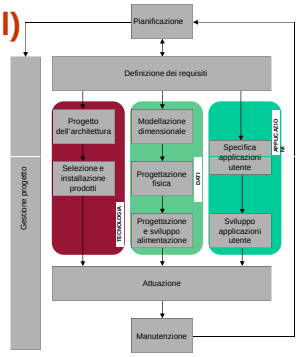
Data warehouse Progettazione

Elena Baralis
Politecnico di Torino

Business Dimensional Lifecycle (Kimball)

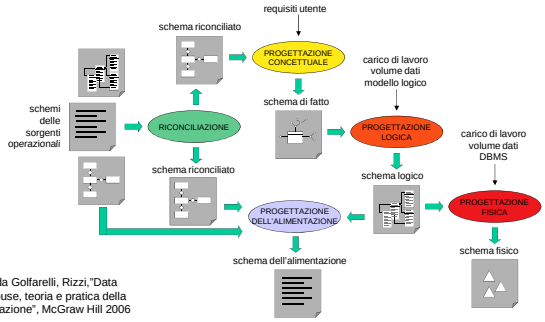


Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Fattori di rischio

- Aspettative elevate degli utenti
  - il data warehouse come soluzione dei problemi aziendali
- Qualità dei dati e dei processi OLTP di partenza
  - dati incompleti o inaffidabili
  - processi aziendali non integrati e ottimizzati
- Gestione "politica" del progetto
  - collaborazione con i "detentori" delle informazioni
  - accettazione del sistema da parte degli utenti finali

Progettazione di data mart



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Progettazione di data warehouse

- Approccio top-down
  - realizzazione di un data warehouse che fornisca una visione globale e completa dei dati aziendali
  - costo significativo e tempo di realizzazione lungo
  - analisi e progettazione complesse
- Approccio bottom-up
  - realizzazione incrementale del data warehouse, aggiungendo data mart definiti su settori aziendali specifici
  - costo e tempo di consegna contenuti
  - focalizzato separatamente su settori aziendali specifici

Analisi dei requisiti

Elena Baralis
Politecnico di Torino

Analisi dei requisiti

- Raccoglie
  - le esigenze di analisi dei dati che dovranno essere soddifatte dal data mart
  - i vincoli realizzativi dovuti ai sistemi informativi esistenti
- Fonti
  - business users
  - amministratori del sistema informativo
- Il data mart prescelto è
  - strategico per l'azienda
  - alimentato da (poche) sorgenti affidabili

Requisiti applicativi

- Descrizione degli eventi di interesse (fatti)
  - ogni fatto rappresenta una categoria di eventi di interesse per l'azienda
    - esempi: (per il CRM) reclami, servizi
  - caratterizzati da dimensioni descrittive (granularità), intervallo di storicizzazione, misure di interesse
  - informazioni raccolte in un glossario
- Descrizione del carico di lavoro
  - esame della reportistica aziendale
  - interrogazioni espresse in linguaggio naturale
    - esempio: numero di reclami per ciascun prodotto nell'ultimo mese

Requisiti strutturali

- Periodicità dell'alimentazione
- Spazio disponibile
  - per i dati
  - per le strutture accessorie (indici, viste materializzate)
- Tipo di architettura del sistema
  - numero di livelli
  - data mart dipendenti o indipendenti
- Pianificazione del deployment
  - avviamento
  - formazione

Progettazione concettuale

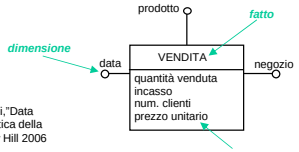
Elena Baralis
Politecnico di Torino

Progettazione concettuale

- Non esiste un formalismo di modellazione comunemente accettato
  - il modello ER non è adatto
- Dimensional Fact Model (Golfarelli, Rizzi)
  - per uno specifico fatto, definisce schemi di fatto che modellano
    - dimensioni
    - gerarchie
    - misure
  - modello grafico a supporto della progettazione concettuale
  - offre una documentazione di progetto utile sia per la revisione dei requisiti con gli utenti, sia a posteriori

Dimensional Fact Model

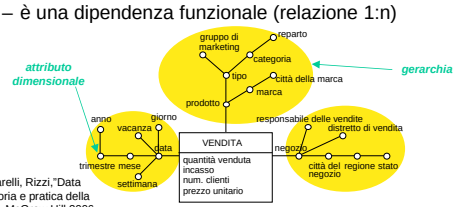
- Fatto
  - modella un insieme di eventi di interesse (vendite, spedizioni, reclami)
  - evolve nel tempo
- Dimensione
  - descrive le coordinate di analisi di un fatto (ogni vendita è descritta dalla data di effettuazione, dal negozio e dal prodotto venduto)
  - è caratterizzata da numerosi attributi, tipicamente di tipo categorico
- Misura
  - descrive una proprietà numerica di un fatto, spesso oggetto di operazioni di aggregazione (ad ogni vendita è associato un incasso)



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Dimensional Fact Model

- Gerarchia
  - rappresenta una relazione di generalizzazione tra un sottoinsieme di attributi di una dimensione (gerarchia geografica per la dimensione negozio)
  - è una dipendenza funzionale (relazione 1:n)

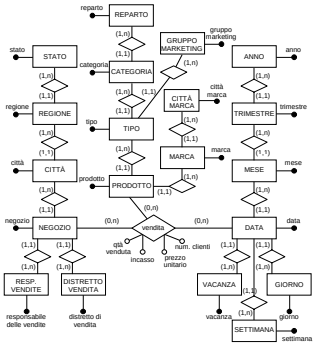


Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Aggregazione

- Processo di calcolo del valore di misure a granularità meno fine di quella presente nello schema di fatto originale
  - la riduzione del livello di dettaglio è ottenuta risalendo lungo una gerarchia
  - operatori di aggregazione standard: SUM, MIN, MAX, AVG, COUNT
- Caratteristiche delle misure
  - additive
  - non additive: non aggregabili lungo una gerarchia mediante l'operatore di somma
  - non aggregabili

Corrispondenza con l'ER

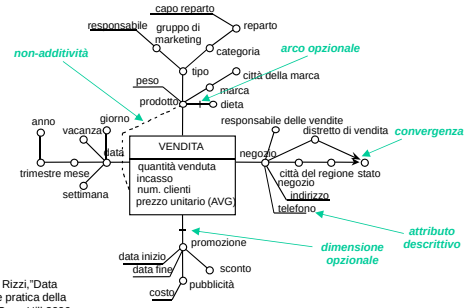


Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Classificazione delle misure

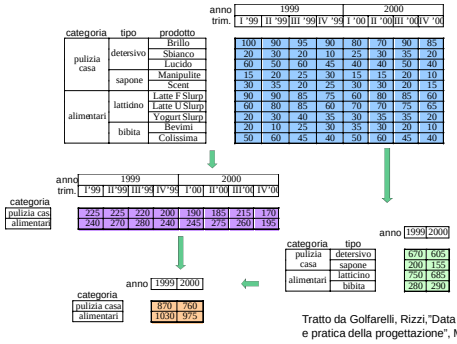
- Misure di flusso
  - possono essere valutate cumulativamente alla fine di un periodo di tempo
  - sono aggregabili mediante tutti gli operatori standard
  - esempi: quantità di prodotti venduti, importo incassato
- Misure di livello
  - sono valutate in specifici istanti di tempo (snapshot)
  - non sono additive lungo la dimensione tempo
  - esempi: livello di inventario, saldo del conto corrente
- Misure unitarie
  - sono valutate in specifici istanti di tempo ed espresse in termini relativi
  - non sono additive lungo nessuna dimensione
  - esempio: prezzo unitario di un prodotto

DFM: costrutti avanzati



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Operatori di aggregazione

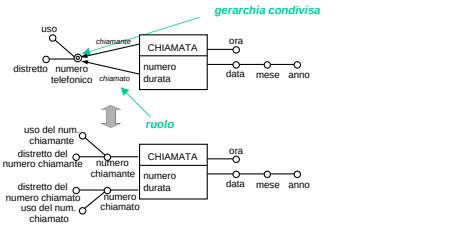


Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Operatori di aggregazione

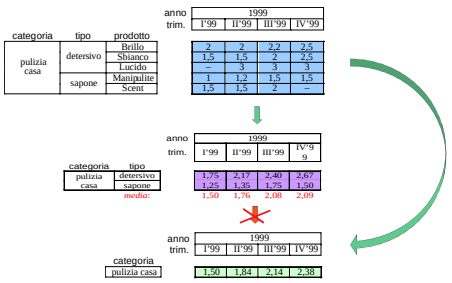
- Distributivi
  - sempre possibile il calcolo di aggregati da dati a livello di dettaglio maggiore
  - esempi: sum, min, max

DFM: costrutti avanzati



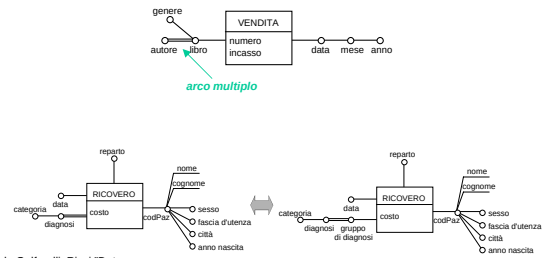
Tratti da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Operatori non distributivi



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

DFM: costrutti avanzati



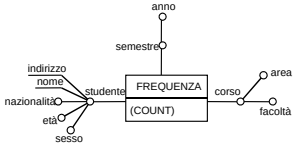
Tratti da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Operatori di aggregazione

- Distributivi
  - sempre possibile il calcolo di aggregati da dati a livello di dettaglio maggiore
  - esempi: sum, min, max
- Algebrici
  - il calcolo di aggregati da dati a livello di dettaglio maggiore è possibile in presenza di misure aggiuntive di supporto
  - esempi: avg (richiede count)
- Olistici
  - non è possibile il calcolo di aggregati da dati a livello di dettaglio maggiore
  - esempi: moda, mediana

Schemi di fatto vuoti

- L'evento può non essere caratterizzato da misure
  - schema di fatto vuoto
  - registra il verificarsi di un evento
- Utile per
  - conteggio di eventi accaduti
  - rappresentazione di eventi non accaduti (insieme di copertura)



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Rappresentazione del tempo

- La variazione dei dati nel tempo è rappresentata esplicitamente dal verificarsi degli eventi
  - presenza di una dimensione temporale
  - eventi memorizzati sotto forma di fatti
- Possono variare nel tempo anche le dimensioni
  - variazione tipicamente più lenta
    - slowly changing dimension [Kimball]
  - esempi: dati anagrafici di un cliente, descrizione di un prodotto
  - necessario prevedere esplicitamente nel modello come rappresentare questo tipo di variazione

Modalità di rappresentazione del tempo (tipo I)

- Fotografia dell'istante attuale
  - esegue la sovrascrittura del dato con il valore attuale
  - proietta nel passato la situazione attuale
  - utilizzata quando non è necessario rappresentare esplicitamente la variazione
  - Esempio
    - il cliente Mario Rossi cambia stato civile dopo il matrimonio
    - tutti i suoi acquisti sono attribuiti al cliente "sposato"

Modalità di rappresentazione del tempo (tipo II)

- Eventi attribuiti alla situazione temporalmente corrispondente della dimensione
  - per ogni variazione di stato della dimensione
    - si crea di una nuova istanza nella dimensione
    - i nuovi eventi sono correlati alla nuova istanza
  - gli eventi sono partizionati in base alle variazioni degli attributi dimensionali
  - Esempio
    - il cliente Mario Rossi cambia stato civile dopo il matrimonio
    - i suoi acquisti sono separati in acquisti attributi a Mario Rossi "celibe" e acquisti attribuiti a Mario Rossi "sposato" (nuova istanza di Mario Rossi)

Modalità di rappresentazione del tempo (tipo III)

- Eventi attribuiti alla situazione della dimensione campionata in uno specifico istante di tempo
  - proietta tutti gli eventi sulla situazione della dimensione in uno specifico istante di tempo
  - richiede una gestione esplicita delle variazioni della dimensione nel tempo
    - modifica dello schema della dimensione
      - introduzione di una coppia di timestamp che indicano l'intervallo di validità del dato (inizio e fine validità)
      - introduzione di un attributo che consenta di identificare la sequenza di variazioni di una specifica istanza (capostipite o master)
    - ogni variazione di stato della dimensione richiede la definizione di una nuova istanza

Modalità di rappresentazione del tempo (tipo III)

- Esempio
  - il cliente Mario Rossi cambia stato civile dopo il matrimonio
  - la prima istanza conclude il suo periodo di validità il giorno del matrimonio
  - la nuova istanza inizia la sua validità nello stesso giorno
  - gli acquisti sono separati come nel caso precedente
  - esiste un attributo che permette di ricostruire tutte le variazioni ascrivibili a Mario Rossi

Carico di lavoro

- Carico di riferimento definito da
  - reportistica standard
  - stime discusse con gli utenti
- Carico reale difficile da stimare correttamente durante la fase di progettazione
  - se il sistema ha successo, il numero di utenti e interrogazioni aumenta nel tempo
  - la tipologia di interrogazioni può variare nel tempo
- Fase di tuning
  - dopo l'avviamento del sistema
  - monitoraggio del carico di lavoro reale del sistema

Volume dei dati

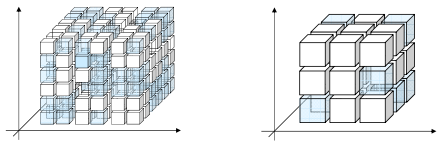
- Stima dello spazio necessario per il data mart
  - per i dati
  - per le strutture accessorie (indici, viste materializzate)
- Si considerano
  - numero di eventi di ogni fatto
  - numero di valori distinti degli attributi nelle gerarchie
  - lunghezza degli attributi
- Dipende dall'intervallo temporale di memorizzazione dei dati
- Valutazione affetta dal problema della sparsità
  - il numero degli eventi accaduti non corrisponde a tutte le possibili combinazioni delle dimensioni
  - esempio: percentuale dei prodotti effettivamente venduti in ogni negozio in un dato giorno pari circa al 10% di tutte le possibili combinazioni

Progettazione logica

- Si considera il modello relazionale (ROLAP)
  - inputs
    - schema (di fatto) concettuale
    - carico di lavoro
    - volume dei dati
    - vincoli di sistema
  - output
    - schema logico relazionale
- Basata su principi diversi rispetto alla progettazione logica tradizionale
  - ridondanza dei dati
  - denormalizzazione delle tabelle

Sparsità

- Si riduce al crescere del livello di aggregazione dei dati
- Può ridurre l'affidabilità della stima della cardinalità dei dati aggregati



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

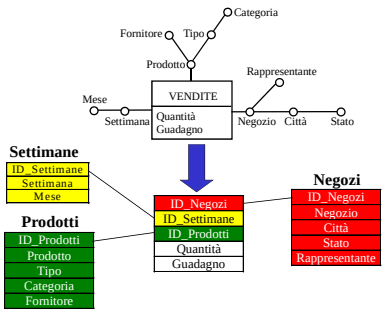
Schema a stella

- Dimensioni
  - una tabella per ogni dimensione
  - chiave primaria generata artificialmente (surrogata)
  - contiene tutti gli attributi della dimensione
  - gerarchie non rappresentate esplicitamente
    - gli attributi della tabella sono tutti allo stesso livello
  - rappresentazione completamente denormalizzata
    - presenza di ridondanza nei dati
- Fatti
  - una tabella dei fatti per ogni schema di fatto
  - chiave primaria costituita dalla combinazione delle chiavi esterne delle dimensioni
  - le misure sono attributi della tabella

Progettazione logica

Elena Baralis
Politecnico di Torino

Schema a stella



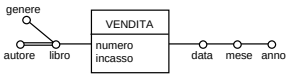
Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Snowflake schema

- Separazione di (alcune) dipendenze funzionali frazionando i dati di una dimensione in più tabelle
  - si introduce una nuova tabella che separa in due rami una gerarchia dimensionale (taglio su un attributo della gerarchia)
  - una nuova chiave esterna esprime il legame tra la dimensione e la nuova tabella
- Si riduce lo spazio necessario per la memorizzazione della dimensione
  - riduzione non significativa
- Aumenta il costo di ricostruzione dell'informazione della dimensione
  - è necessario il calcolo di uno o più join

Elena Baralis
Politecnico di Torino

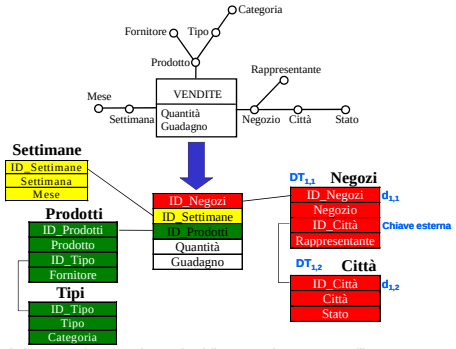
Archi multipli



- Soluzioni realizzative
  - bridge table
    - tabella aggiuntiva che modella la relazione molti a molti
    - nuovo attributo che consenta di pesare la partecipazione delle tuple nella relazione
  - push down
    - arco multiplo integrato nella tabella dei fatti
    - nuova dimensione corrispondente nella tabella dei fatti

Elena Baralis
Politecnico di Torino

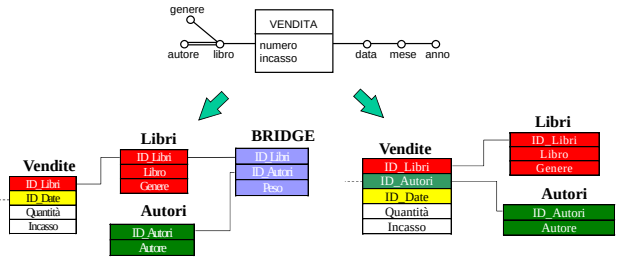
Snowflake schema



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Archi multipli



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Star o snowflake?

- Lo schema snowflake è normalmente sconsigliato
  - la riduzione di spazio occupato è scarsamente benefica
    - l'occupazione maggiore di spazio è dovuta alla tabella dei fatti (la differenza è pari ad alcuni ordini di grandezza)
  - il costo di eseguire più join può essere significativo
- Lo schema snowflake può essere utile
  - quando porzioni di una gerarchia sono condivise tra più dimensioni (esempio: gerarchia geografica)
  - in presenza di viste materializzate che richiedano una rappresentazione "aggregata" anche della dimensione

Elena Baralis
Politecnico di Torino

Archi multipli

- Tipologie di interrogazione
  - pesate: considerano il peso dell'arco multiplo
    - esempio: incasso di ciascun autore
    - con bridge table

```
SELECT ID_Autori, SUM(Incasso*Peso)
...
group by ID_Autori
```

  - di impatto: non considerano il peso
    - esempio: numero di copie vendute per ogni autore
    - con bridge table

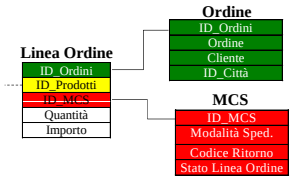
```
SELECT ID_Autori, SUM(Quantità)
...
group by ID_Autori
```


Elena Baralis
Politecnico di Torino

Archi multipli

- Confronto tra le soluzioni realizzative
  - il peso è esplicitato nella bridge table, ma integrato nella tabella dei fatti per push down
    - (push down) difficile eseguire interrogazioni di impatto
    - (push down) calcolo del peso durante l'alimentazione
    - (push down) modifiche successive difficoltose
  - push down introduce una forte ridondanza nella tabella dei fatti
  - costo di esecuzione delle interrogazioni minore per push down
    - numero minore di join

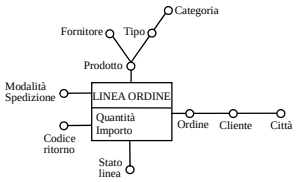
Junk dimension



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Dimensioni degeneri

- Dimensioni rappresentate da un solo attributo



Viste materializzate

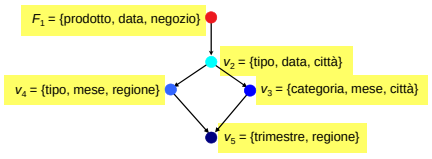
Elena Baralis
Politecnico di Torino

Dimensioni degeneri

- Soluzioni realizzative
  - integrazione nella tabella dei fatti
    - per attributi di dimensione (molto) contenuta
  - junk dimension
    - unica dimensione che integra più dimensioni degeneri
    - non esistono dipendenze funzionali tra gli attributi della dimensione
      - sono possibili tutte le combinazioni
      - attuabile solo per cardinalità limitate del dominio degli attributi

Viste materializzate

- Sommari precalcolati della tabella dei fatti
  - memorizzati esplicitamente nel data warehouse
  - permettono di aumentare l'efficienza delle interrogazioni che richiedono aggregazioni



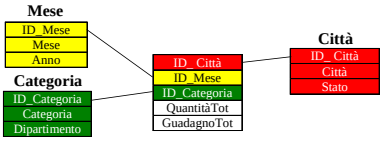
Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Viste materializzate

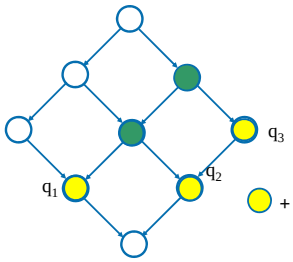
- Definite da istruzioni SQL
- Esempio: definizione di v_3
  - a partire da tabelle di base o viste di granularità superiore

```
group by Città, Mese, Categoria
```

  - aggregazione (SUM) sulle misure `Quantità`, `Guadagno`
  - riduzione dettaglio delle dimensioni



Scelta delle viste

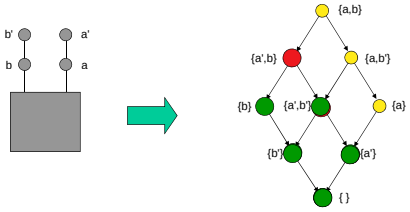


+ = ***viste candidate***, **ossia potenzialmente utili a ridurre il costo di esecuzione del carico di lavoro**

Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Viste materializzate

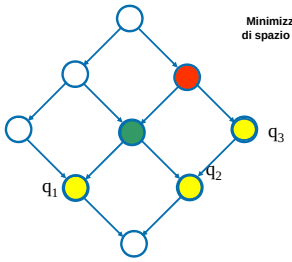
- Una vista materializzata può essere utilizzata per rispondere a più interrogazioni diverse
  - attenzione al tipo di operatore di aggregazione richiesto



Reticolo multidimensionale

Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Scelta delle viste

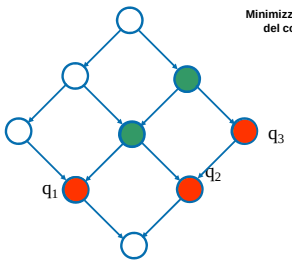


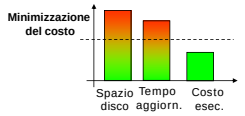
Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Scelta delle viste

- Numero di possibili combinazioni di aggregazioni molto elevato
  - quasi tutte le combinazioni di attributi sono eleggibili
- Scelta dell'insieme "ottimo" di viste materializzate
- Minimizzazione di funzioni di costo
  - esecuzione delle interrogazioni
  - aggiornamento delle viste materializzate
- Vincoli
  - spazio disponibile
  - tempo a disposizione per l'aggiornamento
  - tempo di risposta
  - freschezza dei dati

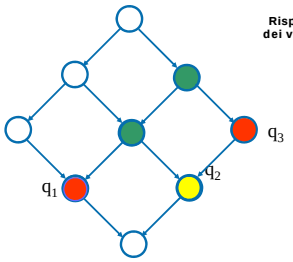
Scelta delle viste





Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Scelta delle viste



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Progettazione fisica

- Caratteristiche dell'ottimizzatore
  - deve considerare le statistiche nella definizione del piano di accesso ai dati (cost based)
  - funzionalità di aggregate navigation
- Procedimento di progettazione fisica
  - selezione delle strutture adatte per supportare le interrogazioni più frequenti (o più rilevanti)
  - scelta di strutture in grado di contribuire al miglioramento di più interrogazioni contemporaneamente
  - vincoli
    - spazio su disco
    - tempo disponibile per l'aggiornamento dei dati

Progettazione fisica

Elena Baralis
Politecnico di Torino

Progettazione fisica

- Tuning
  - variazione a posteriori delle strutture fisiche di supporto
  - richiede strumenti di monitoraggio del carico di lavoro
  - spesso necessario per applicazioni OLAP
- Parallelismo
  - frammentazione dei dati
  - parallelizzazione delle interrogazioni
    - inter-query
    - intra-query
  - le operazioni di join e group by si prestano bene all'esecuzione parallela

Progettazione fisica

- Caratteristiche del carico di lavoro
  - interrogazioni con aggregati che richiedono l'accesso a una frazione significativa di ogni tabella
  - accesso in sola lettura
  - aggiornamento periodico dei dati con eventuale ricostruzione delle strutture fisiche di accesso (indici, viste)
- Strutture fisiche
  - tipologie di indici diverse da quelle tradizionali
    - indici bitmap, indici di join, bitmapped join index, ...
    - l'indice B^+-tree non è adatto per
      - attributi con dominio a cardinalità bassa
      - interrogazioni poco selettive
  - viste materializzate
    - richiedono la presenza di un ottimizzatore che le sappia sfruttare

Scelta degli indici

- Indicizzazione delle dimensioni
  - attributi frequentemente coinvolti in predicati di selezione
  - se il dominio ha cardinalità elevata, indice B-tree
  - se il dominio ha cardinalità ridotta, indice bitmap
- Indici per i join
  - raramente opportuno indicizzare solo le chiavi esterne della tabella dei fatti
  - consigliato bitmapped join index, se disponibile
- Indici per i group by
  - uso di viste materializzate

Alimentazione del data warehouse

Elena Baralis
Politecnico di Torino

Extraction, Transformation and Loading (ETL)

- Processo di preparazione dei dati da introdurre nel data warehouse
  - estrazione dei dati dalle sorgenti
  - pulitura
  - trasformazione
  - caricamento
- semplificato dalla presenza di una staging area
- eseguito durante
  - il primo popolamento del DW
  - l'aggiornamento periodico dei dati

Estrazione

- Acquisizione dei dati dalle sorgenti
- Modalità di estrazione
  - statica: fotografia dei dati operazionali
    - eseguita durante il primo popolamento del DW
  - incrementale: selezione degli aggiornamenti avvenuti dopo l'ultima estrazione
    - utilizzata per l'aggiornamento periodico del DW
    - immediata o ritardata
- Scelta dei dati da estrarre basata sulla loro qualità

Estrazione

- Dipende dalla natura dei dati operazionali
  - storicizzati: tutte le modifiche sono memorizzate per un periodo definito di tempo nel sistema OLTP
    - transazioni bancarie, dati assicurativi
    - operativamente semplice
  - semi-storicizzati: è conservato nel sistema OLTP solo un numero limitato di stati
    - operativamente complessa
  - transitori: il sistema OLTP mantiene solo l'immagine corrente dei dati
    - scorte di magazzino, dati di inventario
    - operativamente complessa

Estrazione incrementale

- Assistita dall'applicazione
  - le modifiche sono catturate da specifiche funzioni applicative
  - richiede la modifica delle applicazioni OLTP (o delle API di accesso alla base di dati)
  - aumenta il carico applicativo
  - necessaria per sistemi legacy
- Uso del log
  - accesso mediante primitive opportune ai dati del log
  - formato proprietario del log
  - efficiente, non interferisce con il carico applicativo

Estrazione incrementale

- Definizione di trigger
  - i trigger catturano le modifiche di interesse
  - non richiede la modifica dei programmi applicativi
  - aumenta il carico applicativo
- Basata su timestamp
  - i record operazionali modificati sono marcati con il timestamp dell'ultima modifica
  - richiede la modifica dello schema della base di dati OLTP (e delle applicazioni)
  - estrazione differita, può perdere stati intermedi se i dati sono transitori

Confronto tra le tecniche di estrazione

	Statica	*Marche temporali*	*Assistita applicazione*	*Trigger*	*Log*
Gestione dati transitori o semi-storicizzati	NO	Incompleta	Completa	Completa	Completa
Supporto per sistemi basati su file	SI	SI	SI	NO	Raro
Tecnica di realizzazione	Prodotti	Prodotti o sviluppo interno	Sviluppo interno	Prodotti	Prodotti
Costi di sviluppo interno	Nessuno	Medi	Alti	Nessuno	Nessuno
Utilizzo in sistemi legacy	SI	Difficile	Difficile	Difficile	SI
Modifiche ad applicazioni	Nessuna	Probabile	Probabile	Nessuna	Nessuna
Dipendenza delle proedure dal DBMS	Limitata	Limitata	Variabile	Alta	Limitata
Impatto sulle prestazioni del sistema operaz.	Nessuna	Nessuna	Medio	Medio	Nessuna
Complessità delle procedure di estrazione	Bassa	Bassa	Alta	Media	Bassa

Tratto da Devlin, Data warehouse: from architecture to implementation, Addisono-Wesley, 1997

Pulitura

- Ogni problema richiede una tecnica specifica di soluzione
  - tecniche basate su dizionari
    - adatte per errori di battitura o formato
    - utilizzabili per attributi con dominio ristretto
  - tecniche di fusione approssimata
    - adatte per riconoscimento di duplicati/correlazioni tra dati simili
      - join approssimato
      - problema purge/merge
  - identificazione di outliers o deviazioni da business rules
- La strategia migliore è la prevenzione, rendendo più affidabili e rigorose le procedure di data entry OLTP

Estrazione incrementale

4/4/2010

Cod	Prodotto	Cliente	Qtà
1	Greco di tufo	Malavasi	50
2	Barolo	Maio	150
3	Barbera	Lumini	75
4	Sangiovese	Cappelli	45

6/4/2010

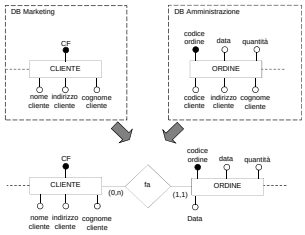
Cod	Prodotto	Cliente	Qtà
1	Greco di tufo	Malavasi	50
2	Barolo	Maio	150
4	Sangiovese	Cappelli	145
5	Vermentino	Maltoni	25
6	Trebbiano	Maltoni	150

Differenza incrementale

Cod	Prodotto	Cliente	Qtà	Azione
3	Barbera	Lumini	75	D
4	Sangiovese	Cappelli	145	U
5	Vermentino	Maltoni	25	I
6	Trebbiano	Maltoni	150	I

Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Join approssimato



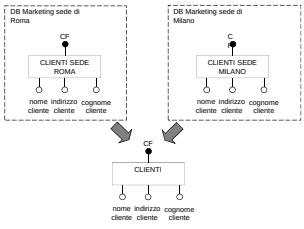
Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

- Il join deve essere eseguito sulla base dei campi comuni, che non rappresentano un identificatore per il cliente

Pulitura

- Operazioni volte al miglioramento della qualità dei dati (correttezza e consistenza)
  - dati duplicati
  - dati mancanti
  - uso non previsto di un campo
  - valori impossibili o errati
  - inconsistenza tra valori logicamente associati
- Problemi dovuti a
  - errori di battitura
  - differenze di formato dei campi
  - evoluzione del modo di operare dell'azienda

Problema purge/merge



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

- I record duplicati devono essere identificati ed eliminati
- E` necessario un criterio per valutare la somiglianza tra due record

Esempio di pulitura e trasformazione

Elena Baralis
C.so Duca degli Abruzzi 24
20129 Torino (I)

Normalizzazione

nome:	Elena
cognome:	Baralis
indirizzo:	C.so Duca degli Abruzzi 24
CAP:	20129
città:	Torino
nazione:	I

Standardizzazione

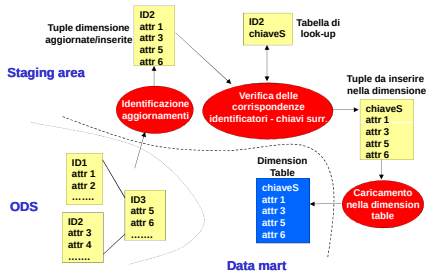
nome:	Elena
cognome:	Baralis
indirizzo:	Corso Duca degli Abruzzi 24
CAP:	20129
città:	Torino
nazione:	Italia

Correzione

nome:	Elena
cognome:	Baralis
indirizzo:	Corso Duca degli Abruzzi 24
CAP:	10129
città:	Torino
nazione:	Italia

Adattato da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Alimentazione delle dimensioni

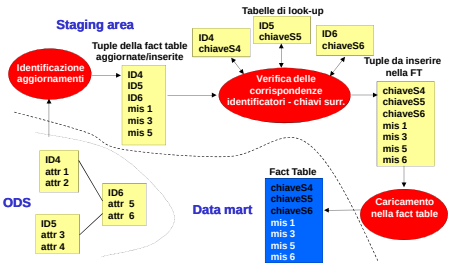


Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Trasformazione

- Conversione dei dati dal formato operazionale a quello del data warehouse (integrazione)
- Richiede una rappresentazione uniforme dei dati operazionali (schema riconciliato)
- Può avvenire in due passi
  - dalle sorgenti operazionali ai dati riconciliati nella staging area
    - conversioni e normalizzazioni
    - matching
    - (eventuale) filtraggio dei dati significativi
  - dai dati riconciliati al data warehouse
    - generazione di chiavi surrogate
    - generazione di valori aggregati

Alimentazione delle fact table

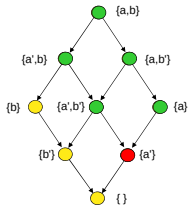


Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Caricamento

- Propagazione degli aggiornamenti al data warehouse
- Per mantenere l'integrità dei dati, si aggiornano in ordine
  1. dimensioni
  2. tabelle dei fatti
  3. viste materializzate e indici
- Finestra temporale limitata per eseguire gli aggiornamenti
- Richiede proprietà transazionali (affidabilità, atomicità)

Alimentazione delle viste materializzate



Tratto da Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006