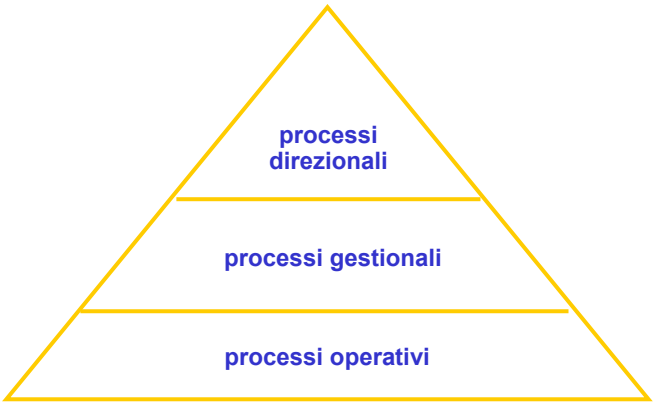
Data Warehousing

Luca Cabibbo, Riccardo Torlone, Paolo Atzeni

Sommario

Processi



Processi, dati e decisioni

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

Processi presso una banca

- Processi operativi
  - gestione di un movimento su un conto corrente bancario, presso sportello tradizionale o automatico
- Processi gestionali
  - concessione di un fido
  - revisione delle condizioni su un conto corrente
- Processi direzionali
  - verifica dell'andamento dei servizi di carta di credito
  - lancio di una campagna promozionale
  - stipula di accordi commerciali

Processi presso una compagnia telefonica

- Processi operativi
  - stipula di contratti ordinari
  - instradamento delle telefonate
  - memorizzazione di dati contabili sulle telefonate (chiamante, chiamato, giorno, ora, durata, instradamento,..)
- Processi gestionali
  - stipula di contratti speciali
  - installazione di infrastrutture
- Processi direzionali
  - scelta dei parametri che fissano il costo delle telefonate
  - definizione di contratti diversificati
  - pianificazione del potenziamento delle infrastrutture

Caratteristiche dei processi dei vari tipi

- processi operativi
  - operano sui dati dipartimentali e dettagliati
  - le operazioni sono strutturate, basate su regole perfettamente definite
- processi gestionali
  - operano su dati settoriali e parzialmente aggregati
  - le operazioni sono semi-strutturate, basate su regole note, ma in cui è spesso necessario un intervento umano "creativo"
- processi direzionali
  - operano su dati integrati e fortemente aggregati
  - le operazioni sono non strutturate, non esistono criteri precisi, e la capacità personale è essenziale

Sistemi informatici: una classificazione

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

Sistemi di supporto alle decisioni

- I sistemi di supporto alle decisioni (DSS) costituiscono la tecnologia che supporta la dirigenza aziendale nel prendere decisioni tattico-strategiche in modo efficace e veloce, mediante particolari tipologie di elaborazione (per esempio OLAP)
- Ma su quali dati?
  - quelli accumulati per i processi operativi e gestionali

Tipi di elaborazione

- Nei Transaction Processing Systems:
  - On-Line Transaction Processing
- nei management information systems
  - On-Line Transaction Processing + applicazioni evolute (“intelligenti”)
- nei Decision Support Systems:
  - On-Line Analytical Processing

OLTP

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

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

OLTP e OLAP

	OLTP	OLAP
Utente	impiegato	dirigente
Funzione	operazioni giornaliere	supporto alle decisioni
Progettazione	orientata all'applicazione	orientata ai dati
Dati	correnti, aggiornati, dettagliati, relazionali, omogenei	storici, aggregati, multidimensionali, eterogenei
Uso	ripetitivo	casuale
Accesso	read-write, indicizzato	read, sequenziale
Unità di lavoro	transazione breve	interrogazione complessa
Record acc.	decine	milioni
N. utenti	migliaia	centinaia
Dimensione	100MB - 1GB	100GB - 1TB
Metrica	throughput	tempo di risposta

Evoluzione dei DSS

- Anni '60 — rapporti batch
  - difficile trovare e analizzare dati
  - ogni richiesta richiede un nuovo programma
- Anni '70 — DSS basato su terminale
  - accesso ai dati operazionali
- Anni '80 — strumenti d'automazione d'ufficio e di analisi
  - fogli elettronici, interfacce grafiche
- Anni '90 — data warehousing
  - strumenti di OLAP

OLTP e OLAP

- I requisiti sono quindi contrastanti
- Le applicazioni dei due tipi possono danneggiarsi a vicenda

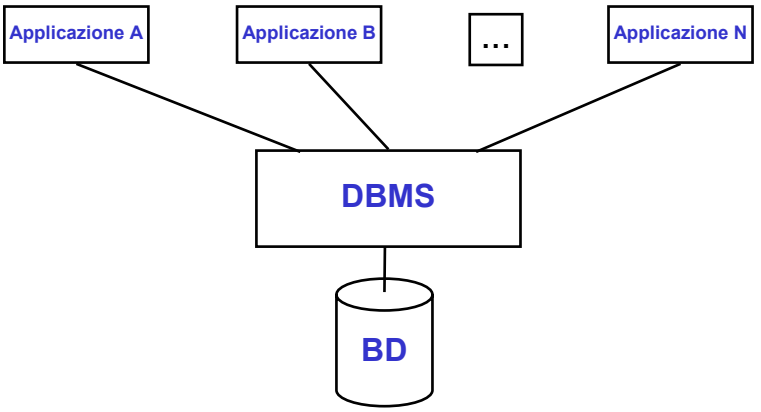
Base di dati

- "Collezione di dati **persistente** e **condivisa**, gestita in modo **efficace**, **efficiente** e **affidabile** (da un **DBMS**)"
- il concetto di base di dati nasce per rispondere alle esigenze di "gestione di una risorsa pregiata", condivisa da più applicazioni

Basi di dati:
"le magnifiche sorti e progressive"

- “ogni organizzazione ha **una** base di dati, che organizza tutti i dati di interesse in forma integrata e non ridondante”
- “ciascuna applicazione ha accesso a tutti i dati di proprio interesse, in tempo reale e senza duplicazione, riorganizzati secondo le proprie necessità”
- “bla bla bla ...”

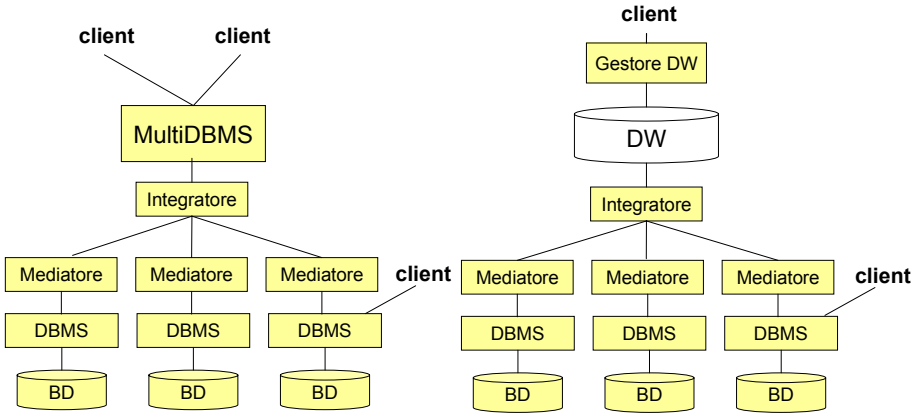
La base di dati “ideale”



L'obiettivo ideale è sensato e praticabile?

- La realtà è in continua evoluzione, non esiste uno "stato stazionario" (se non nell'iperuranio):
  - cambiano le esigenze
  - cambiano le strutture
  - le realizzazioni richiedono tempo
- Il coordinamento forte fra i vari settori può risultare controproducente
- Ogni organizzazione ha di solito diverse basi di dati **distribuite**, **eterogenee**, **autonome**

Multi-database e Data Warehouse (due approcci all'integrazione)



Sommario

- Introduzione
  - Basi di dati integrate, sì, ma …
  - OLTP e OLAP
- ⇨ Data warehousing
  - Data warehouse e data warehousing
  - Dati multidimensionali
- Progettazione di data warehouse

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
- con dati storici — con un ampio orizzonte temporale, e indicazione (di solito) di elementi di tempo
- con dati usualmente aggregati — per effettuare stimee valutazioni
- fuori linea — i dati sono aggiornati periodicamente
- mantenuta separatamente dalle basi di dati operazionali

... integrata ...

- I dati di interesse provengono da tutte le sorgenti informative — ciascun dato proviene da una o più di esse
- Il data warehouse rappresenta i dati in modo univoco — riconciliando le eterogeneità dalle diverse rappresentazioni
  - nomi
  - struttura
  - codifica
  - rappresentazione multipla

... orientata ai dati ...

- Le basi di dati operazionali sono costruite a supporto dei singoli processi operativi o applicazioni
  - produzione
  - vendita
- Il data warehouse è costruito attorno alle principali entità del patrimonio informativo aziendale
  - prodotto
  - cliente

... dati storici ...

- Le basi di dati operazionali mantengono il valore corrente delle informazioni
  - L'orizzonte temporale di interesse è dell'ordine dei pochi mesi
- Nel data warehouse è di interesse l'evoluzione storica delle informazioni
  - L'orizzonte temporale di interesse è dell'ordine degli anni

... dati aggregati ...

- Nelle attività di analisi dei dati per il supporto alle decisioni
  - non interessa "chi" ma "quanti"
  - non interessa un dato ma
    - la somma,
    - la media,
    - il minimo e il massimo, ...

  di un insieme di dati.
- Le operazioni di aggregazione sono quindi fondamentali nel warehousing e nella costruzione/mantenimento di un data warehouse.

... fuori linea ...

- In una base di dati operazionale, i dati vengono
  - acceduti
  - inseriti
  - modificati
  - Cancellati

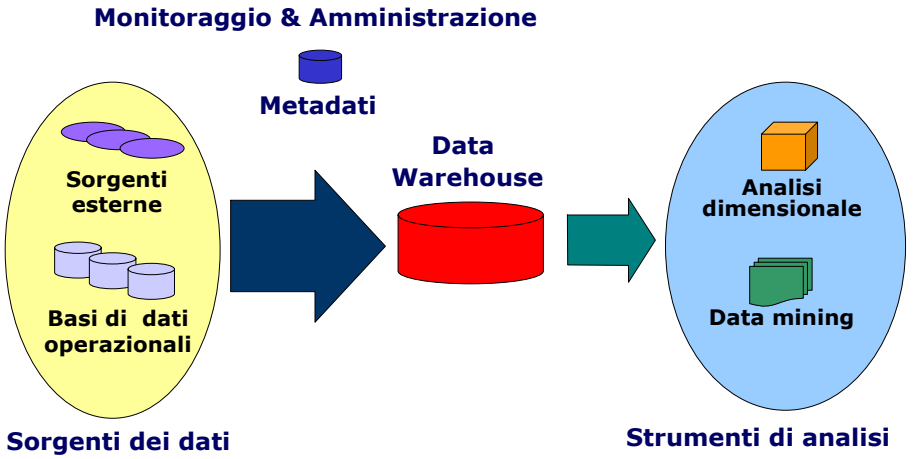
  pochi record alla volta
- Nel data warehouse, abbiamo
  - operazioni di accesso e interrogazione — “diurne”
  - operazioni di caricamento e aggiornamento dei dati — “notturne”

  che riguardano milioni di record

... una base di dati separata ...

- Un data warehouse viene mantenuto separatamente dalle basi di dati operazionali perché
  - non esiste un’unica base di dati operazionale che contiene tutti i dati di interesse
  - la base di dati deve essere integrata
  - non è tecnicamente possibile fare l’integrazione in linea
  - i dati di interesse sarebbero comunque diversi
    - devono essere mantenuti dati storici
    - devono essere mantenuti dati aggregati
  - l’analisi dei dati richiede per i dati organizzazioni speciali e metodi di accesso specifici
  - degrado generale delle prestazioni senza la separazione

Architettura per il data warehousing



Due incisi

- Metadati
- Data mining

Metadati

- "Dati sui dati":
  - descrizioni logiche e fisiche dei dati (nelle sorgenti e nel DW)
  - corrispondenze e trasformazioni
  - dati quantitativi
- Spesso sono non dichiarativi e immersi nei programmi!

Data mining

- Approccio alternativo all'analisi multidimensionale per estrarre informazioni di supporto alle decisioni da un data warehouse
- Insieme di tecniche di ricerca di "informazione nascosta" in una collezione di dati
- Spesso applicate a dati "destrutturati" (collezioni di transazioni)

Problemi classici di data mining

- **associazioni**: individuare regolarità in un insieme di transazioni anonime
- **pattern sequenziali**: individuare regolarità in un insieme di transazioni non anonime, nel corso di un periodo temporale
- **classificazione**: catalogare un fenomeno in una classe predefinita sulla base di fenomeni già catalogati

Associazioni

- Dati di ingresso:
  - sequenze di oggetti (transazioni)
- Obiettivo:
  - trovare delle “regole” che correlano la presenza di un insieme di oggetti con un altro insieme di oggetti

Esempio di regola

Pannolini $\Rightarrow$ Birra

- il 30% delle transazioni che contiene Pannolini contiene anche Birra
- il 2% tra tutte le transazioni contiene entrambi gli oggetti

Rilevanza delle regole

X, Y $\Rightarrow$ Z

- **Supporto S**: la regola è verificata in S% delle transazioni rispetto a tutte le transazioni
  - rilevanza statistica
- **Confidenza C**: C% di tutte le transazioni che contengono X e Y contengono anche Z
  - "forza" della regola

Pattern sequenziali

- Dati di ingresso:
  - insieme di transazioni eseguita da un certo cliente
- Obiettivo:
  - trovare le sequenze di oggetti che compaiono in almeno una certa percentuale data di insiemi di transazioni

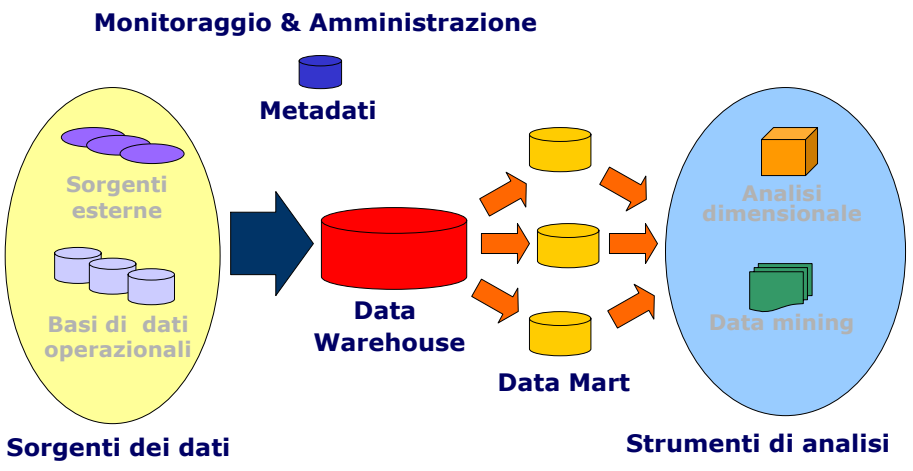
Esempi

- “Il 5% dei clienti ha comprato un lettore di CD in una transazione e CD in un’altra”
  - il 5% è il supporto del pattern
- Applicazioni
  - misura della soddisfazione del cliente
  - promozioni mirate
  - medicina (sintomi - malattia)

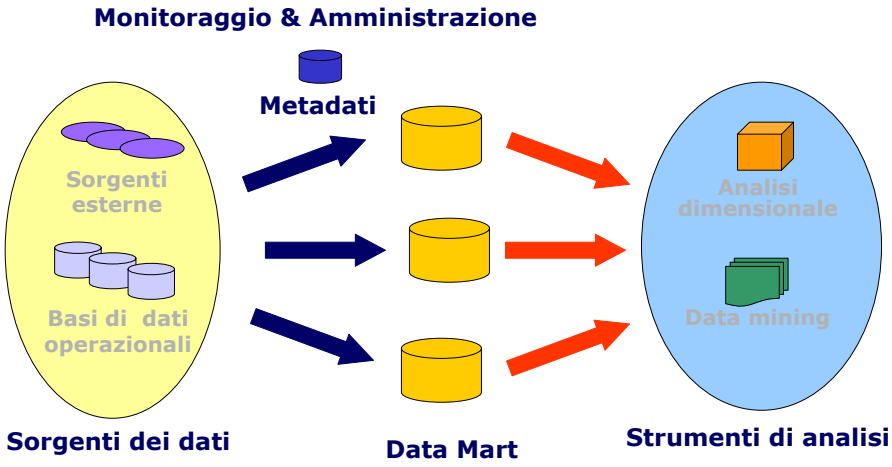
Data mart

- Un sottoinsieme logico dell'intero data warehouse
  - un data mart è la restrizione del data warehouse a un singolo processo
  - un data warehouse è l'unione di tutti i suoi data mart
- Pro e contro dei data mart
  - un data mart rappresenta un progetto solitamente fattibile
    - la realizzazione diretta di un data warehouse completo non è invece solitamente fattibile
  - tuttavia, la realizzazione di un insieme di data mart non porta necessariamente alla realizzazione del data warehouse

Variante dell'architettura



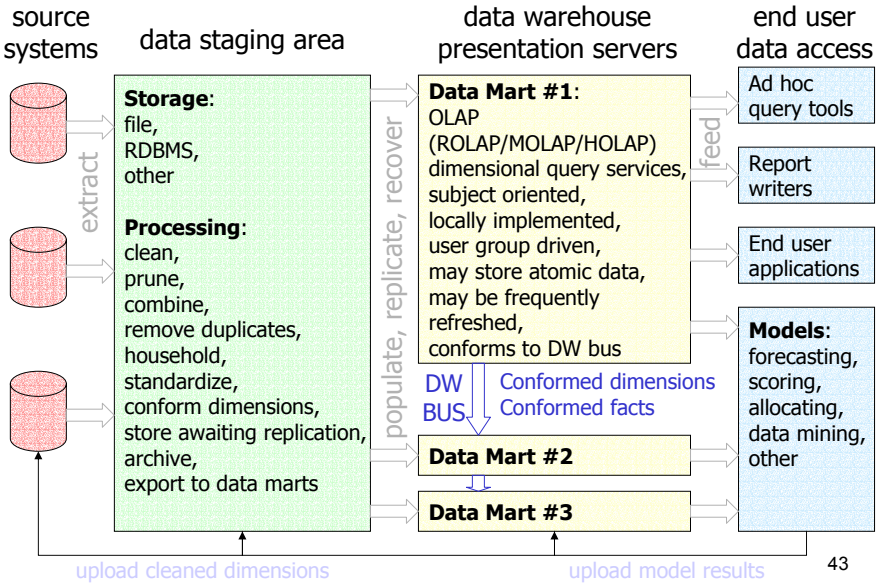
Altra variante



Top-down o bottom-up?

- Prima il data warehouse o prima i data mart?
- Non c'è risposta, o meglio: nessuno dei due!
- Infatti:
  - l'approccio è spesso incrementale
  - è necessario coordinare i data mart: dimensioni conformi

Elementi di un data warehouse



Sorgenti informative

- i sistemi operazionali dell'organizzazione
  - sono sistemi transazionali (OLTP) orientati alla gestione dei processi operazionali
  - non mantengono dati storici
  - ogni sistema gestisce uno o più soggetti (ad esempio, prodotti o clienti)
    - nell'ambito di un processo
    - ma non in modo conforme nell'ambito dell'organizzazione
  - sono sistemi "legacy"
- sorgenti esterne
  - ad esempio, dati forniti da società specializzate di analisi

Area di preparazione dei dati

- L'**area di preparazione** dei dati (**data staging**) è usata per il transito dei dati dalle sorgenti informative al data warehouse
  - comprende ogni cosa tra le sorgenti informative e i server di presentazione
    - aree di memorizzazione dei dati estratti dalle sorgenti informative e preparati per il caricamento nel data warehouse
    - processi per la preparazione di tali dati
      - pulizia, trasformazione, combinazione, rimozione di duplicati, archiviazione, preparazione per l'uso nel data warehouse
  - è un insieme complesso di attività semplici
  - è distribuita su più calcolatori e ambienti eterogenei
  - i dati sono memorizzati prevalentemente su file

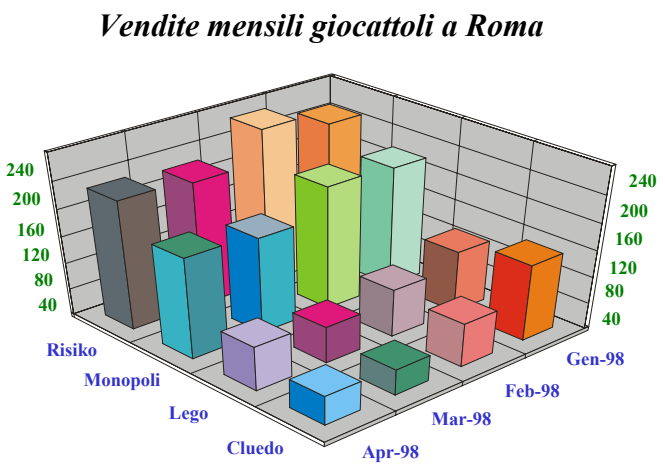
Server di presentazione

- Un **server di presentazione** è un sistema in cui i dati del data warehouse sono organizzati e memorizzati per essere interrogati direttamente da utenti finali, report writer e altre applicazioni
  - i dati sono rappresentati in forma **dimensionale**
    - (secondo i concetti di fatto e dimensione, vediamo fra poco)
  - tecnologie che possono essere adottate
    - RDBMS
      - i dati sono organizzati mediante schemi dimensionali (schemi a stella)
    - tecnologia OLAP
      - i concetti di fatto e dimensione sono espliciti
    - i produttori di RDBMS stanno iniziando a fornire estensioni OLAP ai loro prodotti

Visualizzazione dei dati

- I dati vengono infine visualizzati in veste grafica, in maniera da essere facilmente comprensibili.
- Si fa uso di:
  - tabelle
  - istogrammi
  - grafici
  - torte
  - superfici 3D
  - bolle
  - area in pila
  - forme varie
  - …

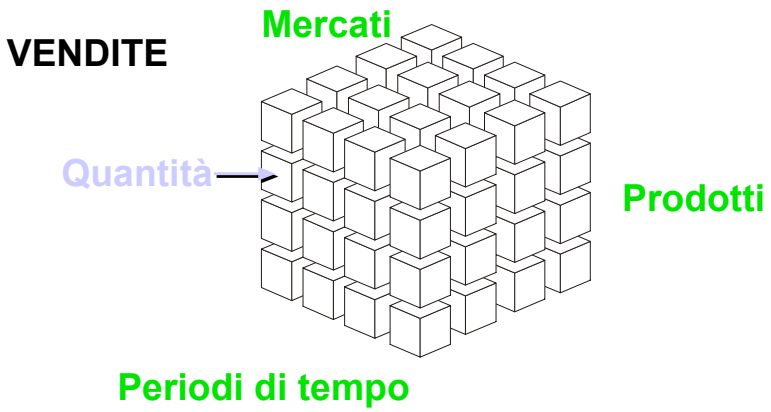
Visualizzazione finale di un'analisi



Rappresentazione multidimensionale

- L'analisi dei dati avviene rappresentando i dati in forma **multidimensionale**
- Concetti rilevanti:
  - fatto — un concetto sul quale centrare l'analisi
  - misura — una proprietà atomica di un fatto da analizzare
  - dimensione — descrive una prospettiva lungo la quale effettuare l'analisi
- Esempi di fatti/misure/dimensioni
  - vendita / quantità venduta, incasso / prodotto, tempo
  - telefonata / costo, durata / chiamante, chiamato, tempo

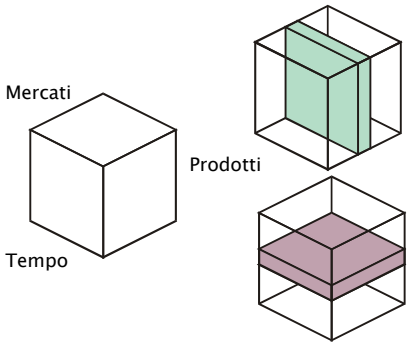
Rappresentazione multidimensionale dei dati



Viste su dati multidimensionali

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

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



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

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

Operazioni su dati multidimensionali

- **Roll up (o drill up)**— aggrega i dati
  - volume di vendita totale dello scorso anno per categoria di prodotto e regione
- **Drill down** — disaggrega i dati
  - per una particolare categoria di prodotto e regione, mostra le vendite giornaliere dettagliate per ciascun negozio
- **Slice & dice** — seleziona e proietta
- **(Pivot** — re-orienta il cubo)

	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze 1	21	4	10	4	6	7
Firenze 2	4	4	4	6	6	3
Roma 1	15	5	8	3	5	20
Roma 2	12	4	7	5	2	4
Roma 3	23	4	9	10	5	5
Latina	3	3	5	1	2	4

	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze 1	21	4	10	4	6	7
Firenze 2	4	4	4	6	6	3
Roma 1	15	5	8	3	5	20
Roma 2	12	4	7	5	2	4
Roma 3	23	4	9	10	5	5
Latina	3	3	5	1	2	4

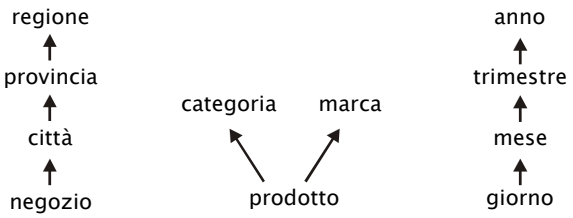
	Gen	Feb	Mar	Apr	Mag	Giu
	90	26	53	32	32	48

	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze 1	21	4	10	4	6	7
Firenze 2	4	4	4	6	6	3
Roma 1	15	5	8	3	5	20
Roma 2	12	4	7	5	2	4
Roma 3	23	4	9	10	5	5
Latina	3	3	5	1	2	4

Pisa	38
Firenze 1	52
Firenze 2	27
Roma 1	56
Roma 2	34
Roma 3	56
Latina	18

Dimensioni e gerarchie di livelli

- Ciascuna dimensione è organizzata in una gerarchia che rappresenta i possibili livelli di aggregazione per i dati
  - negozio, città, provincia, regione
  - prodotto, categoria, marca
  - giorno, mese, trimestre, anno



	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze 1	21	4	10	4	6	7
Firenze 2	4	4	4	6	6	3
Roma 1	15	5	8	3	5	20
Roma 2	12	4	7	5	2	4
Roma 3	23	4	9	10	5	5
Latina	3	3	5	1	2	4

	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze	25	8	14	10	12	10
Roma	50	13	24	18	12	29
Latina	3	3	5	1	2	4

	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze 1	21	4	10	4	6	7
Firenze 2	4	4	4	6	6	3
Roma 1	15	5	8	3	5	20
Roma 2	12	4	7	5	2	4
Roma 3	23	4	9	10	5	5
Latina	3	3	5	1	2	4

	Gen	Feb	Mar	Apr	Mag	Giu
Toscana	37	10	24	13	18	15
Lazio	53	16	29	19	14	33

	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze 1	21	4	10	4	6	7
Firenze 2	4	4	4	6	6	3
Roma 1	15	5	8	3	5	20
Roma 2	12	4	7	5	2	4
Roma 3	23	4	9	10	5	5
Latina	3	3	5	1	2	4

	I trim	II trim
Pisa	24	14
Firenze 1	35	17
Firenze 2	12	15
Roma 1	28	28
Roma 2	23	11
Roma 3	36	20
Latina	11	7

	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze 1	21	4	10	4	6	7
Firenze 2	4	4	4	6	6	3
Roma 1	15	5	8	3	5	20
Roma 2	12	4	7	5	2	4
Roma 3	23	4	9	10	5	5
Latina	3	3	5	1	2	4

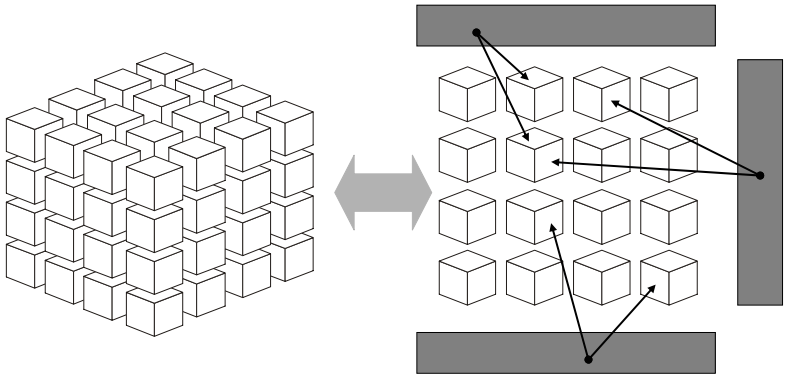
	I trim	II trim
Pisa	24	14
Firenze 1	35	17
Firenze 2	12	15
Roma 1	28	28
Roma 2	23	11
Roma 3	36	20
Latina	11	7

	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze	25	8	14	10	12	10
Roma	50	13	24	18	12	29
Latina	3	3	5	1	2	4

	I trim	II trim
Pisa	24	14
Firenze	47	32
Roma	87	59
Latina	11	7

Implementazione MOLAP

- I dati sono memorizzati direttamente in un formato dimensionale (proprietario). Le gerarchie sui livelli sono codificate in indici di accesso alle matrici



Implementazione ROLAP: schemi dimensionali

- Uno **schema dimensionale** (**schema a stella**) è composto da
  - una tabella principale, chiamata **tabella fatti**
    - la tabella fatti memorizza i fatti misurabili di un processo
      - i fatti più comuni sono numerici, continui e additivi
  - due o più tabelle ausiliarie, chiamate **tabelle dimensione**
    - una tabella dimensione rappresenta una dimensione rispetto alla quale è interessante analizzare i fatti
      - memorizza i membri che caratterizzano la grana dei fatti, nonché gli attributi usati dalle interrogazioni per vincolare e raggruppare i fatti
      - gli attributi sono solitamente testuali, discreti e descrittivi

Schema dimensionale: dimensioni semplici

CodN	Nome
PI	Pisa
FI1	Firenze 1
FI2	Firenze 2
RM1	Roma 1
RM2	Roma 2
RM3	Roma 3
LT	Latina

Negozio	Mese	Vendite
PI	Gen	12
PI	Feb	2
PI	Mar	10
PI	Apr	3
PI	Mag	6
PI	Giu	5
FI1	Gen	21
FI1	Feb	4
FI1	Mar	10
FI1	Apr	4
FI1	Mag	6
FI1	Giu	7
…	…	…

CodM	Mese
Gen	gennaio
Feb	febbraio
Mar	marzo
Apr	aprile
Mag	maggio
Giu	giugno

Schema dimensionale: dimensioni con livelli

CodN	…	Città	Regione	…
PI	…	Pisa	Toscana	…
FI1	…	Firenze	Toscana	…
FI2	…	Firenze	Toscana	…
RM1	…	Roma	Lazio	…
RM2	…	Roma	Lazio	…
RM3	…	Roma	Lazio	…
LT	…	Latina	Lazio	…

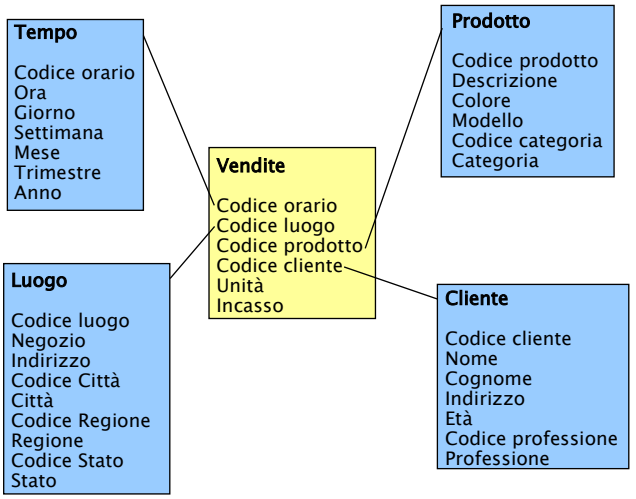
Negozio	Mese	Vendite
PI	Gen	12
PI	Feb	2
PI	Mar	10
PI	Apr	3
PI	Mag	6
PI	Giu	5
FI1	Gen	21
FI1	Feb	4
FI1	Mar	10
FI1	Apr	4
FI1	Mag	6
FI1	Giu	7
…	…	…

CodM	Mese	Trimestre
Gen	gennaio	I trim
Feb	febbraio	I trim
Mar	marzo	I trim
Apr	aprile	II trim
Mag	maggio	II trim
Giu	giugno	II trim

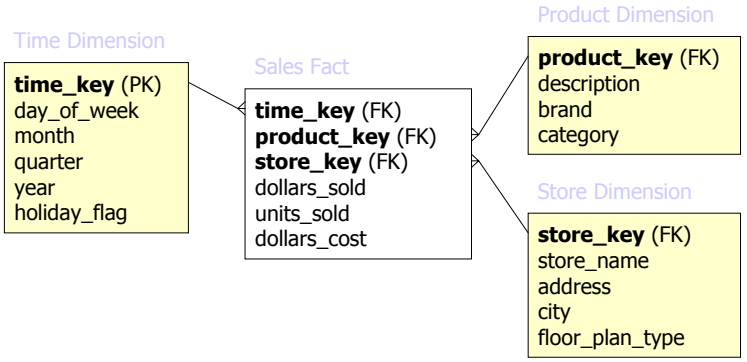
Approccio dimensionale

- uno **schema dimensionale** è uno schema relazionale di forma particolare (**star schema**, o schema a stella)
- lo schema di un data warehouse è un insieme di schemi dimensionali
  - ogni data mart è un insieme di schemi dimensionali
  - tutti i data mart vengono costruiti usando
    - dimensioni conformi
      - ogni dimensione ha lo stesso significato in ciascuno schema dimensionale e data mart
    - fatti conformi
      - anche i fatti hanno interpretazione uniforme

Uno schema dimensionale



Un altro schema dimensionale



- i dati delle vendite di prodotti in un certo numero di negozi nel corso del tempo
  - memorizza i totali giornalieri delle vendite dei prodotti per negozio

Caratteristiche di uno schema dimensionale

- Una tabella dimensione memorizza i membri di una dimensione
  - la chiave primaria è semplice
  - gli altri campi memorizzano gli attributi della dimensione
    - gli attributi sono solitamente testuali, discreti e descrittivi
- La tabella fatti memorizza le misure (fatti) di un processo
  - la chiave è composta da riferimenti alle chiavi di tabelle dimensione
  - gli altri campi rappresentano le misure
    - I fatti (le misure) sono solitamente numerici, continui e additivi

Tabelle dimensione

- Memorizza i membri di una dimensione rispetto alla quale è interessante analizzare un processo (e le relative descrizioni)
  - ciascun record di una tabella dimensione descrive esattamente un elemento della rispettiva dimensione
    - un record di Time Dimension descrive un giorno (nell'ambito dell'intervallo temporale di interesse)
    - un record di Product Dimension descrive un prodotto in vendita nei negozi
  - i campi (non chiave) memorizzano gli attributi dei membri
    - gli attributi sono le proprietà dei membri, che sono solitamente testuali, discrete e descrittive

Tabella fatti

- memorizza le misure numeriche di un processo
  - ogni record della tabella fatti memorizza una ennupla di misure (fatti) relativa a una combinazione dei membri, presa all'intersezione di tutte le dimensioni
- Nell'esempio
  - il processo (i fatti) è la vendita di prodotti nei negozi
  - le misure (i fatti) sono
    - l'incasso in dollari (dollars_sold)
    - la quantità venduta (units_sold)
    - le spese sostenute a fronte della vendita (dollars_cost)
  - la grana è il totale per prodotto, negozio e giorno

Tabella fatti

- I campi della tabella fatti sono partizionati in due insiemi
  - chiave (composta)
    - sono riferimenti alle chiavi primarie delle tabelle dimensione
    - stabiliscono la grana della tabella fatti
  - altri campi: misure
    - spesso chiamati fatti
    - solitamente valori numerici in un dominio continuo e additivi
- Una tabella fatti memorizza una funzione (in senso matematico) dalle dimensioni ai fatti
  - ovvero, una funzione che associa a ciascun fatto un valore per ciascuna possibile combinazione dei membri delle dimensioni

Additività dei fatti

- Un fatto è **additivo** se ha senso sommarlo rispetto a ogni possibile combinazione delle dimensioni da cui dipende
  - l'incasso in dollari è additivo perché ha senso calcolare la somma degli incassi per un certo intervallo di tempo, insieme di prodotti e insieme di negozi
    - ad esempio, in un mese, per una categoria di prodotti e per i negozi in un'area geografica
  - l'additività è una proprietà importante, perché le applicazioni del data warehouse devono solitamente combinare i fatti descritti da molti record di una tabella fatti
    - il modo più comune di combinare un insieme di fatti è di sommarli (se questo ha senso)
    - è possibile anche l'uso di altre operazioni

Semi additività e non additività

- I fatti possono essere anche
  - semi additivi
    - se ha senso sommarli solo rispetto ad alcune dimensioni
      - ad esempio, il numero di pezzi in deposito di un prodotto è sommabile rispetto alle categorie di prodotto e ai magazzini, ma non rispetto al tempo
  - non additivi
    - se non ha senso sommarli
  - può avere senso combinare fatti anche non completamente additivi mediante operazioni diverse dalla somma (ad esempio, medie pesate)

Attributi e interrogazioni

- Gli attributi delle tabelle dimensione sono il principale strumento per l'interrogazione del data warehouse
  - gli attributi delle dimensioni vengono usati per
    - selezionare un sottoinsieme dei dati di interesse
      - vincolando il valore di uno o più attributi
      - ad esempio, le vendite nel corso dell'anno 2000
    - raggruppare i dati di interesse
      - usando gli attributi come intestazioni della tabella risultato
      - ad esempio, per mostrare le vendite per ciascuna categoria di prodotto in ciascun mese

Attributi e interrogazioni

- Dati restituiti dall'interrogazione
  - somma degli incassi in dollari e delle quantità vendute
  - per ciascuna categoria di prodotto in ciascun mese
  - nel corso dell'anno 2000

(product) category	(time) month	(sum of) dollars_sold	(sum of) units_sold
Drinks	gennaio 2000	21.509,05	23.293
Drinks	febbraio 2000	19.486,93	22.216
Drinks	marzo 2000	21.986,43	23.532
Food	gennaio 2000	86.937,77	55.135
Supplies	gennaio 2000	21.554,17	13.541

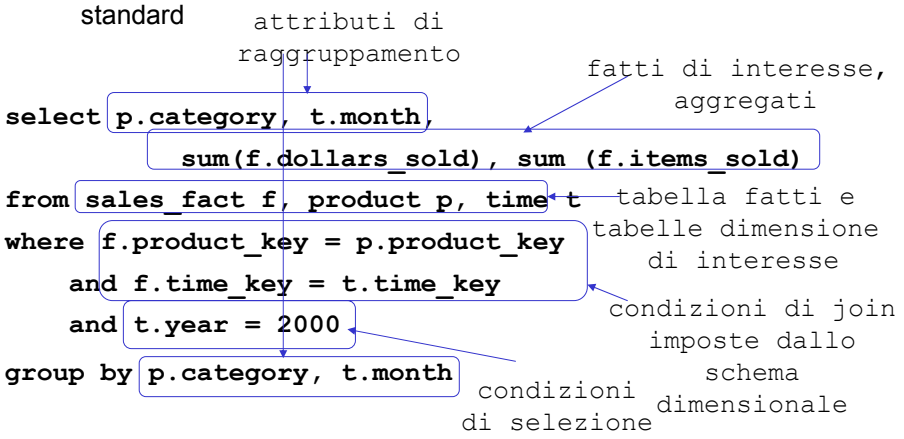
Formato delle interrogazioni

- Le interrogazione assumono solitamente il seguente formato standard

```
select p.category, t.month,
           sum(f.dollars_sold), sum (f.items_sold)
from sales_fact f, product p, time t
where f.product_key = p.product_key
    and f.time_key = t.time_key
    and t.year = 2000
group by p.category, t.month
```

Formato delle interrogazioni

- Le interrogazione assumono solitamente il seguente formato standard



Drill down

- L'operazione di drill down aggiunge dettaglio ai dati restituiti da una interrogazione
  - il drill down avviene aggiungendo un nuovo attributo nell'intestazione di una interrogazione
  - diminuisce la grana dell'aggregazione

(product) category	(time) month	(sum of) dollars_sold	(sum of) units_sold

drill down

(product) category	(time) month	(store) city	(sum of) dollars_sold	(sum of) units_sold

Drill up (roll up)

- L'operazione di drill up riduce il dettaglio dei dati restituiti da una interrogazione
  - il drill up avviene rimuovendo un attributo dall'intestazione di una interrogazione
  - aumenta la grana dell'aggregazione

(product) category	(time) month	(sum of) dollars_sold	(sum of) units_sold

drill up

(product) category	(sum of) dollars_sold	(sum of) units_sold

Discussione

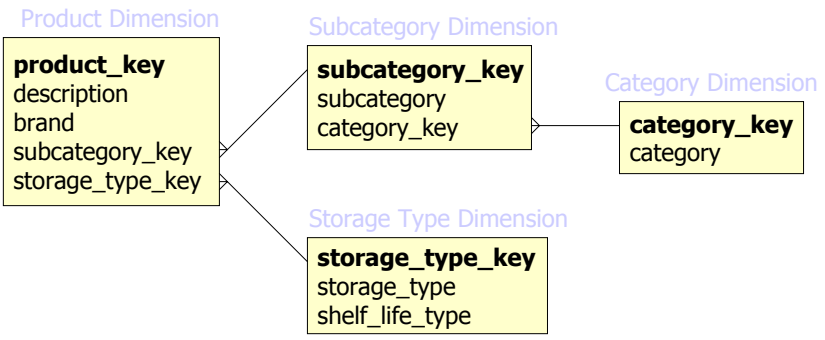
- Per il data warehouse, la modellazione dimensionale presenta dei vantaggi rispetto alla modellazione tradizionale (ER-BCNF) adottata nei sistemi operazionali
  - gli schemi dimensionali hanno una forma standardizzata e prevedibile
    - è facilmente comprensibile e rende possibile la navigazione dei dati
    - semplifica la scrittura delle applicazioni
    - ha una strategia di esecuzione efficiente
  - gli schemi dimensionali hanno una struttura simmetrica rispetto alle dimensioni
    - la progettazione può essere effettuata in modo indipendente per ciascuna dimensione
    - le interfacce utente e le strategie di esecuzione sono simmetriche

Vantaggi della modellazione dimensionale

- gli schemi dimensionali sono facilmente estendibili
  - rispetto all'introduzione di nuovi fatti
  - rispetto all'introduzione di nuovi attributi per le dimensioni
  - rispetto all'introduzione di nuove dimensioni "supplementari"
    - se ogni record della tabella fatti dipende già funzionalmente dai membri della nuova dimensione
- si presta alla gestione e materializzazione di dati aggregati
- sono state già sviluppate numerose tecniche per la descrizione di tipologie fondamentali di fatti e dimensioni

Snowflaking

- Per snowflaking di una dimensione si intende una rappresentazione "più normalizzata" di una tabella dimensione, che evidenzia delle "gerarchie di attributi"



Occupazione di memoria

- Stima dell'occupazione di memoria della base di dati dimensionale di esempio
  - Tempo: 2 anni di 365 giorni, ovvero 730 giorni
  - Negozi: 300
  - Prodotti: 30.000
  - Fatti relativi alle vendite
    - ipotizziamo un livello di sparsità del 10% delle vendite giornaliere dei prodotti nei negozi
      - ovvero, che ogni negozio vende giornalmente 3.000 diversi prodotti
    - 730 x 300 x 3000 = 630.000.000 record

Resistere allo snowflaking

- Lo snowflaking è solitamente svantaggioso
  - inutile per l'occupazione di memoria
    - ad esempio, supponiamo che la dimensione prodotto contenga 30.000 record, di circa 2.000 byte ciascuno
      - occupando quindi 60MB di memoria primaria
    - la tabella fatti contiene invece 630.000.000 record, di circa 10 byte ciascuno
      - occupando quindi 6.3GB di memoria primaria
    - le tabelle fatti sono sempre molto più grandi delle tabelle dimensione associate
      - anche riducendo l'occupazione di memoria della dimensione prodotto del 100%, l'occupazione di memoria complessiva è ridotta di meno dell'1%
  - può peggiorare decisamente le prestazioni

Processi in un data warehouse

- I processi di base in un data warehouse comprendono
  - processi nell'area di preparazione dei dati (attività "notturne")
    - estrazione, trasformazione, caricamento e indicizzazione, controllo di qualità
    - aggiornamento del data warehouse
  - processi utente (attività "diurne")
    - interrogazione
  - processi di amministrazione
    - gestione della sicurezza
    - auditing
    - backup e recovery
    - gestione del feedback

Estrazione

- L'**estrazione** è il primo passo nel transito dei dati dalle sorgenti informative al data warehouse
  - più precisamente, l'attività di estrazione riguarda
    - la comprensione e la lettura delle sorgenti informative
    - la copiatura nell'area di preparazione dei dati delle porzioni di sorgenti informative che sono necessarie al popolamento del data warehouse

Trasformazione

- I dati estratti dalle sorgenti informative, prima di essere caricati nel data warehouse, sono sottoposti a diverse **trasformazioni**
  - pulizia
    - per risolvere errori, conflitti, incompletezze
    - per riportare i dati in un formato standard
  - eliminazione di campi non significativi
  - combinazione
    - per identificare e correlare i dati associati alla rappresentazione di uno stesso oggetto in più sorgenti informative
  - creazione di chiavi
    - le chiavi usate nel data warehouse sono diverse da quelle usate nelle sorgenti informative
  - creazione di aggregati

Caricamento e controllo di qualità

- Dopo il processo di trasformazione, i dati sono organizzati per essere caricati direttamente nel data warehouse
  - il caricamento consiste nella concatenazione (e/o aggiornamento) di un insieme di record per ciascuna tabella (fatti o dimensione) del data warehouse
    - durante il caricamento il data warehouse non è solitamente disponibile per l'accesso e l'interrogazione
  - il caricamento dei dati nel data warehouse viene seguito da una verifica della correttezza delle operazioni di preparazione e caricamento, mediante un'analisi di qualità dei dati
    - se il controllo di qualità ha successo, il nuovo data warehouse è pronto per l'accesso e l'interrogazione

Aggiornamento del data warehouse

- I dati del data warehouse devono essere aggiornati, anche frequentemente
  - aggiornamenti ordinari e periodici
    - caricamento incrementale di nuovi dati nel data warehouse
  - aggiornamenti straordinari
    - correzione di dati (record e/o schemi)
    - sono aggiornamenti orientati al miglioramento della qualità complessiva dei dati

Interrogazione del data warehouse

- L'interrogazione (o analisi) è l'attività prevalente nel data warehouse
  - il data warehouse è stato creato per essere interrogato
  - il data warehouse deve essere ottimizzato per l'esecuzione di interrogazioni complesse
    - ad esempio, mediante la gestione (trasparente) di dati aggregati
  - l'interrogazione avviene mediante diversi strumenti

Processi di amministrazione

- Auditing
  - sull'origine dei dati (ad esempio, per certificarne la qualità)
  - sull'uso del data warehouse (per l'ottimizzazione del data warehouse)
- Gestione della sicurezza
- Backup e recovery
- Gestione del feedback
  - il transito principale dei dati va dalle sorgenti informative al data warehouse e dal data warehouse agli strumenti di analisi
    - dati "puliti" e risultati di analisi significativi possono transitare nella direzione opposta

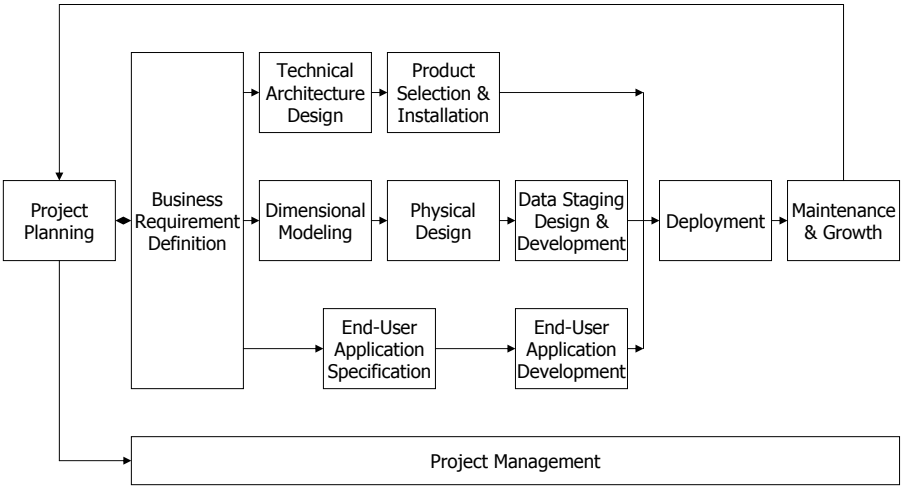
- Introduzione
  - Basi di dati integrate, sì, ma …
  - OLTP e OLAP
- Data warehousing
  - Data warehouse e data warehousing
  - Dati multidimensionali

⇨ Progettazione di data warehouse

Ciclo di vita dimensionale

- Il **ciclo di vita dimensionale (Business Dimensional Lifecycle)** è una metodologia completa di progettazione e realizzazione di data warehouse (Kimball et al.)
  - fornisce il contesto di riferimento per la progettazione e realizzazione di data warehouse dimensionali
  - mediante un insieme di attività e di relazioni tra attività

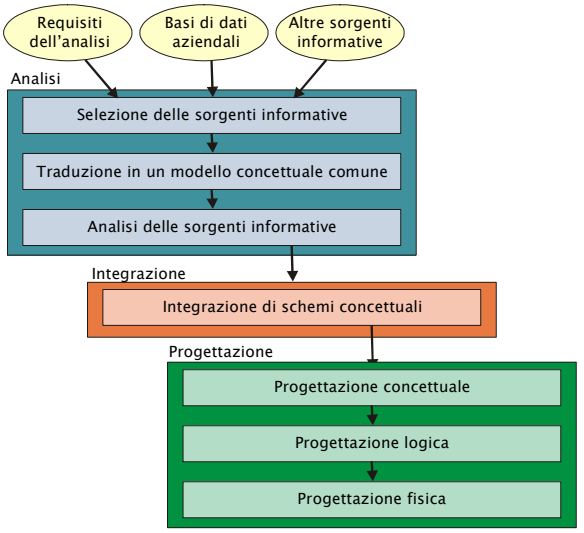
Ciclo di vita dimensionale



Fasi nel ciclo di vita dimensionale

- pianificazione del progetto
- gestione del progetto
- raccolta e analisi dei requisiti
- progettazione del data warehouse
  - progettazione dei dati
    - progettazione dimensionale, progettazione fisica, progetto della preparazione dei dati
  - progettazione tecnologica
    - progettazione dell'architettura tecnica, selezione e installazione dei prodotti
  - progettazione delle applicazioni
    - specifica delle applicazioni, sviluppo delle applicazioni
- installazione e avviamento
- manutenzione e crescita

Progettazione di data warehouse: un altro approccio



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** — 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

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 **fatti** (concetti su cui basare l'analisi), **misure** (proprietà atomiche dei fatti) e **dimensioni** (concetti su cui aggregare le misure)

Reverse engineering di schemi relazionali

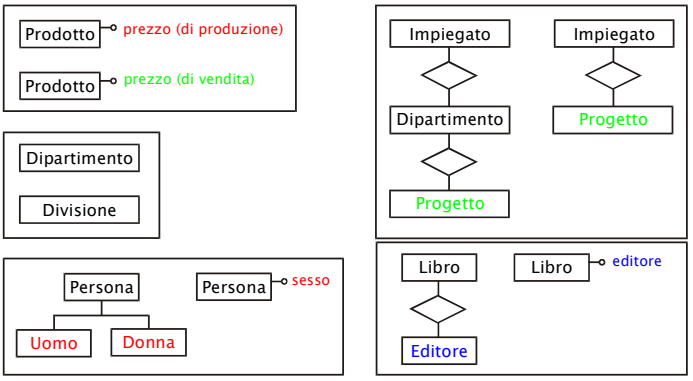
- Il **reverse engineering** è l'attività di comprensione concettuale di uno schema di dati — la rappresentazione di uno schema relazionale in un modello concettuale
- Uno schema ER è più espressivo di uno schema relazionale — è necessario conoscere la realtà di interesse per recuperare la conoscenza persa nella fase di progettazione logica
- Il reverse engineering di schemi relazionali è svolto in modo semiautomatico dagli strumenti di progettazione CASE

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** che rappresenta l'intero patrimonio informativo aziendale
- 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

Integrazione di schemi di sorgenti informative

- Orientata all'analisi e risoluzione di **conflitti** tra schemi — rappresentazioni diverse di uno stesso concetto



Integrazione di sorgenti informative

- L'integrazione di sorgenti informative è guidata da quella dei loro schemi — ma è necessario risolvere anche i conflitti relativi alla codifica delle informazioni
  - un attributo "sesso" può essere rappresentato
    - con un carattere — M/F
    - con una cifra — 0/1
    - implicitamente nel codice fiscale
    - non essere rappresentato
  - il nome e cognome di una persona
    - "Mario", "Rossi"
    - "Mario Rossi"
    - "Rossi, Mario"
    - "Rossi, M."

Integrazione di sorgenti informative

- La parte più problematica è legata alla qualità dei dati disponibili

Mario Rossi è nato il 3 ottobre 1942

Mario Rossi è nato il 10 marzo 1942

Mairo Rossi è nato il 10 marzo 1942

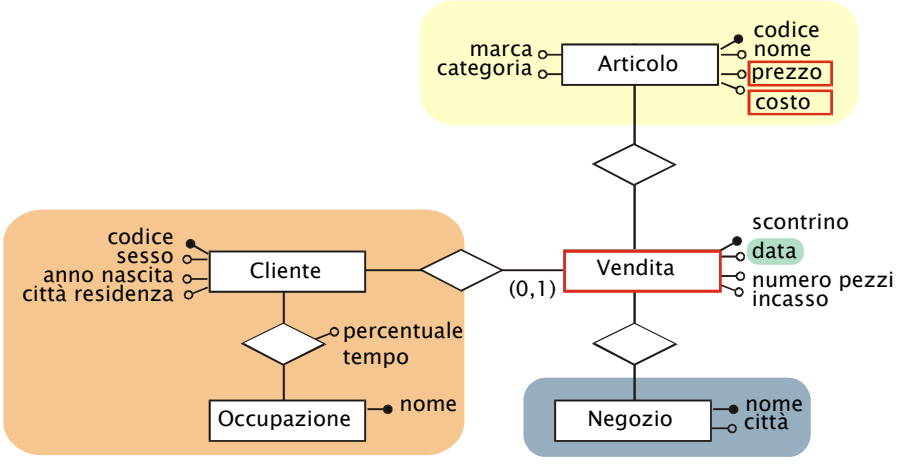
Progettazione del data warehouse

- 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

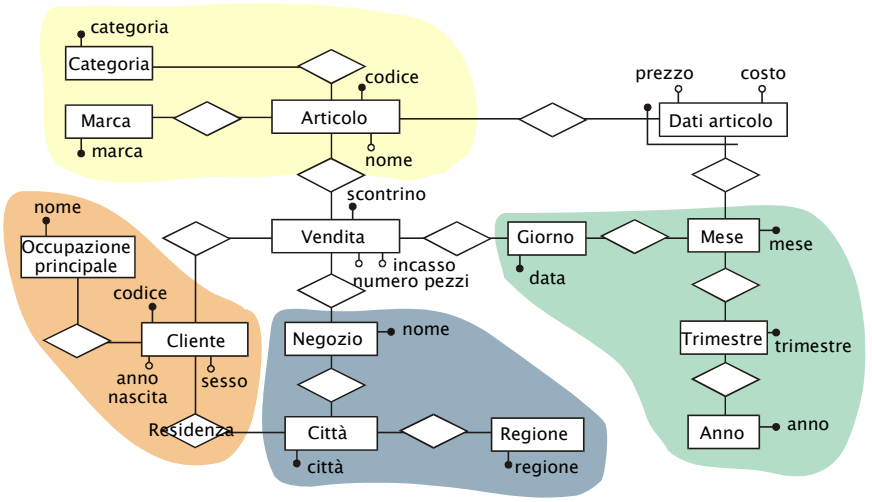
Progettazione del DW e di basi di dati multidimensionali

- 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

Identificazione di fatti e dimensioni

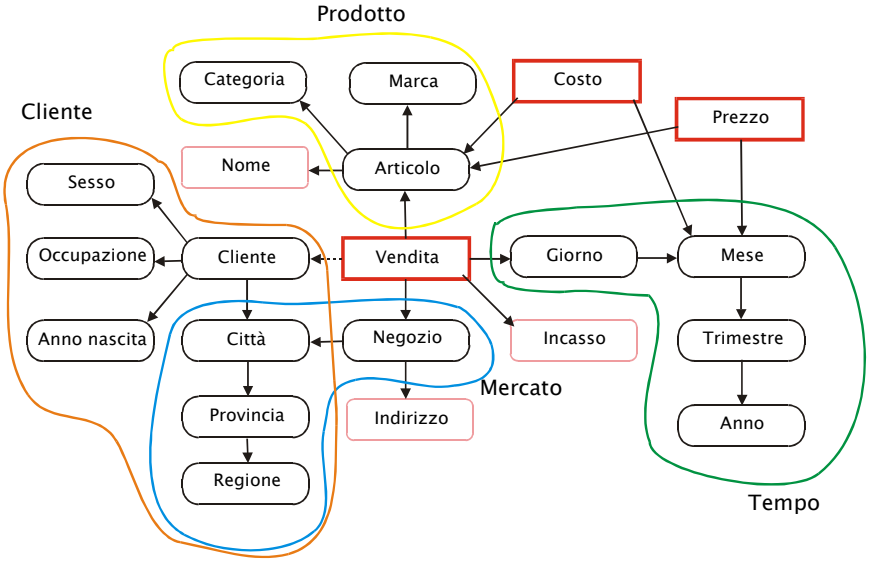


Ristrutturazione dello schema concettuale



E' lo schema concettuale del data warehouse

Derivazione di uno schema dimensionale



Progettazione logica: schema MD per Vendita, Costo e Prezzo

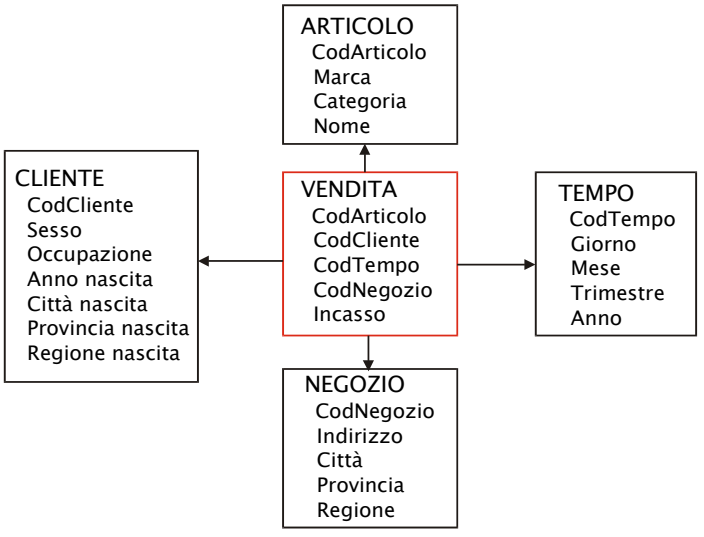
- La traduzione dallo schema dimensionale al modello logico multidimensionale è immediata:
  - dimensioni corrispondono a ipernodi del grafo,
  - livelli e descrizioni corrispondono a nodi del grafo;
  - i fatti corrispondono ai nodi fatto:

    Vendita[Data:giorno,Prod:articolo,A:cliente,Loc:negozio] : [Incasso:numero]

    Costo[Prodotto:articolo,Tempo:mese] : [Valore:numero]

    Prezzo[Prodotto:articolo,Tempo:mese] : [Valore:numero]

Progettazione fisica ROLAP: star schema per Vendita



Progettazione fisica MOLAP: matrice per Vendita

- Si costruiscono matrici a *n* dimensioni le cui celle contengono i dati.
- Le gerarchie sui livelli sono codificate in indici di accesso alle matrici