategici)
- **Amministabilità:** l'attività di amministrazione non deve essere troppo complessa

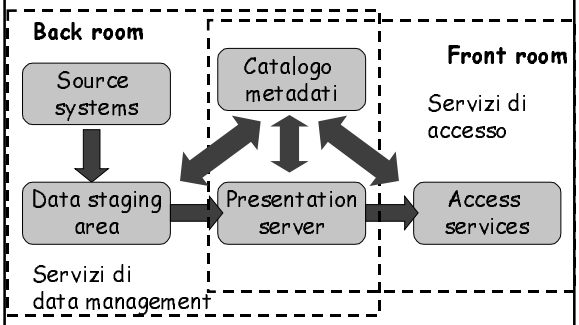
31

Architettura di base



32

Piu' in dettaglio ...



33

Source systems

- Ogni sorgente di informazioni aziendali
- Spesso rappresentate da dati operazionali: insieme di record la cui funzione è quella di catturare le transazioni del sistema organizzativo
- tipico accesso OLTP
- uso di production keys (non vengono usate nel DW)

34

Data staging

- Area di memorizzazione
  - i dati sorgente vengono trasformati
  - tecnologia relazionale ma anche flat files
- insieme di processi che:
  - puliscono, trasformano, combinano, duplicano, archiviano e preparano i dati sorgente per essere usati nel DW

35

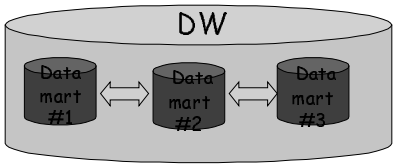
Presentation server

- Componente che permette la memorizzazione e la gestione del data warehouse, secondo un approccio dimensionale
- Può essere basato su:
  - tecnologia relazionale (ROLAP)
  - tecnologia multidimensionale (MOLAP)

36

Presentation server

- Un DW rappresenta spesso l'unione di più data mart
- **Data mart**: restrizione data warehouse ad un singolo processo o ad un gruppo di processi aziendali (es. Marketing)



37

End-user data access tools

- Client del DW, di facile utilizzo
- tools per interrogare, analizzare e presentare l'informazione contenuta del DW a supporto di un particolare bisogno aziendale
- invio specifiche richieste al presentation server in formato SQL

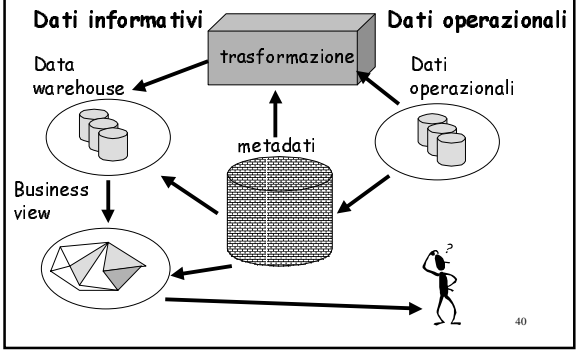
38

I metadati

- = dati sui dati
- Link tra i DB operazionali e il DW
- ogni passo eseguito durante la costruzione del DW genera metadati che possono poi essere utilizzati dalle fasi successive
- **Esempi**: schema, data in cui un dato è stato creato, quale tool l'ha creato, storia delle trasformazioni di un dato nel tempo, statistiche, dimensione tabelle, ecc. ecc.

39

I metadati



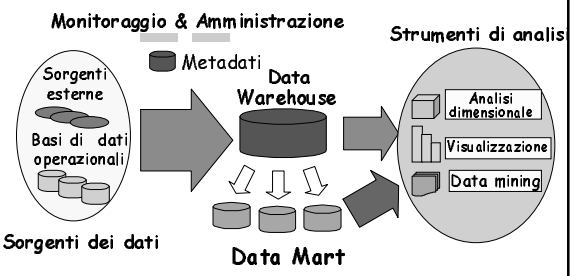
40

Due ritmi diversi ...

- Uso bimodale:
  - 16-22 ore al giorno usati per attività di interrogazione
    - funzionalità front room
  - 2-8 ore al giorno per caricamento, indicizzazione, controllo qualità e pubblicazione
    - funzionalità back room

41

Architettura base per il data warehousing



42

Attività per popolare un data warehouse

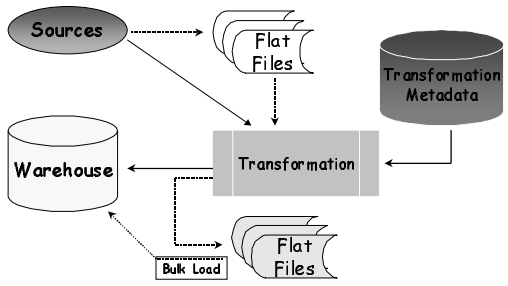
- Processo ETL: Extraction,Transformation, Loading
- Extraction
  - Estrazione dei dati dalle sorgenti informative operazionali
  - Opzioni: tutti i dati / solo dati modificati (incrementale)
- Transformation
  - Pulizia, per migliorare la qualità dei dati
  - Trasformazione di formato, da formato sorgente a quello del DW
  - Correlazione con oggetti provenienti da altre sorgenti
- Loading
  - Caricamento (refresh o update) con aggiunta di informazioni temporali e generazione di dati aggregati

Attività per popolare un data warehouse

- Il ruolo degli strumenti ETL è quello di alimentare una sorgente dati singola, dettagliata, esauriente e di alta qualità che possa a sua volta alimentare il DW
- in caso di architettura a tre livelli questi strumenti alimentano il livello dei dati riconciliati
- la riconciliazione avviene quando il DW viene popolato la prima volta e periodicamente quando il DW viene aggiornato

44

ETL Data Flow



Modelli per il data warehousing

46

Analisi multidimensionale

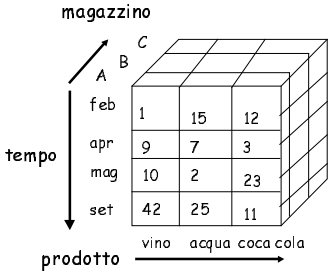
- L'analisi richiede normalmente dimensioni multiple:
  - "quanti items ho venduto
  - per regione
  - per mese
  - per tipo di cliente?"
- Dimensioni normalmente utilizzate per l'analisi:
  - Tempo
  - Prodotto
  - Cliente
  - Area geografica
  - Dipartimento/settore

Il modello Multidimensionale

- Un data warehouse si basa su un modello dei dati multidimensionale che rappresenta i dati sotto forma di data cube
- Un *data cube* permette di modellare e creare viste dei dati rispetto a molteplici dimensioni
- Modello dati multidimensionale
- Detto "Star Schema"
- Implementabile su un DB relazionale
- Consente volumi di dati molto grandi
  - volumi dell'ordine di 100 gbytes forniscono tempi di risposta sotto i 10 sec

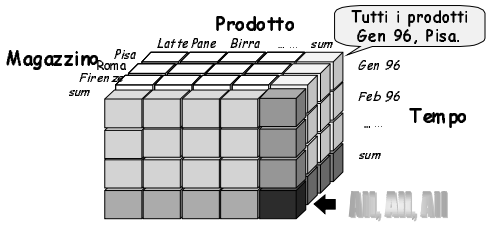
48

Rappresentazione multidimensionale dei dati



49

Data Cube



Ogni dimensione contiene una gerarchia di valori
una cella del cubo contiene valori aggregati (count, sum, max, etc.)

50

Concetti usati per definire un data cube

- **Fatto** un tema di interesse per l'organizzazione (vendite, spedizioni, acquisti)
- **Misura** una proprietà di un fatto da analizzare (numero di unità vendute, prezzo unitario)
- **Dimensione** descrive una prospettiva lungo la quale un'organizzazione vuole mantenere i dati (prodotto, negozio, data)

51

Modello dei dati multidimensionale

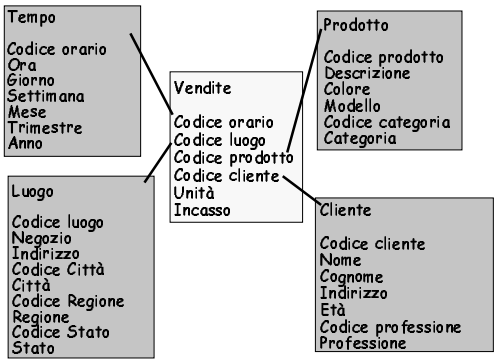
- Ogni dimensione può avere una tabella ad essa associata

Es. item (item_name, brand, type), or
time(day, week, month, quarter, year)

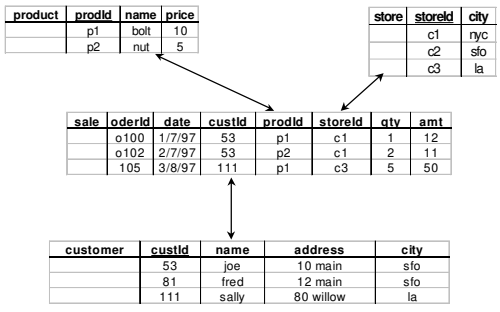
- La Fact table contiene le misure (come dollars_sold) e chiavi esterne per ogni dimension table

52

Organizzazione "star"



Esempio di star schema



product	prodId	name	price
	p1	bolt	10
	p2	nut	5

store	storeId	city
	c1	nyc
	c2	sfo
	c3	la

sale	oderId	date	custId	prodId	storeId	qty	amt
	o100	1/7/97	53	p1	c1	1	12
	o102	2/7/97	53	p2	c1	2	11
	105	3/8/97	111	p1	c3	5	50

customer	custId	name	address	city
	53	joe	10 main	sfo
	81	fred	12 main	sfo
	111	sally	80 willow	la

54

Dimensioni

- Sono le entità rilevanti per l'analisi
- Tipicamente sono caratterizzate da attributi testuali o discreti
- La dimensione temporale esiste sempre

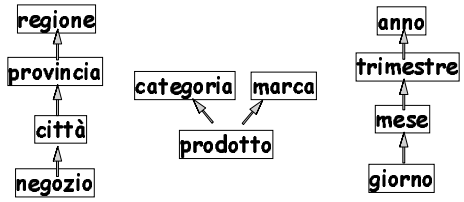
Esempio:

- vendite in una catena di supermercati
  Dimensioni: tempo, prodotti, magazzino
- iscrizioni universitarie
  Dimensioni: tempo, facoltà, tipologia studenti

55

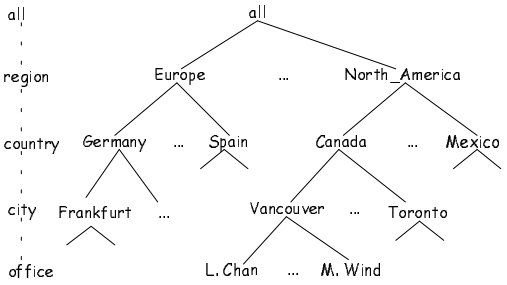
Dimensioni e gerarchie di livelli

- Ciascuna dimensione è organizzata in una gerarchia che rappresenta i possibili livelli di aggregazione per i dati



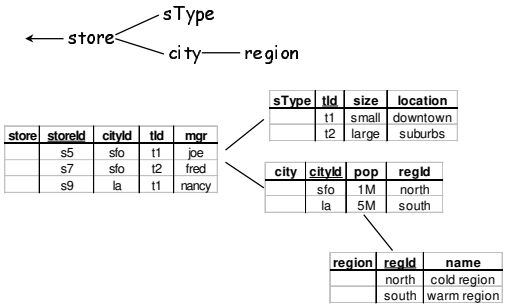
56

Gerarchia di concetti: Dimensione (luogo)



57

Esempio di DW con gerarchie



58

Misure

- Le misure sono tipicamente numeriche
- Esempio
  - Consideriamo le vendite in una catena di supermercati
  - Le misure possono essere
    - N. prodotti venduti
    - Incassi
    - Costi
    - .....

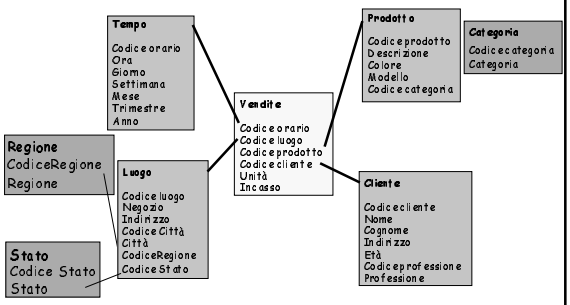
59

Esempi di fatti/misure/dimensioni

- Catena di negozi
  - vendita
  - quantità venduta, incasso
  - prodotto, tempo, zona
- Compagnia telefonica
  - telefonata
  - costo, durata
  - chiamante, chiamato, tempo

60

Generalizzazione: Organizzazione "snowflake"



Accedere al DW: reportistica, OLAP, data mining

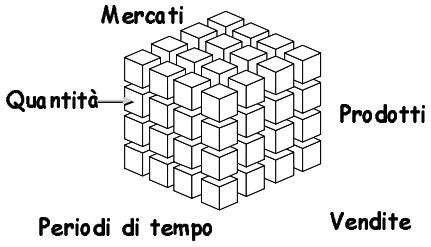
Reportistica

- Approccio orientato ad utenti che hanno necessità di accedere a intervalli di tempo predefiniti a informazioni strutturate in modo pressochè invariabile
- di questi rapporti è nota a priori la forma
- un rapporto è definito da un'interrogazione e da una presentazione
- l'interrogazione comporta in genere la selezione e l'aggregazione di dati multidimensionali
- la presentazione può essere in forma tabellare o grafica
- la reportistica non è nata con il DW, ma ha acquisito con il DW benefici in termini di affidabilità e tempestività dei risultati

OLAP: On-Line Analytical Processing

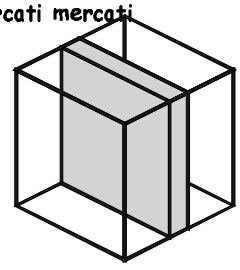
- Una visione multidimensionale, logica, dei dati
- Analisi interattiva dei dati
- Modellazione analitica: derivazione delle proporzioni, delle varianze, etc
- Aggregazioni per ogni sottoinsieme delle dimensioni
- Previsione, trend analysis, e statistical analysis
- Calcola e visualizza i dati in 2D o 3D crosstabs, charts, e grafi, con semplici operazioni di rotazione degli assi

OLAP su data cubes



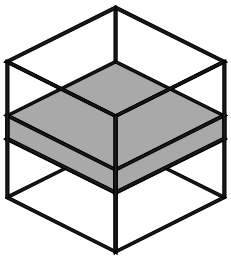
Analisi per segmento di mercato

Il manager regionale esamina la vendita dei prodotti in tutti i periodi relativamente ai propri mercati mercati



Analisi per prodotto

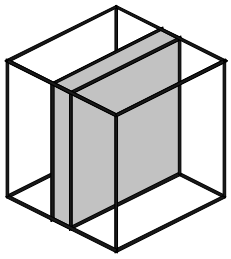
Il manager di prodotto esamina la vendita di un prodotto in tutti i periodi e in tutti i mercati



67

Analisi per periodo di tempo

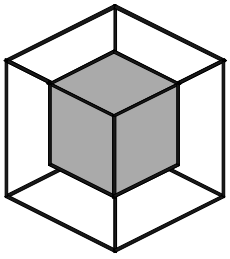
Il manager finanziario esamina la vendita dei prodotti in tutti i mercati relativamente al periodo corrente e quello precedente



68

Analisi multidimensionale

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



69

Operazioni tipiche su data cubes

- Roll up: riassumi i dati, salendo nella gerarchia dei concetti per una dimensione o attraverso una riduzione di una dimensione
  - il volume totale di vendite per categoria di prodotto e per regione
  - si rimuove per esempio la dimensione tempo
- Roll down or drill down: passa da un livello di dettaglio basso ad un livello di dettaglio alto, scendendo nella gerarchia o introducendo una nuova dimensione.
  - per un particolare prodotto, trova le vendite dettagliate per ogni venditore e per ogni data

70

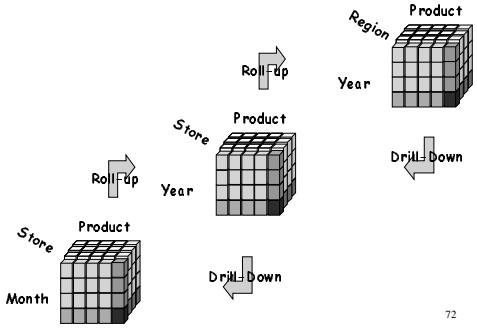
Operazioni tipiche (cont.)

- Slice and dice: select & project
  - L'operazione di Slice esegue una selezione su una dimensione del cubo.
  - L'operazione di Dice definisce un sottocubo eseguendo una selezione su due o più dimensioni

  Vendite delle bevande nel West negli ultimi 6 mesi
- Pivot (rotate): riorienta il cubo

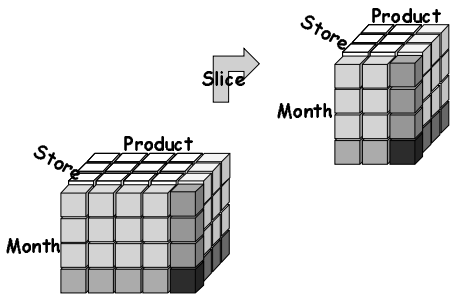
71

Operazioni tipiche: Roll-Up



72

Operazioni tipiche: Slice and Dice



73

Data mining

- Attività orientata a scoprire informazioni nascoste nei dati
- le tecniche di data mining sono utilizzate da anni in applicazioni scientifiche specialistiche (ricerca geologica, medica, astronomica, metereologica, …)
- con il DW il data mining viene trasportato dall'analisi scientifica all'analisi commerciale (ricerche di mercato, segmentazione di mercato, analisi delle abitudini di acquisto, …)
- permette di analizzare automaticamente grosse quantità di dati
- tipologie di pattern estraibili con regole di data mining: regole associative, clustering, alberi di decisione, serie temporali

74

Progettazione di un data warehouse

75

Fattori di rischio

- Tipiche ragioni di fallimento dei progetti di data warehousing:
- Rischi legati alla gestione del progetto
  - necessità di condivisione di informazione tra i reparti
  - definizione dell'ambito e delle finalità del sistema
- Rischi legati alle tecnologie (rapida evoluzione)
- Rischi legati ai dati e alla progettazione
  - qualità dei dati e del progetto realizzato
- Rischi legati all'organizzazione
  - difficoltà di trasformare la cultura aziendale, inerzia organizzativa

76

Metodologie di progettazione

- Approcci top-down e bottom-up
- Approccio top-down
  - + visione globale dell'obiettivo
  - + DW consistente e ben integrato
  - – costi onerosi e lunghi tempi di realizzazione (rischio di scoraggiare la direzione)
  - – complessità dell'analisi e riconciliazione contemporanea di tutte le sorgenti
  - – impossibilità di prevedere a priori nel dettaglio le esigenze delle diverse aree aziendali
  - – impossibilità di prevedere la consegna a breve termine di un prototipo

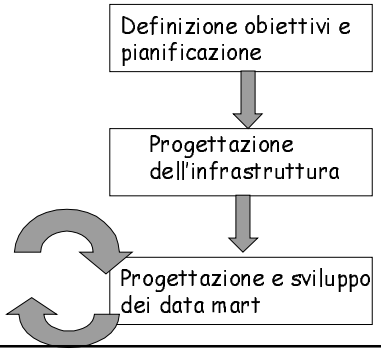
77

Metodologie di progettazione

- È quindi preferibile l'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
- abbinando questo approccio a tecniche di prototipazione veloce si riducono notevolmente tempi e costi necessari per fornire un riscontro sull'effettiva utilità del sistema alla dirigenza aziendale
- rischio: 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 e deve ricoprire un ruolo centrale per l'intero DW

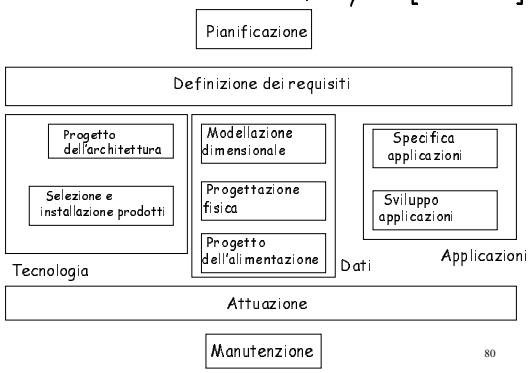
78

Metodologie di progettazione



79

Business Dimensional Lifecycle [Kimball]



80

La progettazione di un data mart

- Fasi
- Analisi e riconciliazione delle fonti dati
  - input: schema delle sorgenti
  - output: schema riconciliato
- Analisi dei requisiti
  - input: schema riconciliato
  - output: fatti, carico di lavoro preliminare
- Progettazione concettuale
  - input: schema riconciliato, fatti, carico di lavoro preliminare
  - ouput: schemi di fatto
- Raffinamento del carico di lavoro, validazione dello schema concettuale
  - input: schemi di fatto, carico di lavoro preliminare
  - ouput: carico di lavoro, schemi di fatto validati

81

La progettazione di un data mart

- Progettazione logica
  - input: schema di fatto, modello logico target, carico di lavoro
  - output: schema logico del data mart
- Progettazione dell'alimentazione
  - input: schemi delle sorgenti, schema riconciliato, schema logico del data mart
  - output: procedure di alimentazione
- Progettazione fisica
  - input: schema logico del data mart, DBMS target, carico di lavoro
  - output: schema fisico del data mart

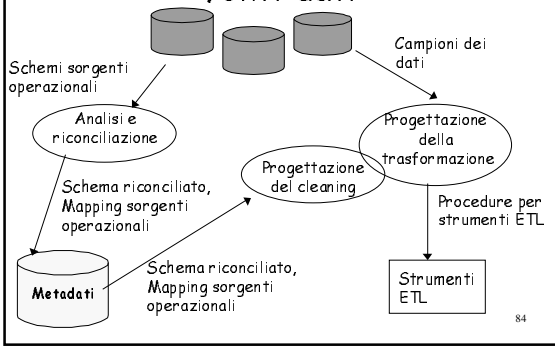
82

La progettazione di un data mart

- Aspetto chiave: basare la modellazione dei data mart sugli schemi operazionali
- uno schema concettuale di massima per il data mart può essere derivato dal livello dei dati riconciliati
- per questo motivo la fase di analisi e riconciliazione delle fonti avviene prima della fase di analisi dei requisiti utente
- se queste due fasi sono invertite i fatti, le misure e le gerarchie vengono ricavate dalle specifiche utente e solo a posteriori si verifica che le informazioni richieste siano effettivamente disponibili nei database operazionali
- rischio di minare la fiducia del cliente verso il progettista

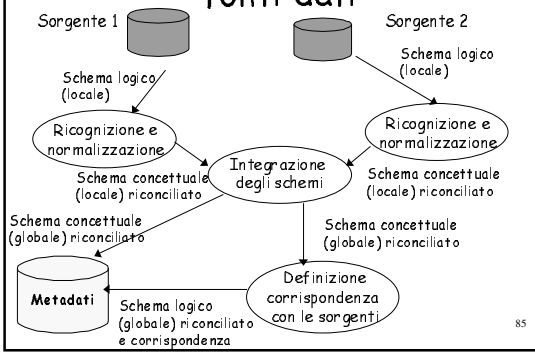
83

Analisi e riconciliazione delle fonti dati



84

Analisi e riconciliazione delle fonti dati



85

Analisi e riconciliazione delle fonti dati

- **Ricognizione:** Esame approfondito degli schemi locali mirato alla piena comprensione del dominio applicativo
- **normalizzazione:** correzione degli schemi locali per modellare in modo più accurato il dominio applicativo (Fasi da svolgere anche se sorgente dati unica)
- **integrazione:** v. quanto detto su integrazione di schemi concettuali
- **definizione delle corrispondenze:** il risultato finale è lo schema riconciliato in cui sono risolti i conflitti e l'insieme delle corrispondenze tra gli elementi degli schemi sorgenti e quelli dello schema riconciliato

86

Analisi dei requisiti utente

- Interviste vs riunioni coordinate
- interviste: fasi preliminari
  - ricerca pre-intervista
  - selezione degli intervistati
  - preparazione dei questionari
  - pianificazione delle interviste
  - preparazione degli intervistati
- interviste: tipologie
  - a risposte aperte
  - a risposte chiuse
  - a risposte probatorie

87

Analisi dei requisiti utente

- interviste: strutturazione
  - a piramide
  - a imbuto
- output della fase di analisi:
  - tabella delle derivazioni (correlazione di ogni attributo con le sorgenti operazionali)
  - tabella degli utilizzi (descrizione testuale di ciascun attributo)
  - tabella di struttura (attributo modellato come dimensione, attributo di dimensione, misura)
  - identificazione dei fatti (categoria di eventi, con aspetti dinamici) e scelta della loro granularità
  - carico di lavoro preliminare (specifica in linguaggio naturale delle interrogazioni di analisi)

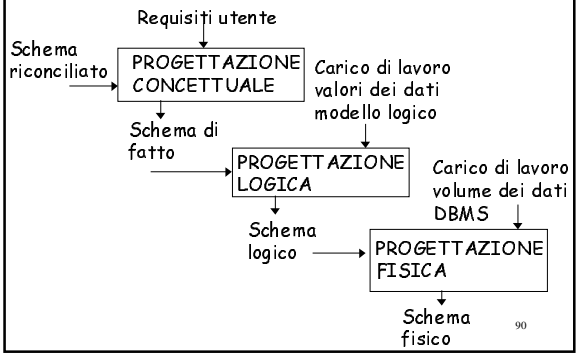
88

Le fasi della progettazione di un data mart

- Progettazione concettuale:
  - fornisce una rappresentazione formale del contenuto informativo del data mart
  - indipendente dal sistema che verrà utilizzato per la sua implementazione
- progettazione logica:
  - lo schema concettuale viene tradotto nel modello dei dati del sistema prescelto
- progettazione fisica:
  - fase in cui vengono scelte le caratteristiche fisiche del sistema

89

Le fasi della progettazione di un data mart



90

Progettazione concettuale di un data warehouse

91

Progettazione concettuale

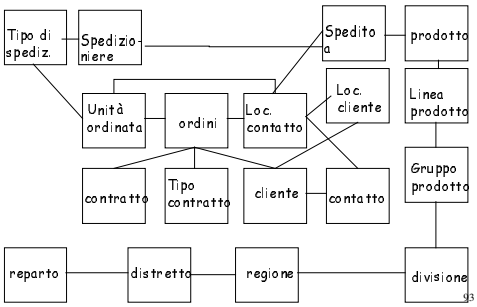
OLTP

- modello entità-relazione
- si cerca di eliminare il più possibile la ridondanza
  - maggiore efficienza delle operazioni di aggiornamento
- schema simmetrico: tutte le entità hanno la stessa importanza
- ci possono essere molti modi per connettere (mediante un'operazione di join) due tabelle
- la rappresentazione dipende dalla struttura dei dati

92

Progettazione concettuale

OLTP: un esempio

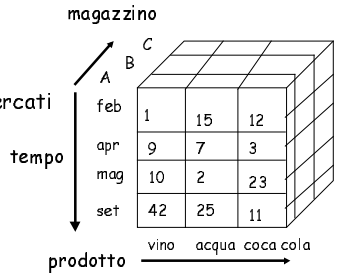


93

Progettazione concettuale

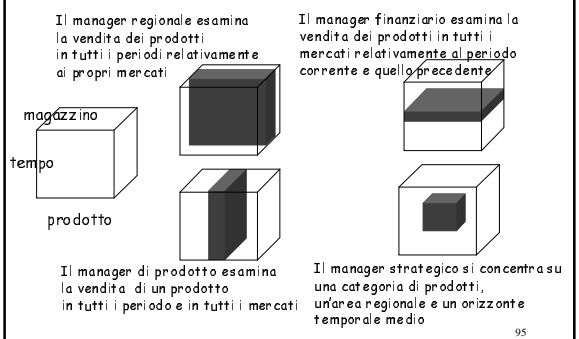
OLAP

Processo:
vendite in una catena di supermercati



94

Progettazione concettuale



95

Progettazione concettuale

OLAP

- Ogni parametro può essere organizzato in una gerarchia che ne rappresenta i possibili livelli di aggregazione:
  - negozio, città, provincia, regione
  - giorno, mese, trimestre, anno

96

Progettazione concettuale

OLAP

- L'eliminazione della ridondanza non è un obiettivo
  - non si devono eseguire operazioni di aggiornamento
  - schemi denormalizzati
- schemi asimmetrici: alcune entità sono più importanti di altre
- un solo modo per connettere (mediante un'operazione di join) due tabelle
  - minore numero di join
  - maggiore efficienza
- la rappresentazione dipende dalla struttura dei dati

97

Progettazione concettuale

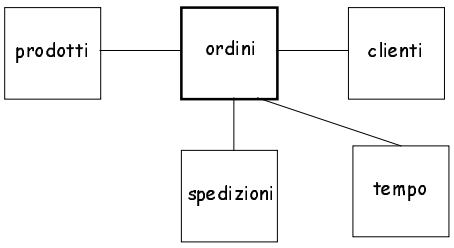
I modelli multidimensionali

- Vari modelli multidimensionali:
  - modello a stella
  - modello a costellazione di fatti
  - modello a fiocco di neve (snowflake)
- possono essere implementati in
  - sistemi relazionali
  - sistemi multidimensionali
  - sistemi ad oggetti
- In genere: implementazione diretta in DBMS relazionali

98

Progettazione concettuale

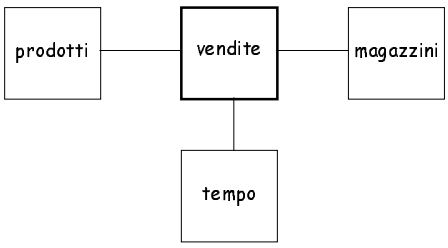
OLAP: un esempio



99

Progettazione concettuale

OLAP: un esempio



100

Il modello a stella

- Nel modello dimensionale, non sono rilevanti i singoli eventi (transazioni) ma il loro accadere durante un determinato intervallo temporale ⇒ granularità dello schema
- Il modello a stella è basato sull'esigenza di vedere dati di dettaglio (fatti) in funzione di più dimensioni

101

Il modello a stella

- **Fatti**: identificano l'attività principale e sono caratterizzati dai dati di dettaglio che si desidera analizzare
- **Dimensioni**: parametri che influenzano i dati di dettaglio e rispetto ai quali si analizzano tali dati
- Fatti e dimensioni collegati attraverso chiavi esterne
  - in generale, uno schema a stella rappresenta una relazione molti a molti
    - il collegamento tra ogni tabella delle dimensioni e la tabella dei fatti rappresenta una relazione uno a molti

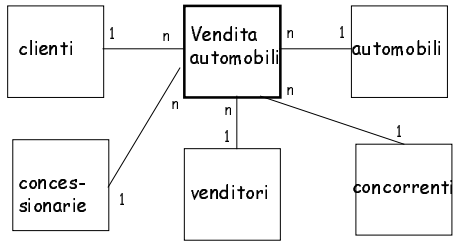
102

Esempio

- Attività principale: vendita automobili
- dimensioni:
  - clienti
  - venditori
  - concorrenti
  - automobili
  - concessionarie

103

Esempio



104

Le dimensioni

- Devono essere scelte solo le entità rilevanti per le analisi che si intendono effettuare
- Le dimensioni sono tipicamente caratterizzate da attributi:
  - testuali
  - discreti

  ma possono anche essere numeriche
  - dimensione di un prodotto
- esiste sempre una dimensione temporale

105

Dimensioni: esempi

- Attività: vendita in una catena di supermercati
  - dimensioni: tempo, prodotti, magazzino
- Attività: ordini
  - dimensioni: tempo, prodotti, clienti, spedizioni
- Attività: iscrizioni universitarie
  - dimensioni: tempo, facoltà, tipologia studenti
- Attività : vendita automobili
  - dimensioni: clienti, venditori, concorrenti, automobili, concessionarie

106

Le dimensioni

- Problema: come si può identificare se un attributo numerico è un fatto o un attributo di una dimensione?
- se è una misura che varia continuamente nel tempo
  - fatto
    - analisi costo di un prodotto nel tempo
- se è una descrizione discreta di qualcosa che è ragionevolmente costante
  - attributo di una dimensione
    - costo di un prodotto visto come informazione descrittiva

107

Le dimensioni

- Le dimensioni utilizzate sono spesso le stesse in vari contesti applicativi:
  - tempo
  - collocazione geografica
  - organizzazione
  - clienti
- il numero di attributi per ogni dimensione è in genere molto elevato (anche nell'ordine del centinaio)

108

La dimensione tempo

- È presente in ogni DW in quanto virtualmente ogni DW rappresenta una serie temporale
- **Domanda**: perché non campo di tipo DATE nella tabella dei fatti?
- **Risposta**: la dimensione tempo permette di descrivere il tempo in modi diversi da quelli che si possono desumere da un campo date in SQL (giorni lavorativi-vacanze, periodi fiscali, stagioni, ecc.)

109

La dimensione tempo

- Alcuni tipici attributi della dimensione tempo:
  - tempo-k (può essere un campo di tipo data in SQL)
  - giorno-della-settimana
  - n-giorno-nel-mese
  - n-giorno-in-anno
  - n-settimana-in-anno
  - mese
  - stagione
  - periodo fiscale
  - ...

110

I fatti

- La tabella dei fatti mette in evidenza una relazione molti-a-molti
- I fatti hanno delle proporietà che sono dette misure
- Le proprietà dei fatti sono tipicamente:
  - numeriche
  - additive
- Numerici, additivi: possono essere aggregati rispetto agli attributi delle dimensioni, utilizzando l'operazione di addizione

111

Fatti e misure: esempi

- Attività (fatti): vendite in una catena di supermercati
  - misure: n. prodotti venduti, incassi, costi, ...
- Attività (fatti): ordini
  - misure: n. spedizioni, n. clienti, importi, ...
- Attività (fatti): iscrizioni universitarie
  - misure: n. studenti, ...

112

Additività delle misure

- Incasso, unità vendute: sono additive in quanto si possono aggregare sommando rispetto ad ogni dimensione:
  - somma incassi/unità su tempo
  - somma incassi/unità su prodotti
  - somma incassi/unità su dipartimenti

113

Semiadditività delle misure

- Numero clienti non è una misura additiva:
  - somma n. clienti su tempo OK
  - somma n. clienti su dipartimenti OK

  MA:
  - somma n. clienti su prodotto genera problemi
  - si supponga che
    - clienti che hanno comprato carne 20
    - clienti che hanno comprato pesce 30
  - il numero di clienti che hanno comprato carne o pesce è un qualunque numero tra 30 e 50

114

Semiadditività delle misure

- Il numero clienti è una misura semiadditiva, poiché può essere sommata solo rispetto ad alcune dimensioni
- Soluzione: cambiare la granularità del database, portandola a livello singola transazione

115

Semiadditività delle misure

- Tutte le misure che memorizzano una informazione statica, quali:
  - bilanci finanziari
  - misure di intensità (temperatura di una stanza)

  sono semiadditive rispetto al tempo
- ciò che comunque si può fare è calcolare la media su un certo periodo di tempo

116

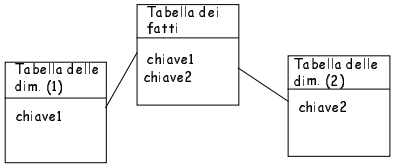
Non additività delle misure

- Le misure non additive sono misure che non possono essere sommate
- Esempi:
  - misure: costo unitario e quantità nel contesto di un ordine
  - dimensioni: clienti, spedizioni, tempo, …
  - i costi unitari non possono essere sommati se prima non sono moltiplicati per le rispettive quantità, quindi tali costi sono misure non additive

117

Collegamenti tra dimensioni e fatti

- Avviene tramite chiavi esterne
- ogni tabella delle dimensioni ha una chiave
- la tabella dei fatti deve contenere come attributi la chiave di ciascuna dimensione
- tali attributi sono chiavi esterne e, complessivamente, rappresentano la chiave della tabella dei fatti



118

Vantaggi nell'uso dello schema a stella

- Permette di fare assunzioni circa i dati, da utilizzare in fase di ottimizzazione
- simmetrico
- facilmente estensibile
- approcci standard alla costruzione
- tipiche query eseguibili efficientemente (vedi oltre)

119

Metodologia di progettazione

- La progettazione dipende da:
  - requisiti utente
    - si determinano attraverso intervistecon gli utenti finali
  - dati disponibili
    - si determinano analizzando la documentazione esistente e attraverso interviste con i DBA

120

Passi nel processo di progettazione

- Scegliere il processo aziendale che si intende modellare (fatti)
- scegliere la granularità con la quale si intende modellare il processo aziendale
- scegliere le dimensioni e i loro attributi
- scegliere le misure dei fatti

121

Un esempio per capire

La gestione di una catena di supermercati

- 500 supermercati distribuiti su un'area che comprende 3 stati negli USA
- ogni supermercato è composto da diversi reparti
- ogni reparto vende molti prodotti, identificati da stock keeping unit (SKU)
- i dati delle vendite vengono raccolti nei punti di vendita (casse)
- i prodotti sono spesso soggetti a promozioni di vendita
- problematiche che si intendono analizzare:logistica degli ordini, massimizzazione profitti in ciascun supermercato

122

Passo 1: scelta del processo aziendale

- Si deve decidere quale processo modellare, combinando la conoscenza aziendale con la conoscenza di quali dati sono disponibili
- Esempio
  movimento giornaliero delle varie unità

123

Passo 2: scelta della granularità

- La granularità identifica il contenuto della tabella dei fatti nel processo considerato
- è importante perché:
  - determina le dimensioni del database
  - condiziona la dimensione del database
- Esempio
  SKU per magazzino per promozione per giorno (granularità a livello di giorno e di singola unità)

124

Passo 2: scelta della granularità

- Perché giornaliero:
  - granularità a livello transazione (operazione di acquisto)
    - il database diventerebbe enorme e quindi ingestibile
  - granularità a livello settimana o mese:
    - molti effetti delle vendite non sarebbero visibili
    - ad esempio: differenza in vendite tra Lunedì e Sabato

125

Passo 2: scelta della granularità

- Perché a livello singola unità:
  - granularità a livello pacco:
    - non sarebbe più possibile rispondere a domande quali:
      - le vendite di quali prodotti si riducono quando un certo prodotto viene messo in promozione di vendita?
      - se confrontiamo le vendite con quelle della concorrenza, quali sono i 10 prodotti che la concorrenza vende e noi non vendiamo?

126

Passo 3: scelta delle dimensioni

- Una definizione accurata della granularità comporta immediatamente la definizione delle dimensioni principali del DW (dimensioni primarie)
- è quindi possibile aggiungere altre dimensioni, purchè queste dimensioni assumano un singolo valore per ogni combinazione delle dimensioni primarie

127

Passo 3: scelta delle dimensioni

- Esempio:
  nel nostro esempio, le dimensioni primarie sono:
  - tempo
  - prodotti
  - magazzini

  dimensioni aggiuntive
  - promozioni

128

Passo 3: scelta delle dimensioni

- Per ciascuna dimensione, devono essere specificati gli attributi che la caratterizzano
- spesso si tratta di attributi alfanumerici
- se una dimensione contiene attributi non correlati, è meglio suddividere la dimensione in due dimensioni distinte

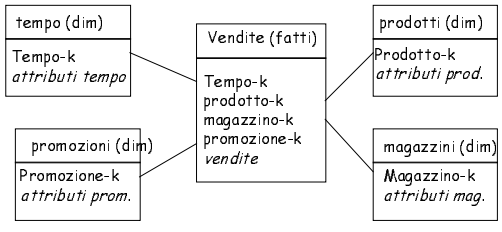
129

Passo 4: scelta delle misure

- Misure tipici sono numerici e additivi
- Esempio: alcuni utili misure:
  - incasso
  - unità vendute
  - numero clienti
- nel caso di granularità transazionale, l'unica misura in genere rilevante è una quantità (livello di granularità più fine)

130

Schema iniziale per l'esempio considerato



Per il momento: assumiamo che lo schema precedente rappresenti sia lo schema logico che lo schema fisico del database, quindi il database contiene 5 tabelle

131

Osservazioni sulla normalizzazione

- La tabella dei fatti è completamente normalizzata
- le tabelle delle dimensioni possono non essere normalizzate, ma:
  - la dimensione delle tabelle delle dimensioni è in genere irrilevante rispetto alla dimensione della tabella dei fatti
  - quindi, ogni sforzo per normalizzare queste tabelle ai fini del DW è una perdita di tempo
  - lo spazio guadagnato è in genere meno dell'1% dello spazio richiesto dallo schema complessivo
- la normalizzazione delle tabelle delle dimensioni può ridurre la capacità di browsing (navigazione) dello schema (si veda oltre)

132

Gerarchie

- Molto spesso una singola dimensione può contenere dati organizzati gerarchicamente
- Esempio: Ogni unità può essere rappresentata all'interno di una gerarchia:
  - sku
  - pacco
  - marca
  - sottocategoria
  - categoria
  - dipartimento

Gerarchie

- Tutti gli attributi della gerarchia devono essere inseriti all'interno della dimensione

⇒ ridondanza accettabile in quanto la dimensione delle tabelle delle dimensioni in genere è inifluente sulla dimensione totale del database

- Esempio: 30000 prodotti distinti
  30 dipartimenti distinti
  ⇒ in media 1000 ripetizioni

Esempio

Prodotti (dim)
Prodotto-k descrizione-SKU numero-SKU tipo-pacco marca sottocategoria categoria dipartimento tipo-pacco peso unità-di-misura ...

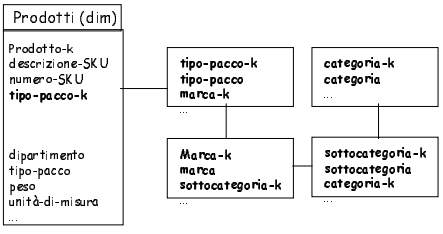
- La tabella dei prodotti è una delle tabelle principali di quasi tutti i DW
- è utile inserire più attributi possibile

Gerarchie

- Gerarchia tipica: gerarchia geografica:
  - comune
  - provincia
  - regione
  - stato
  - continente
- una dimensione può contenere attributi relative a gerarchie multiple

Schemi snowflake

- Una gerarchia rappresenta molte relazioni molti-a-uno
  - si potrebbe pensare di utilizzare una tabella per ogni relazione ⇒ schema snowflake



Schemi snowflake

- Uno schema snowflake rende meno efficienti le operazioni di ricerca, anche se la tabella è grande (+ join)
- è conveniente utilizzare uno schema snowflake solo se questo approccio aumenta la leggibilità dello schema e le prestazioni globali

Esempio

- Dim. tabella dei fatti: 30 GB
- dim. indice tabella dei fatti 20 GB
- dim. max tabella delle dim. 0.1 GB
- usando schema snowflake 0.005 GB
- dim DB senza snowflake 50 GB
- dim DB con snowflake 50 GB

139

Riassumendo

- Concetti base del modello concettuale
- **Fatto:** un fatto è un concetto di interesse per il processo decisionale; tipicamente modella un insieme di eventi che accadono nell'impresa
- **Misura:** una misura è una proprietà numerica di un fatto e ne descrive un aspetto quantitativo di interesse per l'analisi
- **Dimensione:** una dimensione è una proprietà con dominio finito di un fatto e ne descrive una coordinata di analisi

140

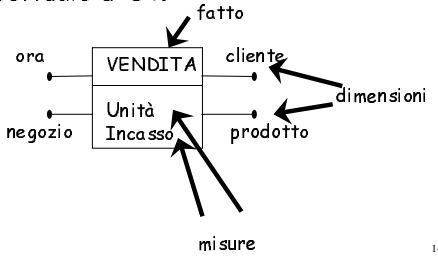
Riassumendo

- **Evento primario:** 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
- **Attributo di dimensione:** proprietà a valori discreti che descrive una dimensione
- **Gerarchia:** un albero i cui nodi sono dimensioni e loro attributi e i cui archi modellano associazioni molti-a-uno tra i nodi (diversi livelli di granularità di una dimensione)

141

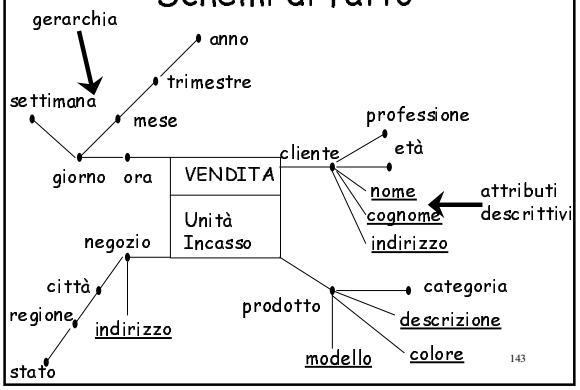
Schemi di fatto

- Una notazione grafica per la modellazione concettuale di DW



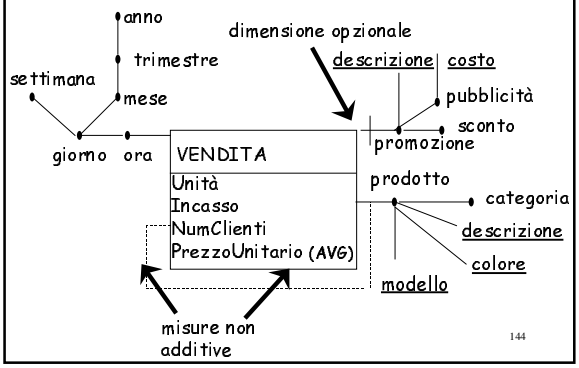
142

Schemi di fatto



143

Schemi di fatto



144

Progettazione concettuale avanzata di un data warehouse

Composizione degli schemi

- Lo schema risultante da ogni processo aziendale può essere visto come lo schema associato ad uno specifico data mart
- problema: combinare i fatti e le dimensioni contenuti negli schemi associati a ciascun processo, cioè contenuti in ciascun data mart

Composizione degli schemi

- Gli schemi associati ai vari processi possono avere dimensioni a comune
  - Una singola tabella delle dimensioni può essere usata in relazione a diverse tabelle dei fatti
- per potere passare dalle informazioni contenute in uno schema alle informazioni contenute in un altro (drill-across): le dimensioni con lo stesso nome devono avere lo stesso significato e contenere gli stessi attributi
  - dimensioni conformate
- Conseguenza: i vincoli su attributi delle dimensioni a comune devono restituire le stesse entità per ogni schema considerato

Esempio: catena di produzione

- inventario dei prodotti
  - dimensioni: tempo, prodotti, warehouse
- spedizione ai centri di distribuzione
  - dimensioni: tempo, prodotti, warehouse, centri di distribuzione, contratti, tipi di spedizione
- inventario del centro di distribuzione
  - dimensioni: tempo, prodotti, centri di distribuzione
- distribuzione ai magazzini
  - dimensioni: tempo, prodotti, centri di distribuzione, magazzini, contratti, tipi di spedizione
- inventario dei magazzini
  - dimensioni: tempo, prodotti, magazzini
- vendite
  - dimensioni: tempo, prodotti, magazzini, promozioni, clienti

Composizione degli schemi: eccezione

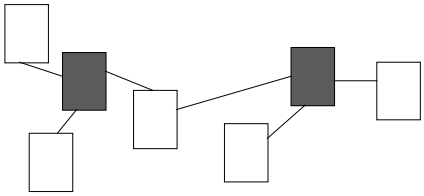
- Eccezione: la stessa dimensione può comparire in schemi diversi con un sottoinsieme di attributi (diversa conoscenza di un particolare aspetto applicativo)
  - drill-across si può fare solo sugli attributi in comune
- **Esempio**: i produttori conoscono i prodotti ad un livello di dettaglio maggiore rispetto a quello noto ai venditori, ma il tipo di prodotto comparirà in entrambe le dimensioni

Fatti conformati

- Anche i fatti devono essere conformati
  - fatti con lo stesso nome in tabelle diverse hanno la stessa granularità e le stesse unità di misura
  - stesso periodo temporale
  - stesso riferimento geografico

Costellazione di fatti

- Schema risultante:
  - costellazione di fatti



151

Aggregazione

- In alcune situazioni, non si hanno vincoli su tutte le dimensioni ma solo per alcune
- Esempi:
  - qual è il rapporto tra vendite effettuate nei week-end e vendite effettuate nei giorni lavorativi in ogni magazzino?
  - quale prodotto è stato maggiormente venduto negli ultimi 3 mesi?
- L'esecuzione di queste interrogazioni è molto costosa se viene effettuata sui dati di base
  - Idea: **precalcolare aggregati**

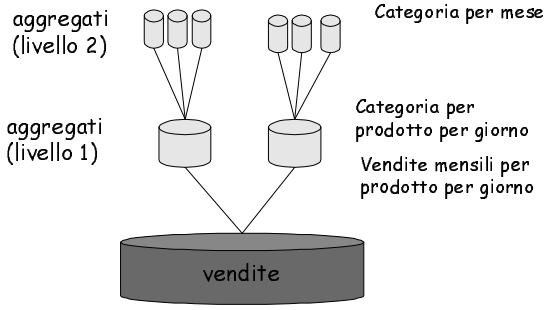
152

Aggregazione

- Un aggregato è un record di una tabella dei fatti che rappresenta una sintesi di vari record contenuti nella tabella dei fatti di base
- una tabella dei fatti aggregata è sempre associata ad una o più tabelle delle dimensioni aggregate
- un aggregato viene utilizzatoper due motivi:
  - efficienza
  - impossibilità di rappresentare gli stessi dati al livello di dettaglio
  - esempio: costi di promozione possono essere espressi a livello categoria e non a livello di singolo prodotto

153

Esempio



154

Due problemi

- Quali dati aggregare?
- Come e dove memorizzare i dati aggregati?

155

Quali dati aggregare?

- È importante considerare:
  - **tipiche richieste aziendali**
    - distribuzione geografica, linee di prodotti, periodicità generazione reportistica
    - per ogni dimensione, identificare gli attributi e le combinazioni di attributi che può essereutile aggregare
  - **distribuzione statistica dei dati**
    - stimare la dimensione delle tabelle aggregate
    - se la dimensione della tabella aggregata non riduce di molto la dimensione della tabella di partenza, forse non conviene aggregare
    - aggregazioni non molto usate possono essere utili come punto di partenza per effettuare altre aggregazioni più significative

156

Come e dove memorizzare i dati aggregati?

- Esistono due approcci di base:
  - nuove tabelle dei fatti
    - vengono create nuove tabelle per i fatti e le dimensioni aggregate
  - nuovo campo
    - vengono aggiunti nuovi campi nelle tabelle dei fatti e delle dimensioni
- Vedremo solo il primo approccio

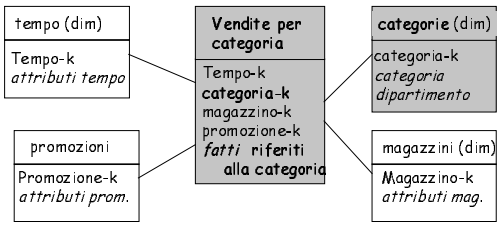
157

Nuove tabelle dei fatti

- Per ogni aggregato di interesse viene generata una nuova tabella dei fatti
- si generano tabelle delle dimensioni derivate da quelle di base ma contenenti solo i dati di interesse per la tabella dei fatti aggregata
  - generazione di chiavi artificiali per le tabelle delle dimensioni aggregate

158

Esempio



159

Nuove tabelle dei fatti

- L'uso di tabelle dei fatti e delle dimensioni aggregate accelera anche l'esecuzione di interrogazioni rispetto ad attributi che generalizzano (in base ad opportune gerarchie) l'attributo aggregato
- Esempio:
  - interrogazioni sui dipartimenti partendo dagli aggregati di categoria
- rispetto a questi attributi, si può evitare di costruire tabelle aggregate ad hoc

160

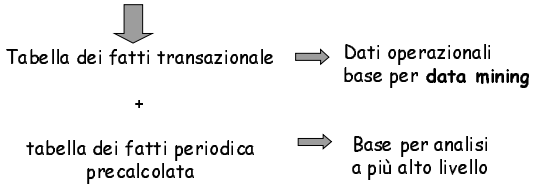
Vantaggi e svantaggi nell'uso degli aggregati

- Svantaggi:
  - L'uso degli aggregati aumenta di molto la dimensione del DB (anche del 300%!)
  - usare aggregazione nel caso in cui ogni aggregato sintetizza almeno 10-20 record di base
- Vantaggi:
  - Miglioramento delle prestazioni
  - possono essere utilizzati in modo trasparente all'utente

161

Granularità transazionale e snapshot periodico

- Nel caso di granularità transazionale, potrebbe capitare di avere anche bisogno di informazioni sintetiche periodiche, ad esempio mensili



162

Progettazione logica di un data warehouse

163

Scelta sistema di gestione dei dati

- DBMS operazionale: in genere relazionale
- DBMS informativo:
  - relazionale (Oracle 8/8i, RedBrick-Informix,…) (ROLAP)
  - multidimensionale (Oracle Express Server) (MOLAP)

164

ROLAP & MOLAP

- ROLAP:
  - sistema di data warehouse in grado di supportare le interrogazioni tipiche
  - presentation server relazionale
    - Oracle 8i + Discoverer
- MOLAP:
  - sistema di data warehouse in grado di supportare le interrogazioni tipiche
  - presentation server multidimensionale
    - Express Server
- DOLAP (Desktop OLAP):
  - i dati vengono recuperati da un DW relazionale o multidimensionale e copiati localmente
    - Business Objects

165

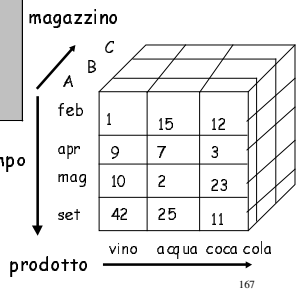
DBMS relazionali

- Tecnologia consolidata
- molto efficienti su dati di dettaglio
- estesi in modo da permettere la materializzazione degli aggregati
  - (Oracle 8i)
- performance
- scalabilità
- general-purpose

166

DBMS multidimensionali

vendite	prodotto	mese	magazzino
1	vino	febbraio	A
2	acqua	febbraio	B
3	coca cola	aprile	A
4	acqua	maggio	A
5	acqua	settembre	C
…	…	…	…



167

DBMS multidimensionali

- Modello dei dati basato su ipercubi (array multidimensionali)
- precalcolo aggregazioni
- aumento prestazioni per le query utente ma
  - … no join
  - … no interfaccia SQL (API)
  - … necessità sistema relazionale per dati dettaglio
  - … problema sparsità dei dati
  - … file molto grandi
  - … limitazioni a circa 10GB (problemi scalabilità)
- Per superare questi problemi:
  - aggiunta capacità di navigare da un MDBMS ad un RDBMS

168

ROLAP & MOLAP

- Performance
  - Query: MOLAP
  - Caricamento: ROLAP
- Analisi: MOLAP
- Dimensione DW: ROLAP
  - MOLAP: problema sparsità
- Flessibilità nello schema: ROLAP
  - MOLAP: minor numero di dimensioni ammesse

169

Progettazione logica

- Durante questa fase, lo schema concettuale del DW viene tradotto in uno schema logico, implementabile sullo strumento scelto
- Il modello logico deve essere il più possibile vicino al modello concettuale, anche se alcune variazioni possono essere rese necessarie dal particolare tool prescelto
- supponiamo che il sistema prescelto sia ROLAP
  - tabella ➡ relazione

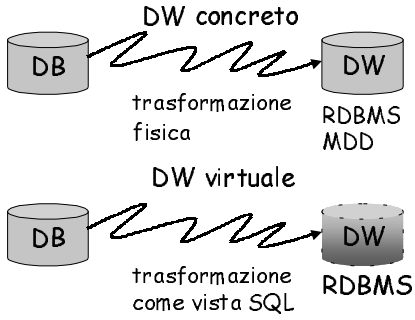
170

Progettazione logica

- Un'eccezione alle regole precedenti è dato dall'eventuale uso di view
- View per creare un DW a stella partendo da un DB normalizzato (basato su un generico schema ER)!
- Questo è utile ed efficiente solo su DW di piccole dimensioni, in cui l'accesso dimensionale è limitato

171

Progettazione logica



172

Progettazione fisica di un data warehouse

173

Problematiche

- Durante questa fase si definiscono le strutture di memorizzazione e indicizzazione da utilizzare per l'implementazione del DW
- Aspetti principali:
  - ➡ **aggregazione**
  - indici
  - parallelizzazione
  - partizionamento
  - ➡ **materializzazione query**
  - ottimizzazione query

174

Influenza aggregati sul codice SQL

- Se gli aggregati sono presenti, per poterli utilizzare bisogna ovviamente scrivere codice SQL opportuno
- partendo da una query sulle tabelle di base, le tabelle aggregate possono essere utilizzate sostituendole alle corrispondenti tabelle di base

175

Esempio query di base

Query sullo schema di base

```
SELECT categoria, SUM(vendite)
FROM vendite, prodotti, magazzini, tempo
WHERE vendite.prodotto-k = prodotti.prodotto-k
  AND vendite.magazzino-k = magazzini.magazzini-k
  AND vendite.tempo-k = tempo.tempo-k
  AND magazzini.città = 'Milano'
  AND tempo.giorno = '1 Gennaio 1996'
GROUP BY prodotti.categoria
```

176

Esempio query aggregata

- Si supponga adesso che esista una tabella aggregata per categoria

  vendite-aggreg-per-cat(categoria-k, magazzino-k,tempo-k, vendite)

177

Esempio query aggregata

Query sullo schema aggregato

```
SELECT descrizione_categoria, SUM(vendite)
FROM vendite-aggreg-per-cat, categoria, magazzini,
  tempo
WHERE
  vendite-aggreg-per-cat.categoria-k =
  categoria.categoria-k AND
  vendite-aggreg-per-cat.magazzino-k =
  magazzini.magazzini-k AND
  vendite-aggreg-per-cat.tempo-k = tempo.tempo-k
  AND magazzini.città = 'Milano' AND
  tempo.giorno = 1 Gennaio, 1996'
GROUP BY categoria. categoria-k
```

178

Influenza sul codice SQL

- Gli utenti finali e i tool di accesso devono generare codice differente in relazione che esistano o meno le tabelle aggregate
  - discontinuità delle applicazioni
- Soluzione: aggregate navigator

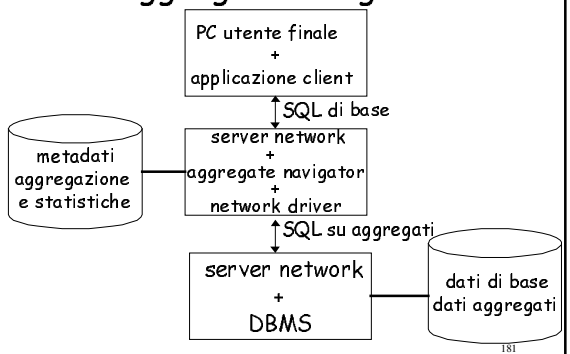
179

Aggregate navigator

- Livello software il cui obiettivo è quello di intercettare le richieste SQL e tradurle utilizzando nel modo migliore le tabelle aggregate
  - si scelgono le più piccole
- le richieste SQL si assumono utilizzare le tabelle di base
- si rende trasparente l'uso degli aggregati all'utente finale

180

Aggregate navigator



181

Strategia di navigazione

1 Si ordinano le tabelle dei fatti aggregate dalla più piccola alla più grande

2 si considera la tabella più piccola
- si verifica se tutti gli attributi dimensionali presenti nella query compaiono nelle tabelle delle dimensioni associate alla tabella dei fatti considerata
- se sì, si rimpiazzano le tabelle dei fatti e delle dimensioni di base con le tabelle aggregate
- se no, si ripete il passo 2 considerando la successiva tabella dei fatti aggregata

3 se nessuna tabella aggregata soddisfa le condizioni precendenti, la query deve essere eseguita utilizzando le tabelle di base

182

View materializzate

- Si materializza la vista, cioè la si calcola una sola volta, la si memorizza e la si usa durante l'esecuzione della query
- Politiche di caricamento:
  - immediato,all'atto della definizione della view
  - differito
- Politiche di aggiornamento (refresh):
  - lazy: la vista è aggiornata ad ogni query, se non è consistente
  - periodica
  - forzata: dopo un certo numero di cambiamenti
- Utilizzo/non utilizzo da parte dell'aggregate navigator

183

View materializzate in Oracle 8i

- Possono essere utilizzate dell'aggregate navigator
- diverse politiche di:
  - caricamento
  - aggiornamento
- il sistema è in grado di suggerire quali view materializzare, in base a statistiche di sistema

184

View materializzate in Oracle 8i

- Caricamento:
  - Immediate: all'atto della definizione (default)
  - Deferred: popolata alla successiva operazione di refresh (che deve essere completo)

185

View materializzate in Oracle 8i

- Refresh:
  - Fast: incrementale (molte restrizioni)
  - Complete: totale
  - Force: incrementale/totale (default)
  - On Commit: fast refresh al commit delle transazioni sulle tabelle di definizione della view (solo per join view e single-table view)
  - On Demand: invocando specifiche procedure
  - Start with <date> Next <date expression>
  - ....

186

View materializzate in Oracle 8i

- For update:
  - se specificato, permette di aggiornare la view e di propagare l'aggiornamento alle tabelle di base
- Query Rewrite:
  - Enable: utilizzata dall'aggregate navigator in fase di riscrittura delle query

187

View materializzate in Oracle 8i

CREATE MATERIALIZED VIEW nome
BUILD <tipo caricamento>
REFRESH <tipo refresh>
[ON UPDATE]
[ENABLE QUERY REWRITE]
AS <sottoquery di definizione>

188

View materializzate in Oracle 8i

```
CREATE MATERIALIZED VIEW
  vendite_mensili
BUILD deferred
REFRESH complete
ENABLE QUERY REWRITE
AS
SELECT mese, SUM(ricavi)
FROM Vendite v, Tempo t
WHERE v.Tempo_k = t.Tempo_k
GROUP by t.mese;
```

189

Dimensioni in Oracle 8i

- Oggetti che permettono di descrivere gerarchie esistenti all'interno delle tabelle
- vengono utilizzate per:
  - riscrivere le query
  - suggerire la creazione di view materializzate
- non contengono nuovi dati ma specificano:
  - gli attributi coinvolti nelle gerarchie (livelli)
  - le gerarchie (anche >= 1 per una stessa tabella)
  - dipendenze funzionali tra livelli ed altri attributi delle tabelle sottostanti
- possono anche coinvolgere più di una tabella (non le vediamo)

190

Dimensioni in Oracle 8i

```
CREATE DIMENSION <nome>
LEVEL <nome_l1>  IS <nome tabella>.<attr>
LEVEL <nome_l2>  IS <nome tabella>.<attr>
…
HIERARCHY <nome gerarchia> (
  <nome_livello> CHILD OF
  <nome_livello> CHILD OF
  …)
ATTRIBUTE <nome livello> DETERMINES
                    <nome<tabella>.<attr>
…
```

191

Dimensioni in Oracle 8i

```
CREATE DIMENSION Prodotti_D
LEVEL prod_l  IS Prodotti.descrizione
LEVEL sottoc_l  IS Prodotti.sottocategoria
LEVEL categ_l  IS Prodotti. categoria
HIERARCHY Prodotti_H (
  prod_l CHILD OF
  sottoc_l CHILD OF
  categ_l)
ATTRIBUTE prod_l DETERMINES marca;
```

192

Interrogazione di un data warehouse

193

Interrogazioni di base

- Le interrogazioni impongono condizioni su una o più tabelle delle dimensioni che si riflettono in restrizioni sulla tabella dei fatti
  - interrogazioni star-join
- Esempio: selezionare tutti i prodotti di colore rosso venduti nell'ultimo trimestre nell'Italia settentrionale

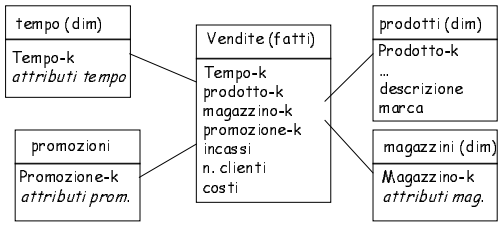
194

Schema interrogazione tipo

SELECT lista attributi e aggregati
FROM tabelle dei fatti e delle dimensioni
WHERE join tabella fatti e tab. dimensione 1
AND
join tabella fatti e tab. dimensione 2
AND
...
AND
condizioni su attributi dimensione
GROUP BY attributo dimensione
ORDER BY attributo dimensione

195

Esempio



196

Esempio

Determinare quante unità sono state vendute e quali incassi sono stati ricavati per ogni prodotto di marca "Findus"

SELECT p.descrizione, sum(f.incassi), sum(f.n-unità)
FROM vendite f, prodotti p
WHERE f.Prodotto-k = p.Prodotto-k
AND
p.marca = "Findus"
GROUP BY p. descrizione
ORDER BY p.descrizione

197

I nuovi tipi di query

- Dipendono dai tool di accesso
- influenzano l'implementazione delle query
- Operazioni di base:
  - aggregazione (noto)
  - drill-down/roll-up
  - pivoting
  - slicing
  - dicing
  - top-n

198

Roll-up e drill-down

- Drill-down:
  - aggiungere informazioni estratte da una tabella delle dimensioni ad un report
- Roll-up:
  - sottrarre informazioni contenute da una tabella delle dimensioni da un report
- Le gerarchie possono essere utilizzate per effettuare operazioni di roll-up e drill-down ma non sono necessarie

199

Esempio drill-down e roll-up

down

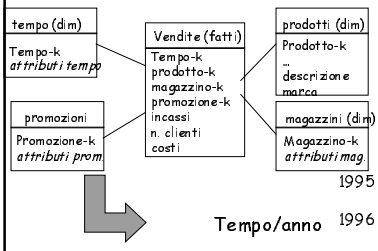
up

Dipartimento	Incassi	Unità vendute
Panificio	Lit. 12100000	5088
Cibo surgelato	Lit. 23000000	15000
…		

Dipartimento	Marca	Incassi	Unità vendute
Panificio	Barilla	6000000	2600
Panificio	Agnesi	6100000	2488
Cibo surgelato	Findus	15000000	6500
Cibo surgelato	Orogel	8000000	8500
…			

200

Pivoting



Tempo/anno

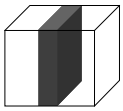
	GE	MI	Totale
1995	63	81	144
1996	38	107	145
1997	75	35	110
Totale	176	223	388

Magazzini/citta`

201

Altre operazioni

- Slicing & dicing:
  - Selezione + proiezione
  - vincoli di uguaglianza o range



- Top-n:
  - Esempio: determinare i 10 prodotti più venduti ad una certa data e in un certo magazzino, ordinati per vendite

202

Impatto sul codice SQL

- Tipiche query OLAP richiedono molte aggregazioni

```
SELECT SUM (vendite)
FROM vendite S, Tempo T, Magazzini M
WHERE S.TId = T.TId AND
    S.Mid = M.Mid
GROUP BY T.anno, M.citta`
```

	GE	MI	Totale
1995	63	81	144
1996	38	107	145
1997	75	35	110
Totale	176	223	388

```
SELECT SUM (vendite)
FROM vendite S, Magazzini M
WHERE S.MId = M.MId
GROUP BY M.citta`
```

```
SELECT SUM (vendite)
FROM vendite S, Tempo T
WHERE S.TId = T.TId
GROUP BY T.anno
```

203

Impatto sul codice SQL

- In genere:
  - fatti con k dimensioni
  - 2^k query SQL aggregate
- Nuovo operatore SQL CUBE per calcolare tutte le possibili aggregazioni

  CUBE Pid, Mid, Tid BY SUM Vendite

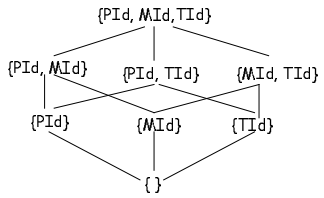
  equivalente ad un insieme di query:

  SELECT SUM (vendite)
  FROM vendite S
  GROUP BY *grouping list*
- Presente in molti DBMS

204

Impatto sul codice SQL

- Relazione tra le varie query calcolate dall'operatore CUBE



205

Impatto sul codice SQL

- Necessità di determinare "i primi n elementi" rispetto ad un certo ordinamento
- Esempio: determinare i 10 prodotti più venduti in un certo magazzino, ordinati per entità delle vendite

```
SELECT P.Pid, P.pnome, S.vendite
FROM vendite S, prodotti P
WHERE S.Pid = P.Pid AND S.Mid = 1
      AND S.Tid = 5/9/00
GROUP BY S.vendite DESC
OPTIMIZE FOR 10 ROWS
```

- Presente in molti DBMS

206

Operatori aggregati in Oracle 8i

- SQL viene esteso con nuovi operatori di aggregazione. Tra i vari operatori:
  - ROLLUP
  - CUBE
  - RANK/TOP-N

207

Roll-up

- SELECT ….
  GROUP BY ROLLUP (elenco colonne)
- calcola l'aggregato standard rispetto all'elenco di colonne specificato
- calcola subtotali di livello più alto, riducendo ad una ad una le colonne da aggregare, procedendo da destra a sinistra nella lista

208

Roll-up

- Esempio:

```
SELECT città, mese, prodotto, SUM(vendite)
FROM   Vendite v, Magazzini m, Tempo t,
       Prodotti p
WHERE m.Magazzino_k = v.Magazzino_k AND
       p.Prodotto_k = v.Prodotto_k AND
        t.Tempo_k = v.Tempo_k
GROUP BY ROLLUP(città,mese,prodotto)
```

209

Roll-up

Città	Mese	Prodotto	Vendite
genova	marzo	p1	120
genova	marzo	p2	320
genova	marzo		440
genova	luglio	p1	220
genova	luglio	p2	110
genova	luglio		330
genova			770
milano	marzo	p1	430
milano	marzo	p2	143
milano	marzo		573
milano	luglio	p1	340
milano	luglio	p2	100
milano	luglio		440
milano			1013

210

Cube

- SELECT ….
  GROUP BY CUBE (elenco colonne)
- calcola l'aggregato standard rispetto all'elenco di colonne specificato e rispetto ad ogni sottoinsieme dell'elenco specificato

211

Cube

- Esempio:

```
SELECT città, mese, prodotto, SUM(vendite)
FROM   Vendite v, Magazzini m, Tempo t,
       Prodotti p
WHERE m.Magazzino_k = v.Magazzino_k AND
       p.Prodotto_k = v.Prodotto_k AND
       t.Tempo_k = v.Tempo_k
GROUP BY CUBE(città,mese,prodotto)
```

212

Cube

Città	Mese	Prodotto	Vendite
genova	marzo	p1	120
genova	marzo	p2	320
genova	marzo		440
genova	luglio	p1	220
genova	luglio	p2	110
genova	luglio		330
genova		p1	340
genova		p2	440
genova			770
milano	marzo	p1	430
milano	marzo	p2	143
milano	marzo		573
milano	luglio	p1	340
milano	luglio	p2	100
milano	luglio		440
milano		p1	770
milano		p2	243
milano			1013
	marzo	p1	550
	marzo	p2	463
	marzo		1113
	luglio	p1	560
	luglio	p2	210
	luglio		770
		p1	1110
		p2	673
			1783

213

Top-N

- SELECT A1,…,An
  FROM
  (SELECT B1,…,Bm,
  RANK OVER(ORDER BY Ai ASC,
  ORDER BY Aj DESC) AS rank
  FROM …
  WHERE …
  GROUP BY A1,…,An)
  WHERE rank <= N;
- permette di ordinare i risultati e restituire solo i primi N rispetto all'ordinamento prescelto

214

Top-N

- Esempio:

```
SELECT città, mese, prodotto, sum_vendite
FROM
  (SELECT città,mese,prodotto, SUM(vendite),
     RANK() OVER (ORDER by SUM(vendite) DESC)
    AS rank
   FROM   Vendite v, Magazzini m, Tempo t, Prodotti p
   WHERE m.Magazzino_k = v.Magazzino_k AND
          p.Prodotto_k = v.Prodotto_k AND
          t.Tempo_k = v.Tempo_k
   GROUP BY (città,mese,prodotto))
 WHERE rank <= 3;
```

215

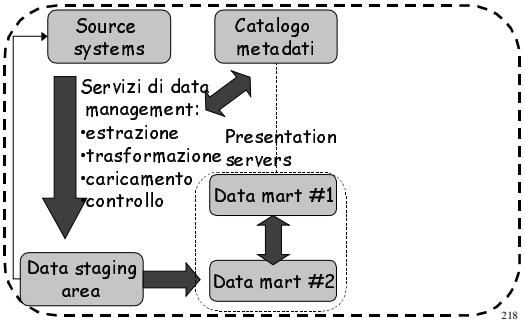
Top-N

Città	Mese	Prodotto	Vendite
milano	marzo	p1	430
milano	luglio	p1	340
genova	marzo	p2	320

216

Architettura della back room

Dove siamo?



Servizi principali

Estrazione

- Estrazione dei dati dai sistemi sorgenti
- copia di parte di essi nell'area data staging
- necessità di interagire con diverse piattaforme e formati
- Due approcci:
  - Transazioni evento:
    - vengono usati timestamp, si identificano e si aggiornano solo i record modificati
  - Rimpiazzamento completo:
    - si ricaricano completamente i dati, utile per DW di piccole dimensioni, in ogni caso, almeno 3 volte

Servizi principali

Trasformazione

- pulizia, cioè eliminazione di conflitti , inserimento di elementi mancanti, formato standard, eliminazione dati non significativi
- identificazione dimensioni che cambiano lentamente e generazione chiavi
- check di integrità referenziale
- denormalizzazione
- conversione tipi di dato e valori nulli
- generazione di dati aggregati (spesso esternamente al DBMS)
- trasformazioni dipendenti dal tool che si intende utilizzare

Servizi principali

- Caricamento:
  - Copia dei dati trasformati nei vari data marts in modalità batch
  - indicizzazione dei nuovi dati:
    - si consiglia un'indicizzazione bulk (cioè complessiva al termine del caricamento)
- Servizi di controllo:
  - Controlla l'intero processo e genera statistiche (metadati)
  - definizione dei processi, schedulazione dei processi, monitoraggio, trattamento errori, notifica

Servizi principali

- Controllo qualità:
  - Verifica consistenza dei dati caricati
  - tecniche applicabili:
    - controllo totali con sistemi di produzione
    - confronti unità periodo precedente e attuale (ad esempio, si contano i magazzini e si aggiunge una piccola variazione addittiva)
- Pubblicazione: Comunicazione dell'avvenuto upload agli utenti
- Data feedback: modifica di dati operazionali riconosciuti come errati durante l'esecuzione dei vari processi

Ruolo dei metadati nell'architettura back room

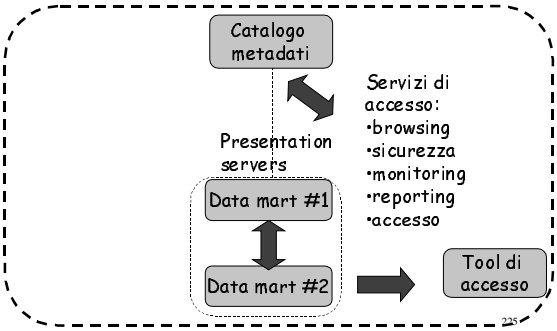
- Metadati che dipendono dai processi che vengono eseguiti:
  - estrazione, trasformazione, caricamento
- Esempi:
  - Schemi sorgente (ad esempio DDL per dati relazional)
  - frequenza di update dei dati sorgente
  - metodi di accesso
  - dimensioni e fatti conformati
  - specifiche per la pulizia dei dati e per la trasformazione
  - definizione aggregazioni
  - aspetti di sicurezza
  - indici DBMS
  - view DBMS

223

Architettura della front room

224

Dove siamo?



225

Servizi principali

- Browsing: deve permettere l'accesso ai dati, anche partendo dai metadati
  - (da una business area ad un folder)
- Sicurezza: autenticazione
- Monitoraggio:
  - performance
  - training nuovi utenti
  - generazione statistiche di uso
  - pianificazione (tempi di caricamento, tempo medio di query)

226

Servizi principali

- Query management: capacità di gestire la formulazione della query, la sua esecuzione e la presentazione del risultato
  - semplificazione vista dati all'utente
  - generazione codice SQL di base
  - aggregate navigation
  - regolamentazione query (es. tempo limite per l'esecuzione)
- Reporting: creazione di report mediante una ridotta interazione con l'utente, distribuzione agli interessati, schedulazione delle esecuzioni

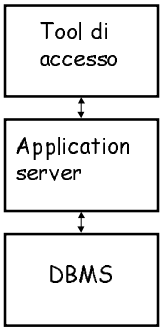
227

Servizi principali

- Supporto di tool di accesso: tool che permettono all'utente di accedere in modo intuitivo ed altamente espressivo ai dati contenuti nel DW:
  - capacità di effettuare confronti
  - presentazione dati avanzata
  - risposte alla domanda: perchè?

228

Architettura fisica



1. Soluzione tipica ma svantaggiosa (vincolati dal tool)
2. Soluzione ottimale (flessibile)
3. Limita la possibilità di usare piattaforme multiple

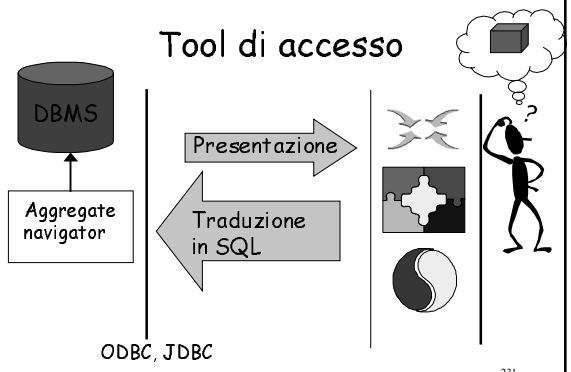
229

Tool di accesso

- Ad hoc
  - permettono all'utente di specificare le proprie query attraverso interfaccie user-friendly
- tools per la generazione di reportistica
- applicazioni avanzate
  - applicazioni che permettono di applicare operazioni molto sofisticate al DW
    - previsione
    - **DATA MINING**
    - …

230

Tool di accesso



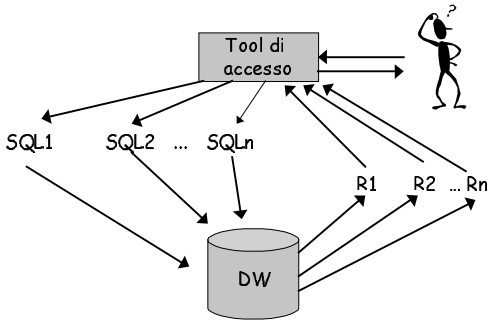
231

Gestione delle interrogazioni

- Le richieste utente richiedono in genere l'esecuzione di query SQL molto complesse
  - difficilmente ottimizzabili
  - complicano l'utilizzo dell'aggregate navigator
- Soluzione: suddividere la query utente in molte query SQL semplici, eseguite indipendentemente e composte direttamente dal tool di accesso
- Utile soprattutto se la parallelizzazione è possibile

232

Gestione delle interrogazioni



233

Operazioni tipiche di accesso

- Browsing: possibilità di visualizzare i valori associati agli attributi delle dimensioni e vincolare la ricerca ad un particolare valore
- campi calcolati: possibilità di visualizzare liste di calcoli comuni, da applicare alle dimensioni
  - vendite --> vendite %

234

Operazioni tipiche di accesso

- drilling down/up:
  - tra gli attributi: si aggiungono/tolgono attributi
  - tra report: si passa da un report ad un altro ad esso collegato
- gestione eccezioni: possibilità di individuare dati anomali, rispetto a
  - valori indicati
  - trend, previsioni
  - determinati eventi

235

Operazioni tipiche di accesso

- Interazione con aggregati: uso aggregate navigator
- drilling across: possibilità di passare da uno schema all'altro, utilizzando dimensioni e fatti comuni
- totali parziali: possibilità di aggiungere al report totali per gruppi di record
- Pivoting: risistemazione righe e colonne nei report
- Ordinamenti: possibilità di ordinare record rispetto a svariati parametri
- Import/export dei risultati in altri tool

236

Ruolo dei metadati nell'architettura front room

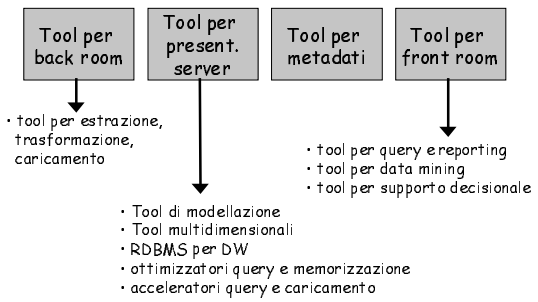
- Metadati che descrivono quali dati sono stati estratti e come devono essere presentati
- Esempi
  - nomi per colonne, tabelle, raggruppamenti
  - definizione dei report
  - documentazione
  - profili sicurezza
  - certificati di autenticazione
  - statistiche relative alla sicurezza
  - profili utente
  - statistiche relative alle query

237

Tool per Data Warehousing

238

Classi di tool



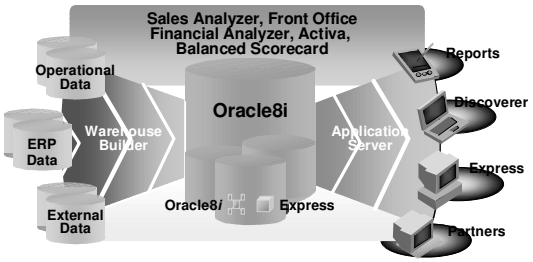
239

In genere

- I produttori di DBMS forniscono in genere tool per ogni componente architetturale
- Tra questi:
  - ORACLE
  - INFORMIX
  - MICROSOFT
  - SYBASE

240

La soluzione ORACLE



Designer and Enterprise Manager

241

DW in Oracle 8/8i

- Indici Bitmap (7.3.2, 8, 8i)
- parallelizzazione (8,8i)
- partizionamento (8, 8i)
- view materializzate (8i)
- operatori di aggregazione estesi (8i)
- ottimizzazioni per star query (7.3.8, 8, 8i)

242

Testi di riferimento

- R. Kimball. The data warehouse toolkit - Practical techniques for building dimensional data warehouses. John Wiley & Sons, Inc. 1996.
- M. Golfarelli, S. Rizzi. Data warehouse - teoria e pratica della progettazione. McGraw-Hill, 2002.

243