

- Illustrare brevemente le tre principali soluzioni per la progettazione logica delle gerarchie temporali (slowly-changing dimension), con particolare riferimento agli scenari temporali che ciascuna di esse supporta.

Il modello multidimensionale assume che gli eventi che istanziano un fatto siano **dinamici**, e che i valori degli attributi che popolano le gerarchie siano **statici**.

- Questa visione non è realistica poiché anche i valori presenti nelle gerarchie variano nel tempo dando vita alle gerarchie dinamiche (*slowly-changing dimension*)
- L'adozione di gerarchie dinamiche implica un sovraccostio in termini di spazio e può comportare una forte riduzione delle prestazioni Oggi per ieri (*attualizzazione*)
- ✓ I dati vengono interpretati in base all'attuale configurazione della gerarchia
- ✓ Implementabile sullo schema a stella
- Ieri per oggi (*retrodatazione*)
- ✓ I dati vengono interpretati in base alla configurazione della gerarchia valida in un particolare istante
- ✓ Richiede la storicizzazione dei dati
- Oggi o ieri (*verità storica*)
- ✓ I dati vengono interpretati in base alla configurazione valida al momento in cui sono stati registrati
- ✓ Implementabile sullo schema a stella

- Descrivere sinteticamente le tre architetture “data mart indipendenti”, “data mart bus” e “hub-and-spoke”, indicando pregi e difetti di ciascuna di esse.

- Spiegare in che modo le architetture “data mart bus” e “hub-and-spoke” permettono di ricostruire una visione “enterprise” integrata dei dati.

- Cosa si intende per architettura federata in ambito data warehouse e in quali situazioni può risultare utile? - Cosa si intende per architettura federata in ambito data warehouse e in quali situazioni può risultare utile?

Data mart indipendenti Data mart bus

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

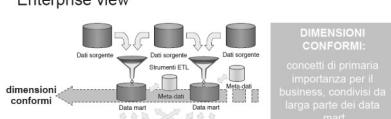


Hub-and-spoke

- Una delle architetture più usate in contesti medio-grandi



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



Federazione

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



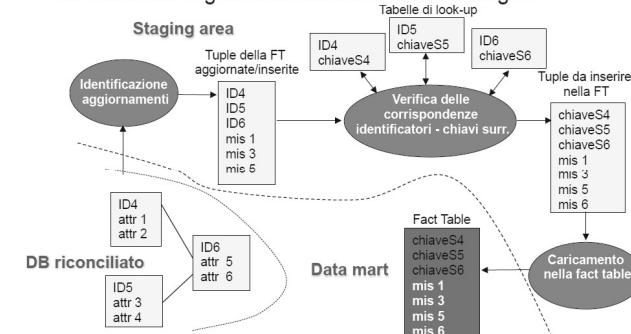
Spiegare, eventualmente con l'ausilio di un diagramma, come si effettua il caricamento delle dimension table e delle fact table.

- Illustrare a grandi linee la procedura per il caricamento delle dimension table e delle fact table, spiegando in particolare il ruolo delle look-up table.

- Si spieghi il ruolo della tabella di look-up nell'ETL, con particolare riferimento all'alimentazione di una dimension table gestita come slowly-changing dimension di tipo 1.

Alimentazione delle fact table

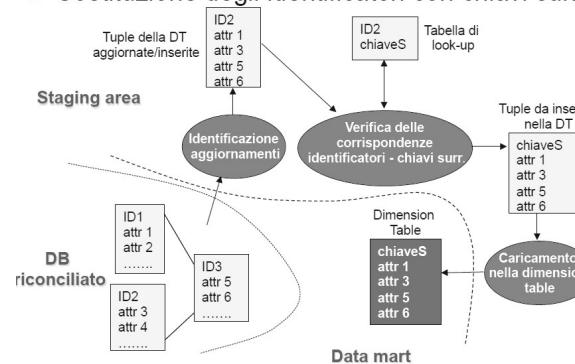
- Segue l' alimentazione delle dimension table per poter rispettare i vincoli di integrità referenziale
- Identificazione dei dati da caricare
- Sostituzione degli identificatori con chiavi surrogate



102

Alimentazione delle dimension table

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



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

- Illustrare la struttura e il funzionamento di un bitmap index, evidenziandone in particolare i pro e i contro rispetto a un B+-tree.

I vantaggi degli indici bitmap

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

Esempio: "Quanti maschi in Emilia-Romagna sono assicurati ?"

RID	Sesso	Assic.	Regione	
1	M	No	LO	
2	M	Sì	E/R	
3	F	No	LA	
4	M	Sì	E/R	

→

1	0	0
1	1	1
0	0	0
1	1	1

= 2

Occupazione su disco

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

Esempio:

NR=10.000.000

Len(Pointer) = 4x8 bit

B-tree

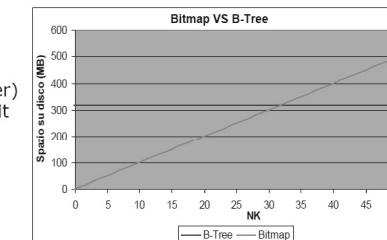
NRxLen(Pointer)

Bitmap

NR × NK × 1 bit

Si ha un risparmio di spazio se:

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



La compressione delle matrici riduce il fattore di crescita della dimensione

- Illustrare il paradigma OLAP per l'analisi dei dati multidimensionali

OLAP

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



interfaccia flessibile, facile
da usare ed efficace

OLAP: sessione

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

- Descrivere la tecnica di materializzazione delle viste spiegandone pro e contro, e discutere in particolare il ruolo del reticolo multidimensionale.

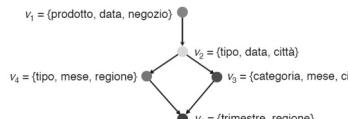
Le viste

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

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

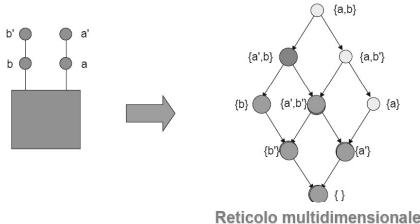
Le viste

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



Risolvibilità delle interrogazioni

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



- Spiegare l'utilizzo del costrutto di gerarchia ricorsiva nel DFM, descrivendone le possibili implementazioni su piattaforma ROLAP.

Gerarchie ricorsive

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



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

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



Navigazione		ID Padre	ID Figlio	Livello	Foglia
1	1	1	1	0	FALSE
1	2	1	2	1	FALSE
1	3	1	3	1	TRUE
1	4	1	4	1	FALSE
1	5	1	5	2	TRUE
2	2	2	2	0	FALSE
2	5	2	5	1	TRUE
2	6	2	6	1	TRUE
5	5	5	5	0	TRUE

- Illustrare la struttura di uno star index, spiegandone pro e contro

- Illustrare i criteri qualitativi che il progettista può seguire per la scelta degli indici in uno star schema.

Gli indici a stella

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

Settimana

IDSett	Sett	Mese	
13	Jan1	Jan.	1
24	Jan2	Jan.	2

Negozi

IDNegozio	Negozio	Città	Stato
123	N1	RM	I
367	N3	MI	I

Vendite

IDNeg	IDSett	IDProd	Quantità	Guadagno
123	13	41	100	10,00
123	24	17	150	15,00
367	13	17	350	35,00
367	24	41	120	12,00

Prodotti

IDProd	Prodotto	Tipo	Categoria	Fornitore
41	P1	A	X	F1
17	P2	A	X	F1

Gli indici a stella

- Rappresentano esplicitamente l' aspetto multidimensionale dei dati e dipendono fortemente dall' ordinamento delle colonne.



- Sono molto efficienti quando utilizzati in interrogazioni che coinvolgono tutte o le sole colonne iniziali dell' indice.
- Forniscono prestazioni sub-ottime in caso contrario.

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

Definire i concetti di data warehouse, data mart e operational data store.

Il Data Warehouse

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

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

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

DATA MART:

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

OPERATIONAL DATA STORE o DATABASE RICONCILIATO:

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

- Illustrare il ruolo delle procedure di ETL nel processo di data warehousing e dettagliarne le singole fasi.

- Elencare e descrivere brevemente le tecniche per l'estrazione incrementale dei dati dalle sorgenti.

- Si illustri, anche attraverso un esempio, la differenza tra pulitura e trasformazione dei dati.

■ Il ruolo degli strumenti di *Extraction, Transformation and Loading* è quello di alimentare una sorgente dati singola, dettagliata, esaurente e di alta qualità che possa a sua volta alimentare il DW (*riconciliazione*)

■ Durante il processo di alimentazione del DW, la riconciliazione avviene in due occasioni: quando il DW viene popolato per la prima volta, e periodicamente quando il DW viene aggiornato.

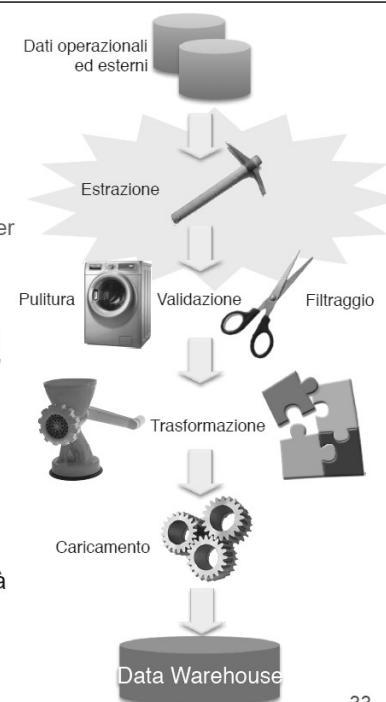
- ✓ estrazione
- ✓ pulitura
- ✓ trasformazione
- ✓ caricamento

Estrazione

■ I dati rilevanti vengono estratti dalle sorgenti

- ✓ L'estrazione statica viene effettuata quando il DW deve essere popolato per la prima volta e consiste concettualmente in una fotografia dei dati operazionali
- ✓ L'estrazione incrementale viene usata per l'aggiornamento periodico del DW, e cattura solamente i cambiamenti avvenuti nelle sorgenti dall'ultima estrazione
 - basata sul log mantenuto dal DBMS operazionale
 - basata su time-stamp
 - guidata dalle sorgenti

■ La scelta dei dati da estrarre avviene principalmente in base alla loro qualità



33

Trasformazione

■ Converte i dati dal formato operazionale sorgente a quello del DW. La corrispondenza con il livello sorgente è complicata dalla presenza di fonti distinte eterogenee, che richiede una complessa fase di integrazione

- ✓ presenza di testi liberi che nascondono informazioni importanti
- ✓ utilizzo di formati e convenzioni differenti per lo stesso dato

■ Per l'alimentazione dei dati riconciliati:

- ✓ conversione e normalizzazione (operano a livello di formato di memorizzazione e di unità di misura per uniformare i dati)
- ✓ matching (stabilisce corrispondenze tra campi equivalenti in sorgenti diverse)
- ✓ selezione (riduce il numero di campi e di record rispetto alle sorgenti)

■ Per l'alimentazione del DW:

- ✓ la normalizzazione è sostituita dalla denormalizzazione
- ✓ si introduce l'aggregazione, che realizza le opportune sintesi dei dati



35

Caricamento

■ Il caricamento dei dati nel DW

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

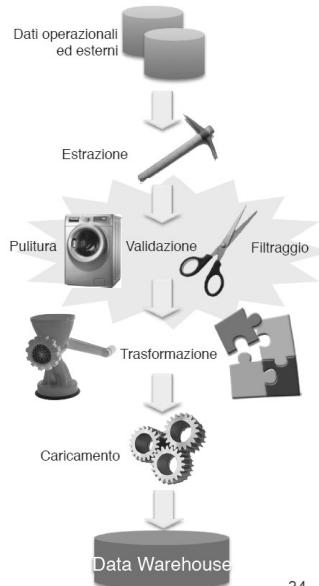


37

Pulitura

- Si incarica di migliorare la qualità dei dati delle sorgenti

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



34

- Illustrare le differenze tra i due principali approcci alla progettazione di data mart, ossia **datadriven** (guidato dai dati) e **requirement-driven** (guidato dai requisiti), discutendo brevemente per ciascuno i vantaggi, gli svantaggi e l'applicabilità.

- Quali sono i vantaggi metodologici che si conseguono prevedendo una fase di **progettazione concettuale**?

Approcci guidati dai dati (*supply-driven*)

- ✓ progettano il data mart a partire da una dettagliata analisi delle sorgenti operazionali
- ✓ i requisiti utente impattano sul progetto guidando il progettista nella selezione delle porzioni di dati considerate rilevanti per il processo decisionale, e determinando la loro strutturazione secondo il modello multidimensionale

Approcci guidati dai requisiti (*demand-driven*)

- ✓ iniziano determinando i requisiti informativi degli utenti del data mart
- ✓ il problema di come creare una mappatura tra questi requisiti e le sorgenti dati disponibili viene affrontato solo in seguito, attraverso l'implementazione di procedure ETL adatte

Approccio guidato dai dati

Vantaggi

- ✓ uno schema concettuale di massima per il data mart può essere derivato algoritmicamente a partire dal livello dei dati riconciliati, ossia in funzione della struttura delle sorgenti
- ✓ la progettazione dell'ETL risulta notevolmente semplificata, poiché ciascuna informazione nel data mart è direttamente associata a uno o più attributi delle sorgenti

Svantaggi

- ✓ ai requisiti utente viene assegnato un ruolo secondario nel determinare i contenuti informativi per l'analisi
- ✓ al progettista viene dato un supporto limitato per l'identificazione di fatti, dimensioni e misure

Applicabilità

- ✓ E' applicabile quando:
 1. è disponibile preliminarmente, oppure ottenibile con costi e tempi contenuti, una conoscenza approfondita delle sorgenti da cui il data mart si alimenterà;
 2. gli schemi delle sorgenti mostrano un buon grado di normalizzazione;
 3. la complessità degli schemi delle sorgenti non è eccessiva
- ✓ Quando l'architettura prescelta prevede l'adozione di un livello riconciliato questi requisiti sono soddisfatti: la normalizzazione e la conoscenza approfondita sono garantite dalla riconciliazione. Lo stesso vale nel caso in cui la sorgente si riduca a un singolo database, ben progettato e di dimensioni limitate
- ✓ L'esperienza di progettazione mostra che, qualora applicabile, l'approccio guidato dai dati risulta preferibile agli altri poiché permette di raggiungere i risultati prefissati in tempi estremamente contenuti

23

Approccio guidato dai requisiti

- Vantaggi
 - ✓ i desiderata degli utenti vengono portati in primo piano
- Svantaggi
 - ✓ è richiesto al progettista uno sforzo consistente durante il disegno dell' alimentazione
 - ✓ fatti, misure e gerarchie vengono desunte direttamente dalle specifiche dettate dagli utenti, e solo a posteriori si verifica che le informazioni richieste siano effettivamente disponibili nei database operazionali
 - ✓ la fiducia del cliente verso il progettista e verso l' utilità del data mart può venir meno
- Applicabilità
 - ✓ Questo approccio costituisce l' unica alternativa nei casi in cui non sia fattibile a priori un' analisi approfondita delle sorgenti (per esempio quando il data mart viene alimentato da un sistema ERP), oppure qualora le sorgenti siano rappresentate da sistemi legacy di tale complessità da sconsigliarne la ricognizione e la normalizzazione
 - ✓ E' più difficilmente perseguitibile dell' approccio guidato dai dati

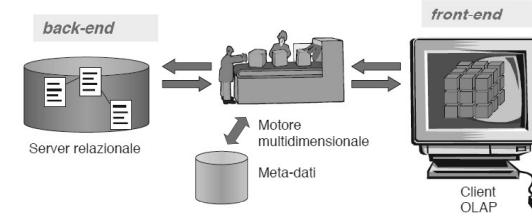
- Spiegare le principali differenze tra le piattaforme **ROLAP**, **MOLAP** e **HOLAP**.

MOLAP (Multidimensional OLAP)

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

ROLAP (Relational OLAP)

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

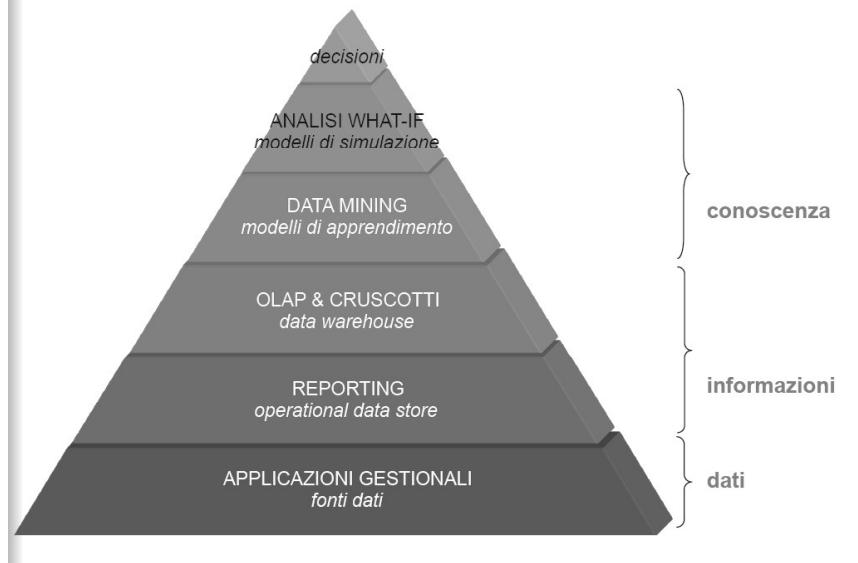


HOLAP (Hybrid OLAP)

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

- Definire la *business intelligence* e descrivere brevemente i livelli della piramide.

La piramide della BI



- Spiegare le principali differenze tra i carichi di lavoro di tipo **OLAP** e **OLTP**.

■ OLTP (On-Line Transactional Processing):

- Le interrogazioni eseguono transazioni che leggono e scrivono un ridotto numero di record da diverse tabelle legate da semplici relazioni
- Il nucleo sostanziale del carico di lavoro è “congelato” all’interno dei programmi applicativi

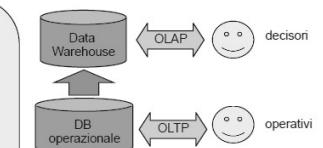
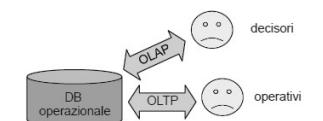
■ OLAP (On-Line Analytical Processing):

- Le interrogazioni effettuano un’analisi multidimensionale che richiede la scansione di un’enorme quantità di record per calcolare un insieme di dati numerici di sintesi che quantifichino le prestazioni dell’azienda
- L’interattività è una caratteristica irrinunciabile delle sessioni di analisi e fa sì che il carico di lavoro effettivo vari continuamente nel tempo

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



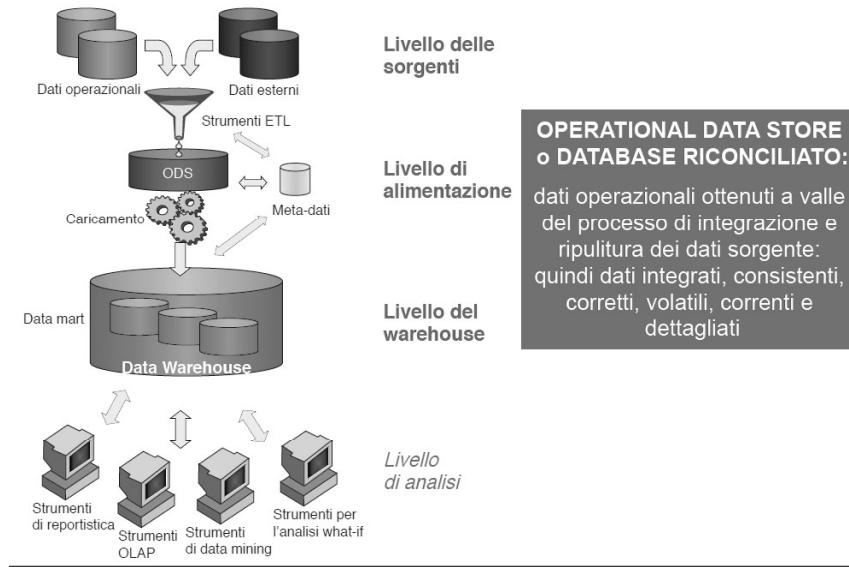
separare l’elaborazione di tipo analitico (OLAP) da quella legata alle transazioni (OLTP), costruendo un nuovo raccoglitrice di informazioni che integri i dati provenienti da sorgenti di varia natura, li organizzi in una forma appropriata e li renda disponibili per scopi di analisi e valutazione finalizzate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



- Illustrare il ruolo e l'utilità dell'**ODS** (*Operational Data Store* o *database riconciliato*) all'interno di un'architettura a 3 livelli.

■ Database riconciliato o Operational Data Store (ODS)

- ✓ Se presente, è parte integrante dell'architettura
- ✓ Espone un modello coerente del business, contiene dati normalizzati e può essere usato per la reportistica operativa
- ✓ Non è sinonimo di "data warehouse" (anche se alcuni lo chiamano erroneamente "data warehouse")



- Spiegare i concetti di base del *modello multidimensionale classico*

Il modello multidimensionale

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

- Illustrare il costrutto DFM di *arco multiplo*, spiegando in particolare i problemi che esso induce sull'aggregazione e come può essere tradotto in un'implementazione ROLAP

- Un *arco multiplo* modella un' associazione multi-a-molti tra due attributi dimensionali



Il DFM	Golfarelli, Rizzi	3
Mi Sembra Logico	Golfarelli	5
La Giusta Misura	Rizzi	10
Un Fatto Come e Perché	Golfarelli, Rizzi	4
La Quarta Dimensione	Golfarelli	8

Quanto ha venduto Rizzi?

- Un arco multiplo corrisponde a un' associazione a-molti R da un' entità E a un' entità G; nello schema di fatto, esso potrà allora connettere l' identificatore di E o il fatto con un attributo di R o di G

- Spiegare la differenza tra misure di flusso, di livello e unitarie, con particolare riferimento agli aspetti di additività.

- L'aggregazione richiede di definire un operatore adatto per comporre i valori delle misure che caratterizzano gli eventi primari in valori da abbinare a ciascun evento secondario
- Da questo punto di vista è possibile distinguere tre categorie di misure:
 - ✓ Misure di flusso: si riferiscono a un periodo, al cui termine vengono valutate in modo cumulativo (il numero di prodotti venduti in un giorno, l'incasso mensile, il numero di nati in un anno)
 - ✓ Misure di livello: vengono valutate in particolari istanti di tempo (il numero di prodotti in inventario, il numero di abitanti di una città)
 - ✓ Misure unitarie: vengono valutate in particolari istanti di tempo, ma sono espresse in termini relativi (il prezzo unitario di un prodotto, la percentuale di sconto, il cambio di una valuta)

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

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

- Illustrare il costrutto DFM di attributo cross-dimensional, spiegando in particolare i problemi che esso induce sull'aggregazione e come può essere tradotto in un'implementazione ROLAP.

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

Un attributo cross-dimensional corrisponde in genere a un attributo posto su un' associazione molti-a-molti R dello schema E/R; i suoi padri nello schema di fatto corrisponderanno allora agli identificatori delle entità coinvolte in R

- Spiegare la differenza tra operatori di aggregazione distributivi, algebrici e olistici.

- ✓ Distributivi: permettono di calcolare dati aggregati a partire direttamente da dati parzialmente aggregati (es. somma, massimo, minimo)
- ✓ Algebrici: richiedono un numero finito di informazioni aggiuntive (*misure di supporto*) per calcolare dati aggregati a partire da dati parzialmente aggregati (es. media – richiede il numero dei dati elementari che hanno contribuito a formare un singolo dato parzialmente aggregato)
- ✓ Olistici: non permettono di calcolare dati aggregati a partire da dati parzialmente aggregati utilizzando un numero finito di informazioni aggiuntive (es. mediana, moda)

- Illustrare il costrutto DFM di gerarchia incompleta, spiegando in particolare i pro e contro delle varie soluzioni di bilanciamento.

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

- Si elenchino e si definiscano almeno tre fattori che caratterizzano la qualità dei dati in un data warehouse

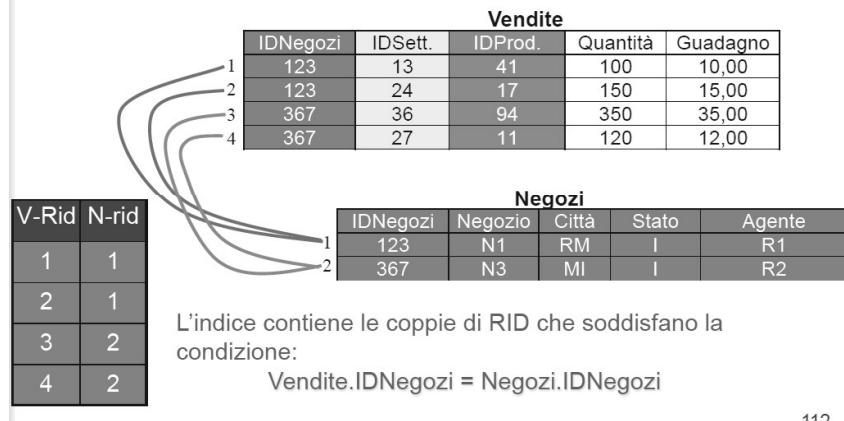
Fattori che caratterizzano la qualità dei dati in un DW:

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

- Illustrare la struttura e il funzionamento di un *join index*

Gli indici di join

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



112

- Si discutano pro e contro dei tre principali approcci all'analisi dei dati multidimensionali: *reportistica statica*, *OLAP* e *reportistica semi-statica*.

Reportistica semi-statica

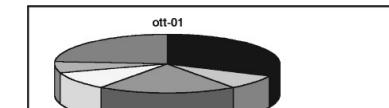
- In molti contesti applicativi, è utile un approccio intermedio tra reportistica statica e OLAP: la *reportistica semi-statica*
 - ✓ Un rapporto semi-statico, pur essendo focalizzato su un'insieme di informazioni predefinite, permette all'analista alcuni gradi di libertà, che si concretizzano nella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

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

Reportistica

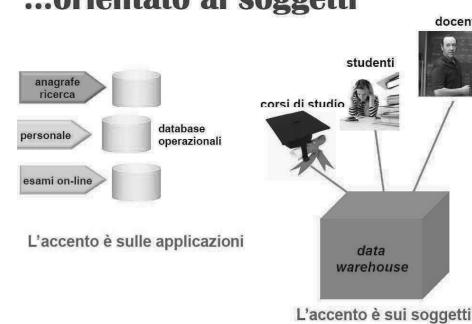
orientato agli utenti
che hanno necessità
di accedere, a
intervalli di tempo
predefiniti, a
informazioni
strutturate in modo
pressoché invariabile

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



- Si spieghi cosa intende, nella definizione classica di data warehouse, per "contenitore di dati orientato ai soggetti e non volatile".

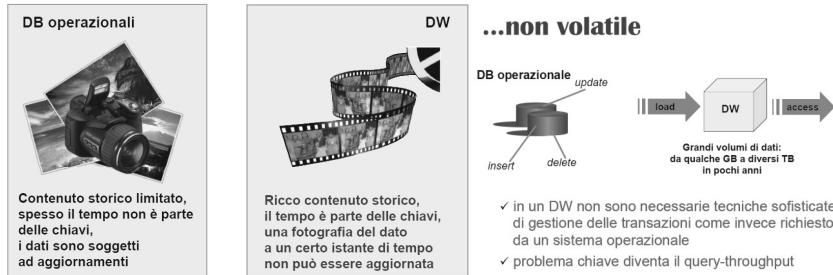
...orientato ai soggetti



...integrato e
consistente

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

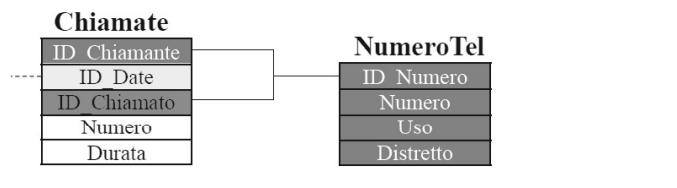
...rappresentativo dell'evoluzione temporale



- Si elenchino i costrutti DFM attraverso i quali è possibile creare dei cicli nelle gerarchie, illustrandone le differenze semantiche.

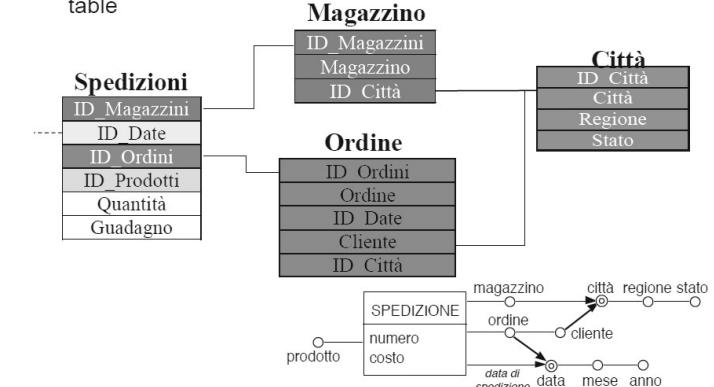
Gerarchie condivise

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



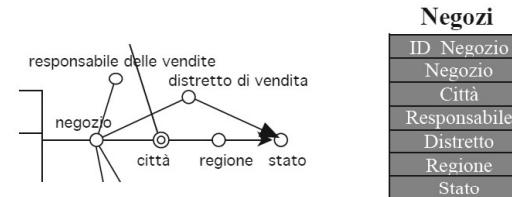
Gerarchie condivise

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



Convergenza

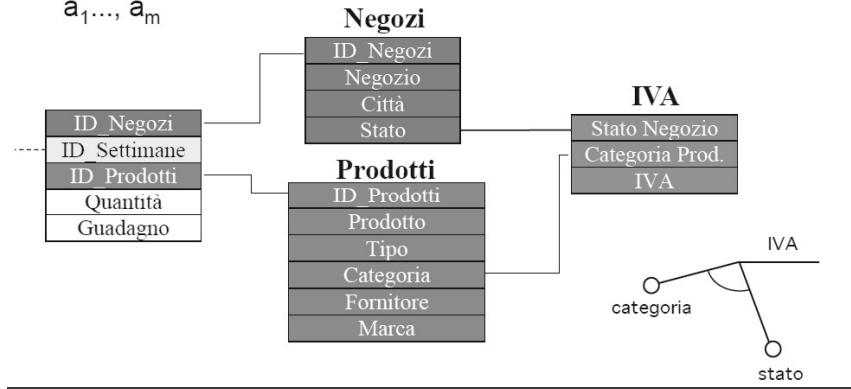
- Gli attributi di convergenza si includono nella stessa dimension table dei loro attributi padri, senza particolari accorgimenti



Attributi cross-dimensional

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

- Discutere pro e contro degli approcci *top-down* e *bottom-up* alla progettazione di data warehouse.



Approccio top-down

- Analizza i bisogni globali dell' intera azienda e pianifica lo sviluppo del DW per poi progettarlo e realizzarlo nella sua interezza
 - ↳ Promette ottimi risultati poiché si basa su una visione globale dell' obiettivo e garantisce in linea di principio di produrre un DW consistente e ben integrato
 - ↳ Il preventivo di costi onerosi a fronte di lunghi tempi di realizzazione scoraggia la direzione dall' intraprendere il progetto
 - ↳ Affrontare contemporaneamente l' analisi e la riconciliazione di tutte le sorgenti di interesse è estremamente complesso
 - ↳ Riuscire a prevedere a priori nel dettaglio le esigenze delle diverse aree aziendali impegnate è pressoché impossibile, e il processo di analisi rischia di subire una paralisi
 - ↳ Il fatto di non prevedere la consegna a breve termine di un prototipo non permette agli utenti di verificare l' utilità del progetto e ne fa scemare l' interesse e la fiducia

Approccio bottom-up

- Il DW viene costruito in modo incrementale, assemblando iterativamente più data mart, ciascuno dei quali incentrato su un insieme di fatti collegati a uno specifico settore aziendale e di interesse per una certa categoria di utenti
 - ↳ Determina risultati concreti in tempi brevi
 - ↳ Non richiede elevati investimenti finanziari
 - ↳ Permette di studiare solo le problematiche relative al data mart in oggetto
 - ↳ Fornisce alla dirigenza aziendale un riscontro immediato sull' effettiva utilità del sistema in via di realizzazione
 - ↳ Mantiene costantemente elevata l' attenzione sul progetto
 - ↳ Determina una visione parziale del dominio di interesse

- Si discutano pro e contro delle diverse tecniche per l'estrazione incrementale dei dati dalle sorgenti.

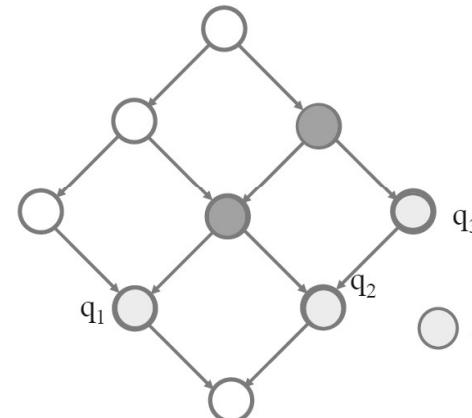
Estrazione statica → Riscrittura completa

Estrazione incrementale

- Livello riconciliato non storizzato: memorizzo solo il tipo di operazione che ha determinato la variazione
- Livello riconciliato storizzato: memorizzo anche una coppia di marche temporali che indicano l' intervallo di validità della tupla

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

- Spiegare in che modo, dato un reticolo multidimensionale e un carico di lavoro, si possono determinare le viste candidate alla materializzazione.



$\text{white circle} + \text{dark grey circle} = \text{viste candidate, ossia potenzialmente utili a ridurre il costo di esecuzione del carico di lavoro}$

- Illustrare, anche con l'aiuto di un esempio, il ruolo delle misure di supporto nell'utilizzo degli operatori di aggregazione algebrici.

Algebrici: richiedono un numero finito di informazioni aggiuntive (*misure di supporto*) per calcolare dati aggregati a partire da dati parzialmente aggregati (es. media – richiede il numero dei dati elementari che hanno contribuito a formare un singolo dato parzialmente aggregato)

- Spiegare in che modo uno scenario temporale di retrodatabilità può essere implementato in una piattaforma ROLAP.

Ieri per oggi (*retrodatabilità*)

- ✓ I dati vengono interpretati in base alla configurazione della gerarchia valida in un particolare istante
- ✓ Richiede la storicizzazione dei dati

- Illustrare il costrutto di gerarchia condivisa nel DFM, descrivendone le possibili implementazioni su piattaforma ROLAP.

- Illustrare le differenze tra i costrutti di gerarchia condivisa e convergenza.

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

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

- Cos'è il group-by set di un'interrogazione OLAP e in che modo è implicitamente rappresentato nel DFM?

Dato un insieme di attributi dimensionali (*group-by set*), ciascuna ennupla di loro valori individua un *evento secondario* che aggrega tutti gli eventi primari corrispondenti. A ciascun evento secondario è associato un valore per ciascuna misura, che riassume in sé tutti i valori della stessa misura negli eventi primari corrispondenti

- ✓ Pertanto, le gerarchie definiscono il modo in cui gli eventi primari possono essere aggregati e selezionati significativamente per il processo decisionale; mentre la dimensione in cui una gerarchia ha radice ne definisce la granularità più fine di aggregazione, agli altri attributi dimensionali corrispondono granularità via via crescenti

- Spiegare punti in comune e differenze tra i costrutti di gerarchia incompleta e gerarchia ricorsiva nel DFM.

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

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

- Discutere pro e contro degli schemi snowflake rispetto agli schemi a stella.

- Si elenchino le differenti situazioni in cui può risultare vantaggioso effettuare lo snowflaking di una dimensione.

Esistono pareri contrastanti sull' utilità dello snowflaking:

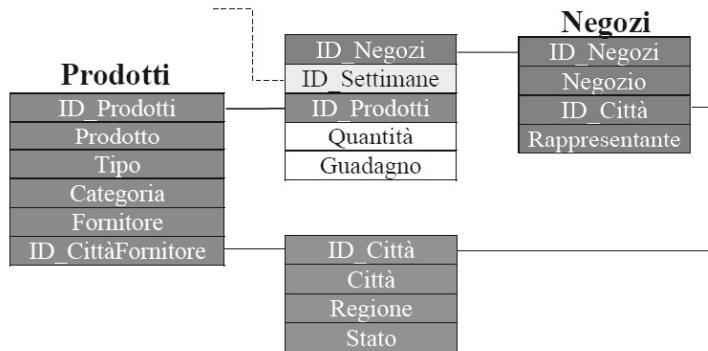
- ✓ Contrasta con la filosofia del data warehousing
- ✓ Rappresenta un inutile "abbellimento" dello schema

Può essere utile

1. Quando il rapporto tra le cardinalità della dimension table primaria e secondaria è elevato, poiché determina un forte risparmio di spazio

Può essere utile

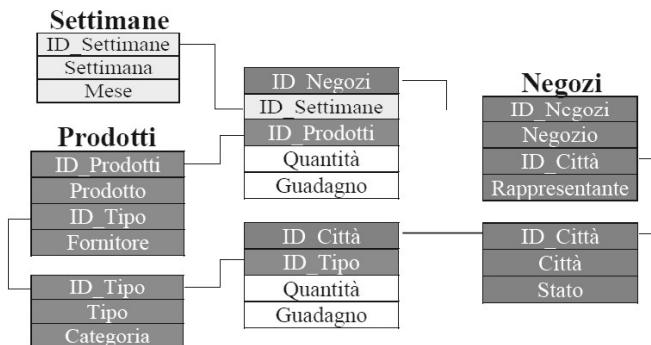
Quando una porzione di una gerarchia è comune a più dimensioni



La dimension table secondaria è riutilizzata per più gerarchie

Può essere utile

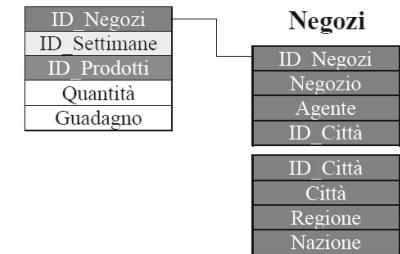
3. In presenza di viste aggregate



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

Può essere utile

4. Quando una parte della gerarchia è soggetta a frequenti aggiornamenti



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

- Si elenchino i fattori che incidono sulla scelta dell'architettura in un progetto di data warehouse.

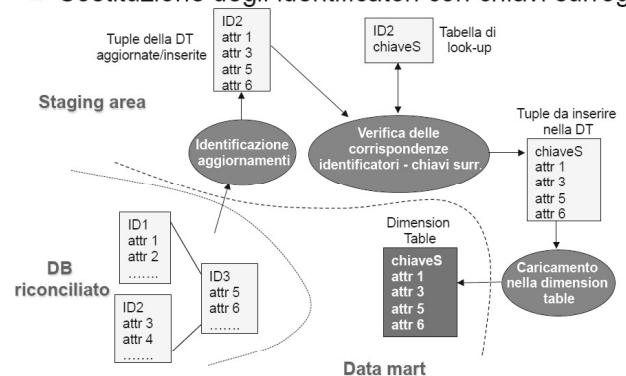
Architettura del sistema di data warehousing (tipo di architettura da implementare, numero dei livelli, presenza di data mart dipendenti o indipendenti, materializzazione del livello riconciliato)

- Spiegare l'importanza dell'uso di *chiavi surrogate* nella progettazione logica di data mart.

Alimentazione delle dimension table

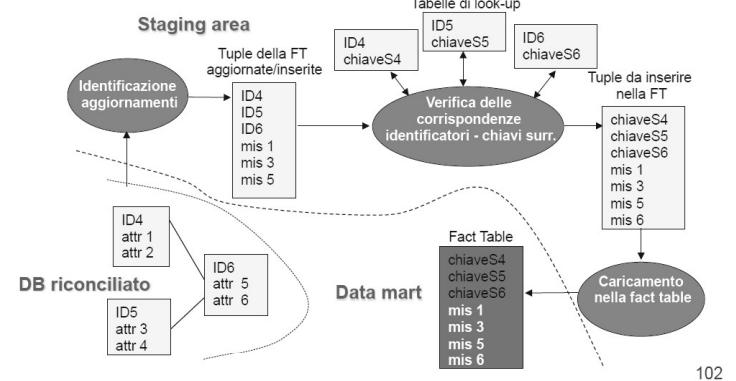
table

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



Alimentazione delle fact table

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



102

Lo snowflake schema: considerazioni

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

- Illustrare la differenza tra *operational data store* e *staging area* e i rispettivi ruoli architettonici.

Data Warehouse

- ✓ Contiene dati denormalizzati (tipicamente star e snowflake schema)

Data Mart

- ✓ Corrisponde a una porzione del data warehouse
- ✓ Include più cubi multidimensionali, per cui "data mart" non è sinonimo di "cubo multidimensionale"

Cubo multidimensionale

- ✓ Implementa un fatto che lo modella a livello concettuale

Misura

- ✓ Chiamata anche "metrica" o "KPI"

Database riconciliato o Operational Data Store (ODS)

- ✓ Se presente, è parte integrante dell'architettura
- ✓ Espone un modello coerente del business, contiene dati normalizzati e può essere usato per la reportistica operativa
- ✓ Non è sinonimo di "data warehouse" (anche se alcuni lo chiamano erroneamente "data warehouse")

Staging area

- ✓ Area di lavoro dell'ETL, spesso erroneamente confusa con l'ODS
- ✓ Ha un ruolo di servizio all'interno dell'architettura

- Si illustrano le possibili soluzioni di progettazione logica a supporto della *materializzazione di viste*.

Scelta delle viste

- È utile materializzare una vista quando:
 - ✓ Risolve direttamente una interrogazione frequente
 - ✓ Permette di ridurre il costo di esecuzione di molte interrogazioni
- Non è consigliabile materializzare una vista quando:
 - ✓ Il suo group-by set è molto simile a quello di una vista già materializzata
 - ✓ Il suo group-by set è molto fine
 - ✓ La materializzazione non riduce di almeno un ordine di grandezza il costo delle interrogazioni

Frammentazione verticale

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

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

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

- Spiegare cos'è la *business intelligence*.

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

- Si parla di *piattaforma* di BI poiché per consentire ai manager analisi potenti e flessibili è necessario definire un'apposita infrastruttura hardware e software di supporto composta da:

- ✓ Hardware dedicato
- ✓ Infrastrutture di rete
- ✓ DBMS
- ✓ Software di back-end
- ✓ Software di front-end

- Il ruolo chiave di una piattaforma di business intelligence è la trasformazione dei *dati* aziendali in *informazioni* fruibili a diversi livelli di dettaglio e, quindi, in *conoscenza*

- Spiegare attraverso quali architetture di data warehouse e con quali accorgimenti metodologici è possibile ottenere l'integrabilità dei *data mart* sviluppati.

- Si spieghi il ruolo della *tabella di look-up* nell'ETL, con particolare riferimento all'alimentazione di una dimension table gestita come slowly-changing dimension di tipo 1.

- Spiegare le differenze concettuali tra i due costrutti del DFM che permettono di rappresentare gerarchie con istanze a lunghezza variabile.

- Spiegare il ruolo della fase di *progettazione fisica* all'interno del ciclo di vita di un *data mart*.

- Si elenchino i fattori che incidono sulla scelta dell'architettura in un progetto di data warehouse.