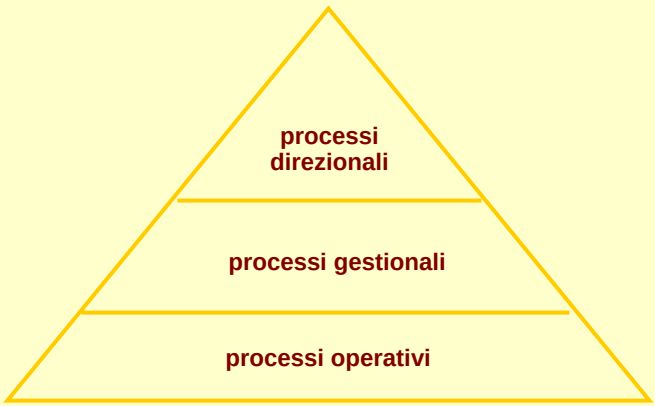
Tecnologie per i sistemi informativi

I data warehouse e la loro progettazione

Docente: Letizia Tanca
Politecnico di Milano
tanca@elet.polimi.it

Processi



Le query che vorremmo poter formulare ai livelli superiori

- Incassi registrati lo scorso anno per ciascuna regione e ciascuna categoria di prodotto
- Correlazione tra l'andamento dei titoli azionari dei produttori di computer e i profitti trimestrali degli ultimi 5 anni
- Quali sono gli ordini che massimizzano gli incassi?
- Quale di due nuove terapie risulterà in una diminuzione della durata media dei ricoveri?

Il problema

- Una promessa della tecnologia relazionale: "flexible data access":
  - Uno strumento per l'utente finale in cui tutte le query siano ugualmente formulabili
- Ma la tecnologia relazionale non ha mantenuto questa promessa:
  - Complessità e rigidità delle applicazioni
  - Enfasi su OLTP
- Le conseguenze:
  - Masse di dati per la gestione operativa
  - Scarso utilizzo dei dati per la gestione strategica

OLTP

Tradizionale elaborazione di transazioni, che realizzano i processi operativi dell'azienda-ente
- Operazioni spesso predefinite e relativamente semplici
- Ogni operazione coinvolge "pochi" dati
- Dati di dettaglio, aggiornati
- Le proprietà "acide" (atomicità, correttezza, isolamento, durabilità) delle transazioni sono essenziali

Le dimensioni delle basi di dati sono dell'ordine dei gbyte

La principale metrica di prestazione e' il throughput delle transazioni

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

Le dimensioni del warehouse raggiungono facilmente i terabyte

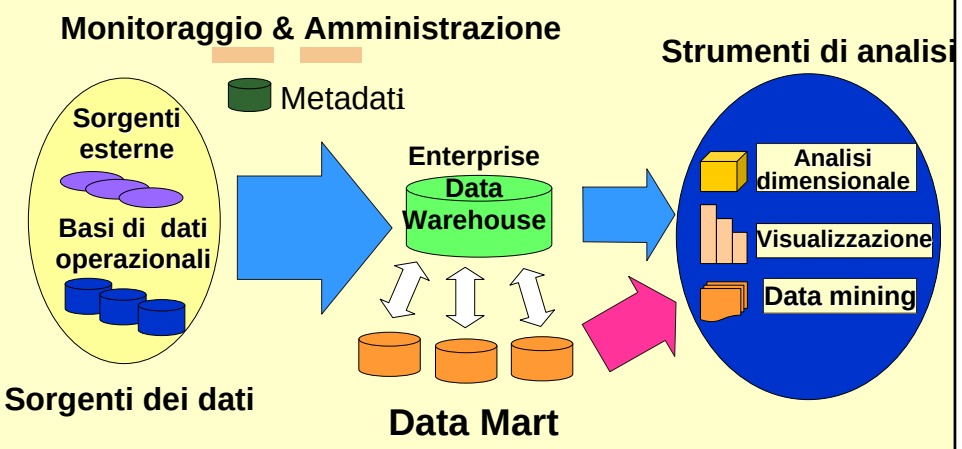
Le prestazioni considerate sono il throughput delle interrogazioni e il loro tempo di risposta

Data warehouse

Una base di dati

- utilizzata principalmente per il supporto alle decisioni direzionali
- integrata — aziendale e non dipartimentale
- orientata ai dati — non alle applicazioni
- Dati storici — con un ampio orizzonte temporale, e indicazione (di solito) di elementi di tempo
- non volatile — i dati sono caricati e acceduti fuori linea
- mantenuta separatamente dalle basi di dati operazionali

Architettura per il data warehousing



DW e Data Mart

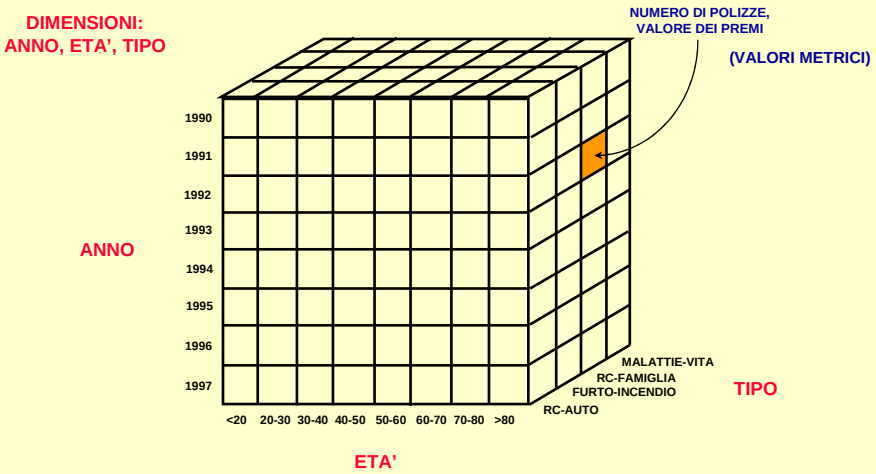
- Una DW spesso integra diversi Data Mart
- Gli utenti normalmente si rivolgono a un particolare Data Mart
- I Data Mart condividono dati tra di loro
- Ciascun Data Mart è responsabile di un particolare aspetto della realtà aziendale, **non necessariamente** di un reparto.
- Es: data mart della rilevazione e rendicontazione di tutte le attività di un ospedale

MODELLI DEI DATI PER OLAP

- devono supportare analisi e calcoli sofisticati su diverse dimensioni e gerarchie
- il modello logico dei dati piu' adatto e' una struttura multidimensionale - il data cube
- le dimensioni del cubo sono costituite dagli attributi secondo i quali si vogliono fare le ricerche (chiavi)
- ogni dimensione puo' "contenere" a sua volta una gerarchia
  - DATA {GIORNO - MESE – TRIMESTRE - ANNO}
  - PRODOTTO {NOME - TIPO - CATEGORIA}
    (LAND ROVER - FUORISTRADA - AUTOVEICOLI)
- le celle del cubo contengono i valori metrici relativi ai valori dimensionali

MODELLI LOGICI DEI DATI PER OLAP

ESEMPIO PER UNA COMPAGNIA DI ASSICURAZIONI



Rappresentazione multidimensionale

Concetti rilevanti:

- **fatto —** un concetto sul quale centrare l'analisi - modella un evento che accade nell'azienda
- **misura —** una proprietà atomica di un fatto da analizzare – ne descrive un aspetto quantitativo
- **dimensione —** descrive una prospettiva lungo la quale effettuare l'analisi

Dimensioni e gerarchie di livelli

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

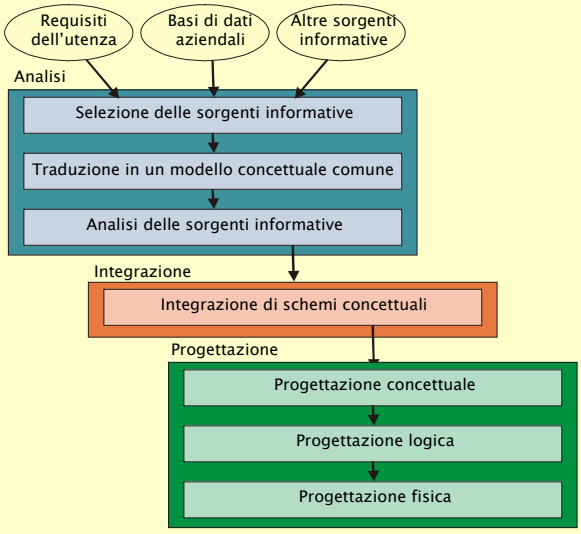
Progettazione di Data Warehouse

- La progettazione di una data warehouse è diversa dalla progettazione di una base di dati operazionale
  - i dati da memorizzare hanno caratteristiche diverse
  - vincolata dalle basi di dati esistenti
  - guidata da criteri progettuali diversi
- Enfasi sulla generalizzazione e chiarezza concettuale
  - poche entità
  - ampia copertura

Progettazione di data warehouse

- Limitata frammentazione (denormalizzazione): l'analisi ha bisogno di una visione sintetica dei dati
- Nella progettazione sarà importante mettere in evidenza gli aspetti comuni
- Bisogna associare a ciascuna entità un significato che corrisponda all'intuizione dell'utente
- Fondamentale il ruolo dei metadati nel guidare all'uso del sistema
- Attività principali
  - analisi — delle sorgenti informative esistenti
  - integrazione
  - progettazione — concettuale, logica e fisica

Progettazione di data warehouse



Dati in ingresso

Le informazioni in ingresso necessarie alla progettazione di un data warehouse

- requisiti — le esigenze aziendali di analisi
- descrizione delle basi di dati operazionali — con una documentazione sufficiente per la loro comprensione
- descrizione di altre sorgenti informative — l'analisi richiede spesso la correlazione con dati non di proprietà dell'azienda ma comunque da essa accessibili — ad esempio, dati ISTAT o sull'andamento dei concorrenti

COSTRUZIONE DEL WAREHOUSE

- **i dati provengono da sorgenti diverse, probabilmente "sporche"**
  - **sistemi legacy non documentati**
  - **sistemi di produzione senza check di integrita' interni**
  - **sorgenti esterne di dubbia qualita'**
    - scarsa qualità del data entry
    - assenza di dati in alcuni campi

e' indispensabile restituire la qualita' ai dati per potervi basare decisioni affidabili

COSTRUZIONE DEL WAREHOUSE

- **strumenti per la qualita' dei dati**
  - **per la migrazione**
    trasformano e riformattano i dati dalle diverse fonti
  - **per la pulizia (scrubbing)**
    usano la conoscenza del dominio per pulire e omogeneizzare

    `<jerry l. jonson, 16 clarke st., altuna, pa>  =`

    `<gerry l. johnson, 16 clark street, altoona, penn> ???`
  - **per il controllo (auditing)**
    scoprono regole e relazioni tra i dati e ne verificano il rispetto
- **strumenti per il caricamento dei dati**
  verificano violazioni di integrita' referenziale; ordinano, aggregano, costruiscono dati derivati; costruiscono indici e altri percorsi di accesso

Analisi delle sorgenti informative esistenti

- **Selezione delle sorgenti informative**
  - analisi preliminare del patrimonio informativo aziendale — analisi di qualità delle singole sorgenti
  - correlazione del patrimonio informativo con i requisiti
  - identificazione di priorità tra schemi
- **Traduzione in un modello concettuale di riferimento**
  - attività preliminare alla correlazione e all'integrazione di schemi — che si svolge meglio con riferimento a schemi concettuali
- **Analisi delle sorgenti informative**
  - identificazione di massima dei fatti, o concetti su cui basare l'analisi, delle loro misure (proprietà atomiche), e delle dimensioni (concetti su cui aggregare le misure)

Integrazione di sorgenti informative

- *L'integrazione di sorgenti informative* è l'attività di fusione dei dati rappresentati in più sorgenti in un'unica *base di dati globale* (fisica o virtuale) che rappresenta l'intero patrimonio informativo a disposizione
- Lo scopo principale dell'integrazione è l'*identificazione* di tutte le porzioni delle diverse sorgenti informative che si riferiscono a uno stesso aspetto della realtà di interesse, per *unificare* la loro rappresentazione
- L'approccio è orientato alla *identificazione, analisi e risoluzione di conflitti* — terminologici, strutturali, di codifica

Progettazione del data warehouse

- L'integrazione delle sorgenti informative ha prodotto una descrizione globale del patrimonio informativo aziendale
- Questo è però solo il risultato dell'integrazione di dati operazionali — non descrive tutti i dati di interesse per il DW
- Progettazione del data warehouse
  - **concettuale** — completare la rappresentazione dei concetti dimensionali necessari per l'analisi — ad esempio, dati storici e geografici
  - **logica** — identificare il miglior compromesso tra la necessità di aggregare i dati e quella di normalizzarli
  - **fisica** — individuare la distribuzione dei dati e le relative strutture di accesso

Progettazione del DW

- Introduzione di elementi dimensionali nella base di dati integrata
- Attività
  - identificazione di fatti, misure e dimensioni
  - ristrutturazione dello schema concettuale
    - rappresentazione di fatti mediante entità
    - individuazione di nuove dimensioni
    - raffinamento dei livelli di ogni dimensione
  - derivazione di un grafo dimensionale
  - progettazione logica e fisica

Strategie di progetto

- **Approccio top-down:**
  - **Interessante perché garantisce la coerenza interna del progetto**
  - **Spesso fallimentare perché impresa lunga e ardua che scoraggia l'utenza**
  - **L'analisi contemporanea di tutte le fonti informative è compito molto complesso**
  - **La previsione contemporanea delle esigenze informative di tutti gli utenti è difficile e rischia di paralizzare il processo**
- **Approccio bottom-up:**
  - **Costruzione incrementale assemblando più data mart**
  - **Si abbina a tecniche di prototipazione veloce**
  - **Incoraggia l'utenza che vede velocemente il prodotto operativo**

Approccio bottom up

Si sceglie di progettare per primo il data mart più strategico per l'azienda, che guida l'integrazione successiva

Ciclo di vita progetto bottom up

- Definizione degli obiettivi e pianificazione
- Progettazione dell'infrastruttura tecnologica
- Progettazione dei data mart e loro progressiva integrazione

La progettazione dei data mart

- **Partire dagli schemi dei dati operazionali esistenti:**
  - **In questo modo non si rischia di promettere all'utente delle funzionalità che è impossibile fornire**
- **Si riesce a tradurre lo schema concettuale dei dati operazionali in quello del data mart in modo quasi algoritmico**

Progettazione di data warehouse

Raccolta e analisi dei requisiti utente

Scopo fondamentale di questa fase è la scelta dei FATTI del data mart

- **Tecniche:**
  - **A risposte aperte:**
    - **Pro: Più coinvolgenti, insegnano al progettista il linguaggio del dominio applicativo…**
    - **Contro: possono comportare risposte lunghe e dispersive**
  - **A risposte chiuse:**
    - **Pro: riducono i tempi di intervista, permettono il confronto tra risposte…**
    - **Contro: possono risultare noiose per l'intervistato**
  - **A risposte probatorie**
    - **Pro: permettono di capire il livello di competenza dell'intervistato…**
    - **Contro: possono comportare ansia nell'intervistato che si sente indagato**

Esempi di domande

- **Per un dirigente:**
  - Quali sono gli obiettivi aziendali?
  - Come misuri il successo dell'azienda?
- **Per un capo reparto:**
  - Quali sono gli obiettivi del tuo reparto?
  - Quali sono i colli di bottiglia nell'accesso ai dati?
  - Quali analisi eseguite di solito sui dati? Quali vorreste eseguire?
- **Per l'amministratore del sistema informativo:**
  - Quali sono le caratteristiche delle fonti di dati disponibili?
  - Che strumenti vengono attualmente usati per analizzare i dati?
  - Quali analisi vi vengono richieste di solito sui dati?

Il glossario dei requisiti utente

FATTO	POSSIBILI DIMENSIONI	POSSIBILI MISURE	STORICITA'
inventario di magazzino	*prodotto, data, magazzino*	*quantità in magazzino*	*1 anno*
vendite	*prodotto, data, negozio*	*quantità venduta, importo, sconto*	*5 anni*
linee d'ordine	*prodotto, data, fornitore*	*quantità ordinata, importo, sconto*	*3 anni*

Il carico di lavoro preliminare

FATTO	POSSIBILI INTERROGAZIONI AGGREGATE
inventario di magazzino	Scorte medie di ciascun prodotto presenti mensilmente Prodotti per i quali è stata esaurita la scorta almeno una volta in tutti i magazzini durante l'ultima settimana
vendite	Quantità totali di ciascun tipo di prodotto vendute ogni mese per regione Incasso totale giornaliero di ciascun negozio Per ciascun negozio, riepilogo delle vendite di ciascuna categoria di prodotto per ogni giorno
linee d'ordine	Quantità totale ordinata annualmente presso un dato fornitore Importo giornaliero ordinato nell'ultimo mese per ciascuna categoria di prodotto

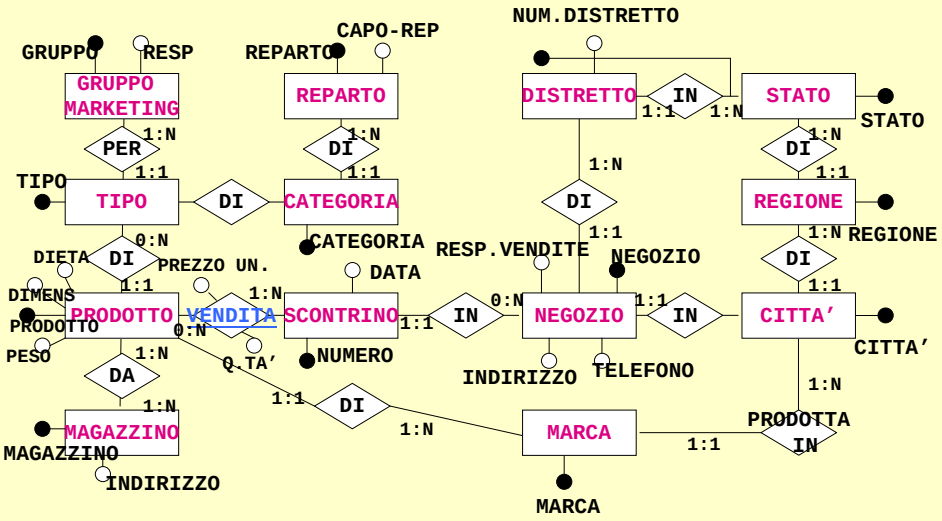
Ci serve a capire, ad esempio, che:

- **Per il magazzino:**
  - Serve sapere, per ogni prodotto, il tipo
  - La gerarchia temporale deve includere almeno il mese e la settimana
  - Occorre poter calcolare le medie delle quantità
- **Per le vendite:**
  - Serve sapere, per ogni prodotto, il tipo e la categoria
  - La gerarchia della dimensione negozio deve includere l'attributo regione
- **Per le linee d'ordine:**
  - Serve sapere, per ogni prodotto, il tipo e la categoria
  - Durante l'aggregazione la misura importo viene sommata sulle gerarchie dei prodotti e dei fornitori

Progettazione di schemi concettuali di DW a partire da schemi ER

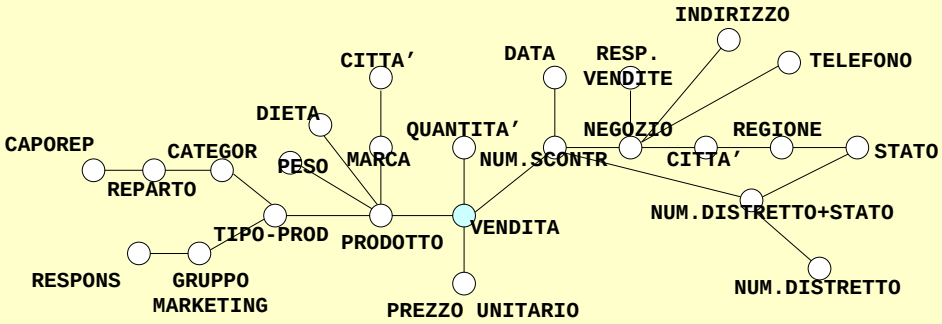
1) Definizione dei **fatti**
   - I **fatti** sono concetti di interesse per il processo decisionale, e corrispondono a eventi che accadono nel mondo aziendale
   - Le entità che rappresentano dati aggiornati frequentemente (es. VENDITE) sono **buoni candidati** per essere dei fatti - es. NEGOZIO, CITTÀ non lo sono
- Per ogni fatto:
   - Costruzione dell'**albero degli attributi**

BASE DI DATI OPERAZIONALE



L'albero degli attributi

Fatto: VENDITE



VENDITA = PRODOTTO + NUM. SCONTRINO

Costruzione dell'albero degli attributi

Fatto: VENDITE

- ogni vertice corrisponde a un attributo dello schema sorgente
- La radice corrisponde all'identificatore del fatto
- Per ogni vertice v dell'albero, l'attributo corrispondente determina funzionalmente gli attributi corrispondenti ai discendenti di v
- Le dipendenze funzionali si riconoscono, sono:
  - Identificatori
  - Associazioni "a-uno"

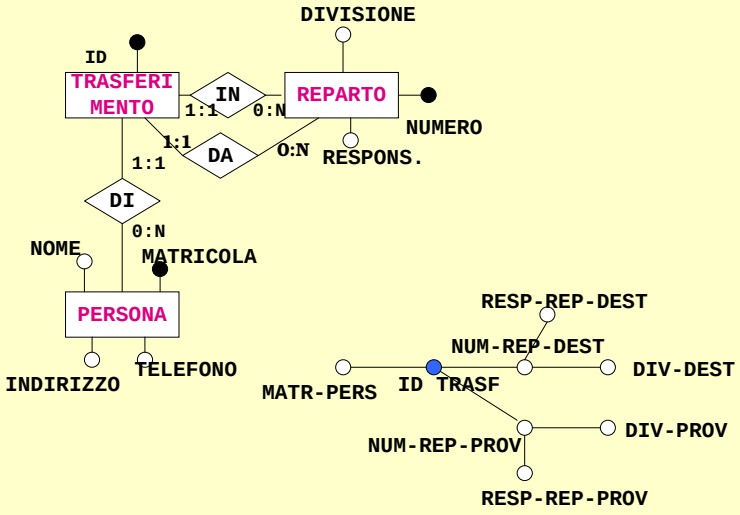
Si costruisce ricorsivamente attraversando il diagramma ER

- Si parte dall'entità corrispondente al fatto scelto (*radice*)
- Per ogni entità E attraversata, si crea nell'albero un vertice v corrispondente all'identificatore di E
- Si aggiunge un vertice figlio per ogni attributo di E (inclusi gli attributi che compongono l'id)
- Per ogni relazione "a-uno" uscente da E verso una entità G, si aggiungono anche tutti i suoi attributi come figli di v
- Si ripete il procedimento per G

Casi particolari

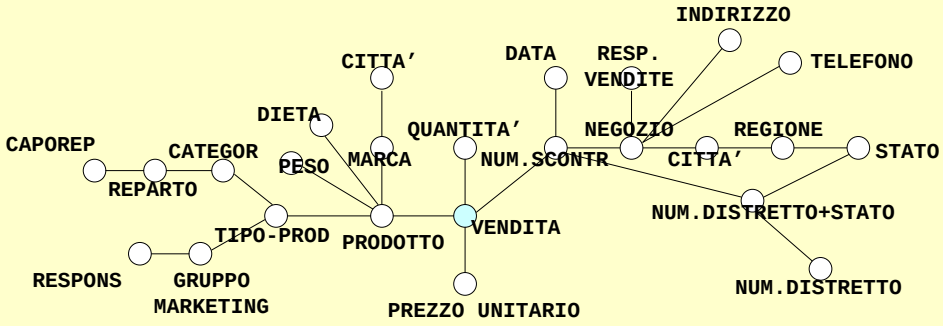
- Relazioni cicliche (es. parte-sottoparte) → si spezzano dopo un certo num. di iterazioni
- Cicli nello schema → si spezzano legando all'entità + conveniente (event. duplicando)
- associazioni n-arie o molti-a molti si possono *reificare,* cioè trasformare in entità
- Attributi opzionali: segnati da un trattino
- Gerarchie: equiv. ad associazioni 0-1 (trattino di opzionalità)
- Attributo composto: un vertice con figli

Cicli nello schema



L'albero degli attributi

Fatto: VENDITE →
relazione reificata



VENDITA = PRODOTTO + NUM. SCONTRINO

Potatura e innesto

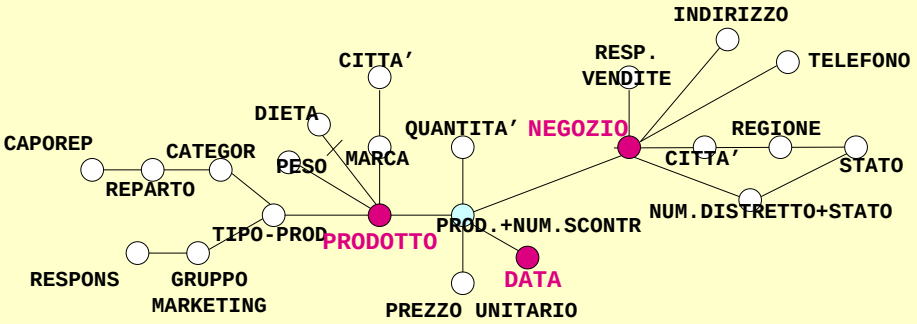
- **Potatura** di un vertice *v*: si elimina *v* con tutti i discendenti
- **Innesto** di *v* con padre *v'* : si elimina *v* e si collegano i figli di *v* direttamente a *v'*
- Es. potatura di *num-scontrino* e innesto di *data* e *negozio* su *vendite*

Progettazione di schemi concettuali di DW

2) Identificazione delle dimensioni

- **Le dimensioni vanno scelte nell'albero degli attributi tra i vertici figli della radice**
- **Possono corrispondere ad attributi discreti o ad intervalli di attributi continui**
- **Il tempo è una dimensione sempre significativa**
  - **→ Se non appare come figlio della radice, vale la pena di ristrutturare l'albero (potare e innestare) per farlo comparire**

Esempio DIMENSIONI



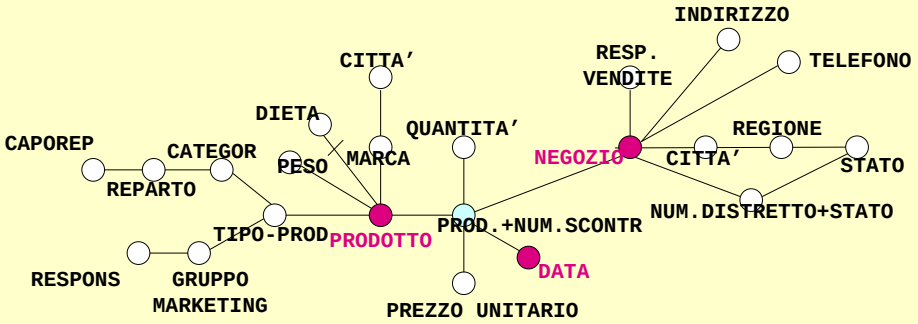
SCEGLIAMO COME DIMENSIONI
LA TERNA <DATA,NEGOZIO,PRODOTTO>

Progettazione di schemi concettuali di DW

3) Identificazione delle misure

- Di solito le misure vanno scelte nell'albero degli attributi tra i vertici figli della radice
- Questo non è assolutamente necessario, nel caso non si verificasse si può potare e innestare l'albero
- Le misure di solito si ottengono aggregando rispetto a tutte le dimensioni

Esempio MISURE



Quantità venduta = SUM(VENDITA.quantità)
Incasso=SUM(VENDITA.quantità*VENDITA.prezzoUnitario)
prezzoUnitario=AVG(VENDITA.prezzoUnitario)
NumClienti=COUNT(*)
IN CORRISPONDENZA DELLA STESSA TERNA <DATA,NEGOZIO,PRODOTTO>

Tecnologie target per la progettazione logica

MOLAP (**M**ultidimensional-**OLAP**) in alternativa a **ROLAP** (**R**elational-**OLAP**)

- **MOLAP:** usa strutture interne non relazionali, prestazioni migliori
- **ROLAP:** in grado di gestire grandi quantità di dati mediante le classiche tecnologie relazionali

SERVER OLAP MULTIDIMENSIONALE (MOLAP)

- implementa direttamente il modello a cubo
  - strutture a matrice multidimensionale
- prestazioni elevate e costanti per l'elaborazione delle interrogazioni
  - metodi di accesso specializzati
  - aggregazione e compilazione eseguite in precedenza
- limitata scalabilita' a causa delle preelaborazioni
- richiede maggiori capacita' da parte dell'amministratore dei dati

SERVER OLAP RELAZIONALE (ROLAP)

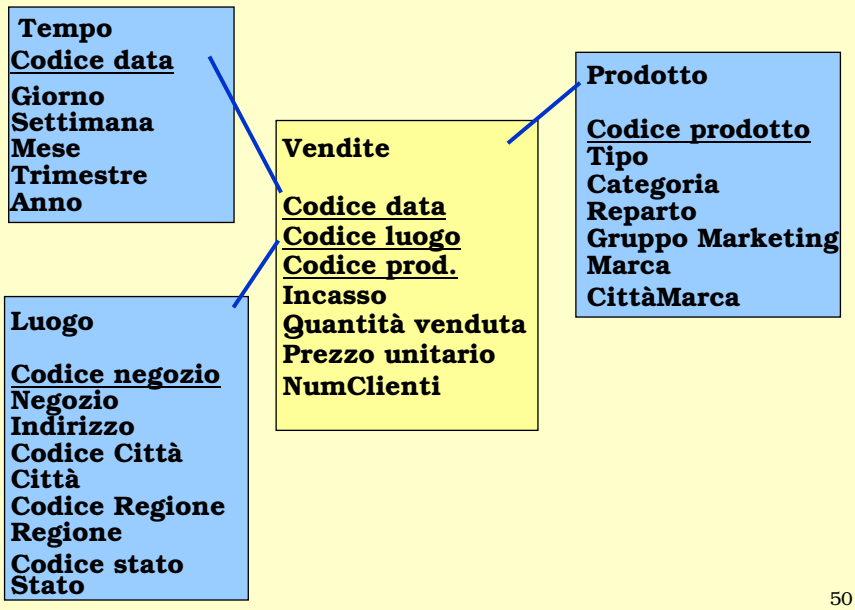
- utilizza un RDBMS standard per realizzare la struttura multidimensionale, applicando operatori aggregati
- lo schema assume una configurazione a stella
- Centro stella: fatto
- Punte stella: dimensioni
- Vantaggi:
  - Si possono usare opportune interfacce di interrogazione
  - Si ottengono buone prestazioni
  - Si ha un'immediata definizione dello schema logico

SERVER OLAP RELAZIONALE (ROLAP)

Lo schema assume una configurazione a stella o a fiocco di neve

- tabella centrale dei fatti
  - le tuple sono costituite dai puntatori (=chiavi esterne) alle tabelle di dimensione e dai valori per le misure descritte $f = (k_1, \ldots, k_n, v_1, \ldots v_m)$
- tabelle di dimensione
  - contengono le tuple con gli attributi relativi a quella dimensione $d_1=(k_1, a_1, \ldots, a_n)$
- costellazione di fatti
  - piu' tabelle dei fatti condividono tabelle di dimensione di uguale struttura

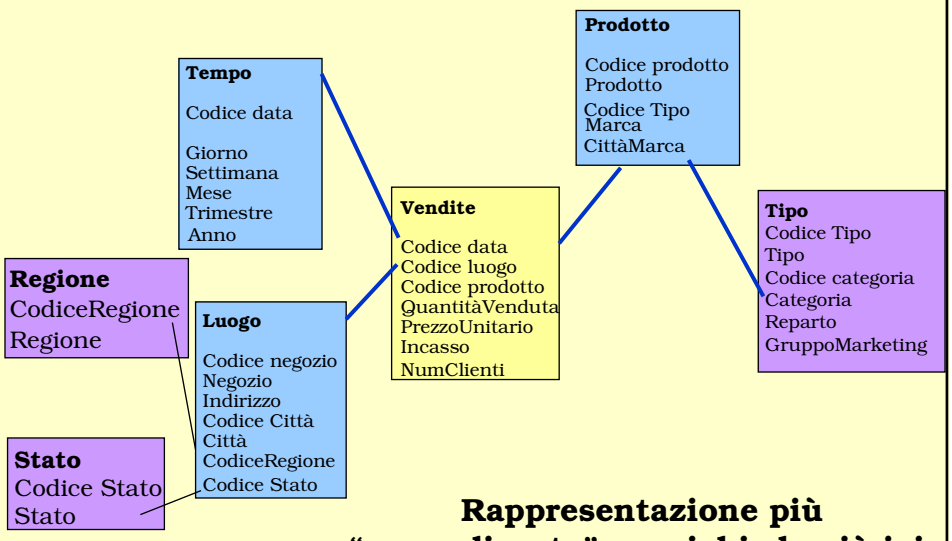
Organizzazione a stella



Modello a fiocco di neve (snow flake)

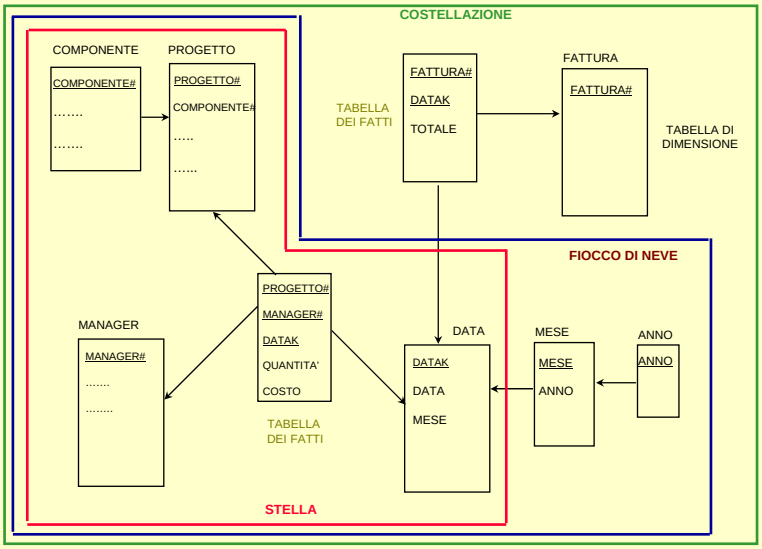
- Costituisce un'estensione del modello a stella
- Permette di evitare ridondanze eccessive nelle dimensioni
- Dalla tabella dei fatti si raggiungono tutte le tabelle delle dimensioni, sempre muovendosi lungo associazioni n:1

Organizzazione "snowflake"



Rappresentazione più "normalizzata" ma richiede più join

SCHEMI A STELLA



Strategia di implementazione

- Si realizza dapprima uno schema di massima e si individuano i dati comuni ai diversi data mart, definendo per essi uno schema condiviso (con l'approccio appena visto)
- Si realizzano poi uno o più Data Mart, con l'obiettivo di familiarizzare gli sviluppatori e gli utenti con la tecnologia OLAP
- Infine, si integrano i sistemi già sviluppati e si procede alla realizzazione della Enterprise Data Warehouse

Altri requisiti

Altri requisiti vengono raccolti per le fasi successive:

- **Per la progettazione logica e fisica: requisiti di spazio e tempo – ottimizzazione degli accessi tramite viste materializzate**
- **Per il progetto dell'alimentazione: periodicità dell'alimentazione, grado di "freschezza" richiesto**

Tecnologie per il progetto fisico

- Indici join
  - Precomputano il join tra la dimensione e la tabella dei fatti
- Materializzazione di viste
  - Vengono precalcolate le viste che possono essere utilizzate come base per rispondere alle query più frequenti