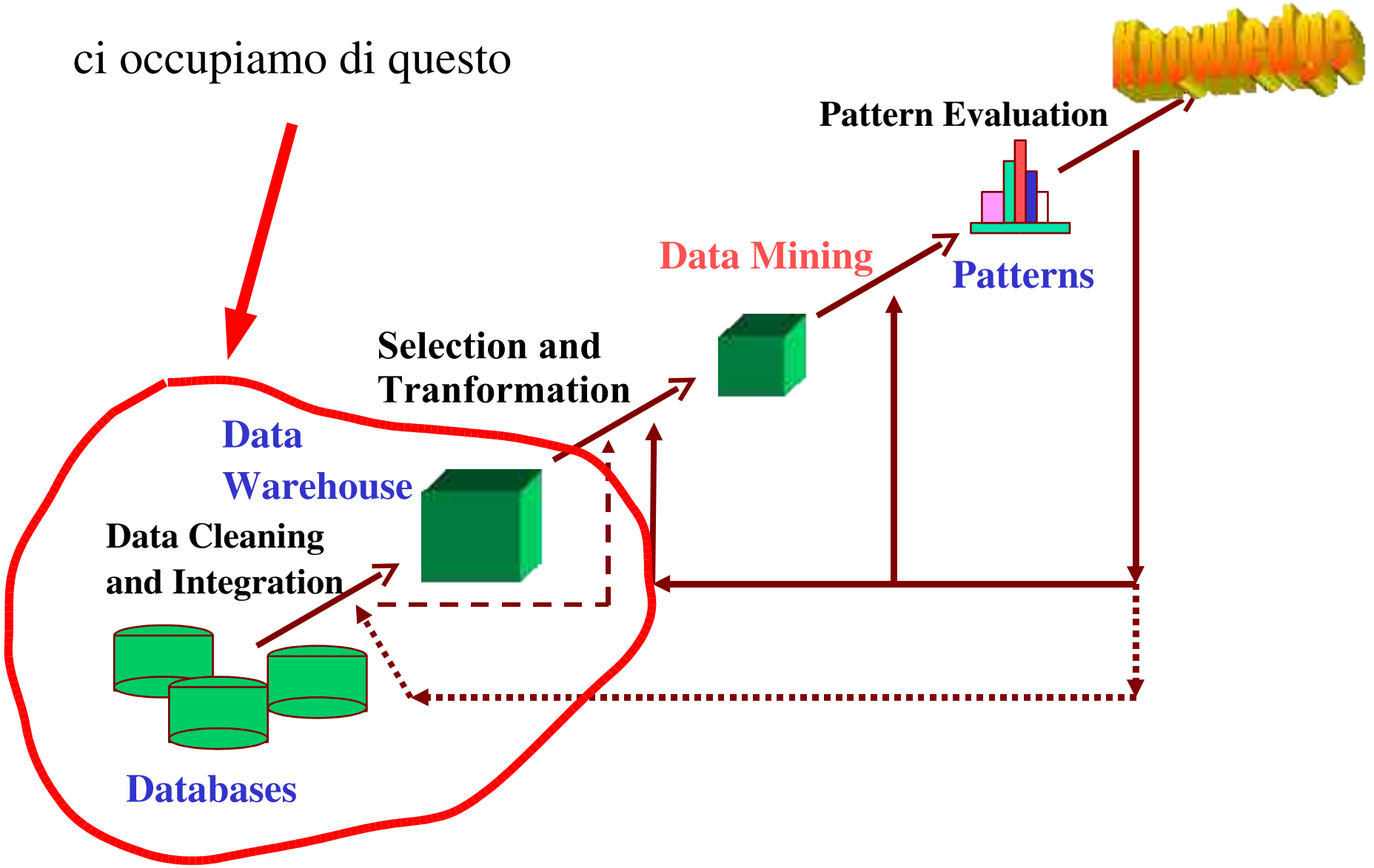
Data Warehouse e OLAP

Gianluca Amato

Corso di Laurea in Economia Informatica
Università “G. D'Annunzio”di Chieti-Pescara

Knowledge Discovery in Databases



Data Warehouse e OLAP

Cosa è un data warehouse

Un modello dei dati multidimensionale

Architettura dei wata warehouse

Dai data warehouse al data mining

Cosa è un Data Warehouse?

- Definito in molti modi diversi, mai in maniera rigorosa.
- Una raccolta organica di informazioni da più sorgenti anche eterogenee (database aziendali, database di altre aziende, internet, flat file) che
  - è mantenuta separatamente dal database principale della organizzazione;
  - serve da supporto per le attività decisionali, fornendo una serie di dati storici consistenti.
- Esempio: un data warehouse per una catena di supermercati.

Esempio di Data Warehouse

- Una catena di supermercati può avere database diversi, uno per ogni punto vendita
- Occorre metterli assieme e correggere eventuali incongruenze
  - una possibile incongruenza: il campo “settore merceologico” contiene “alimentari” per un supermercato e “generi alimentari” per un altro
- I dati sono tipicamente
  - storici: non sono quelli su cui si lavora attualmente ma l'archivio degli ultimi anni
  - aggregati: non contengono dettagli sulle singole le transazioni, ma solo la quantità di ogni prodotto venduta in un giorno

Altra definizione di Data Warehouse

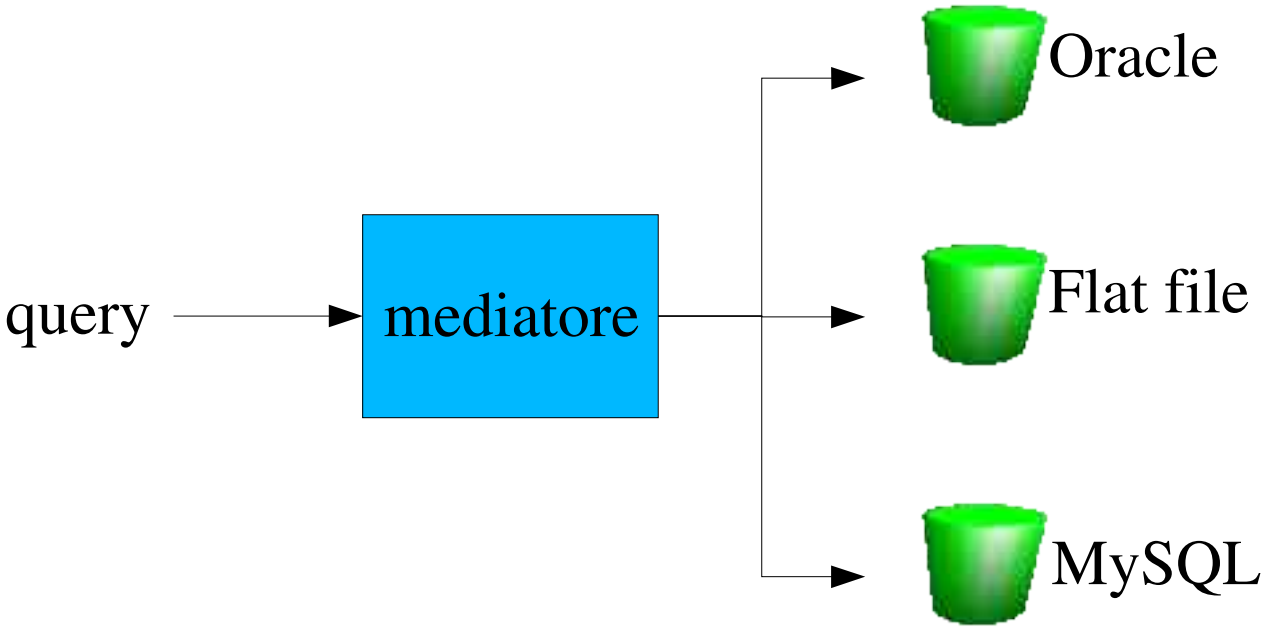
- W.H.Immon definisce un data warehose come
  - A subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making.
- Subject-oriented: un data-warehouse è organizzata attorno a degli specifici aspetti dell'azienda (clienti, vendite) utili per i processi decisionali. Gli altri dati sono trascurati.
- Integrated: come già detto, integra dati da sorgenti diverse, assicurando la consistenza dei dati finali. Le tecniche utilizzate a questo scopo sono note con il nome di data cleaning e data integration.

Altra definizione di Data Warehose

- Time-variant: i dati non forniscono solo informazioni attuali ma hanno una prospettiva storica (per esempio, dati sugli ultimi 5-10 anni)
- Nonvolatile: è un archivio fisicamente separato dalle basi di dati usate per le operazioni quotidiane. Non richiede operazioni di elaborazioni di transazioni e controlli di concorrenza. Le uniche operazioni effettuabili su un data warehouse sono il caricamento iniziale dei dati e l'accesso in lettura.

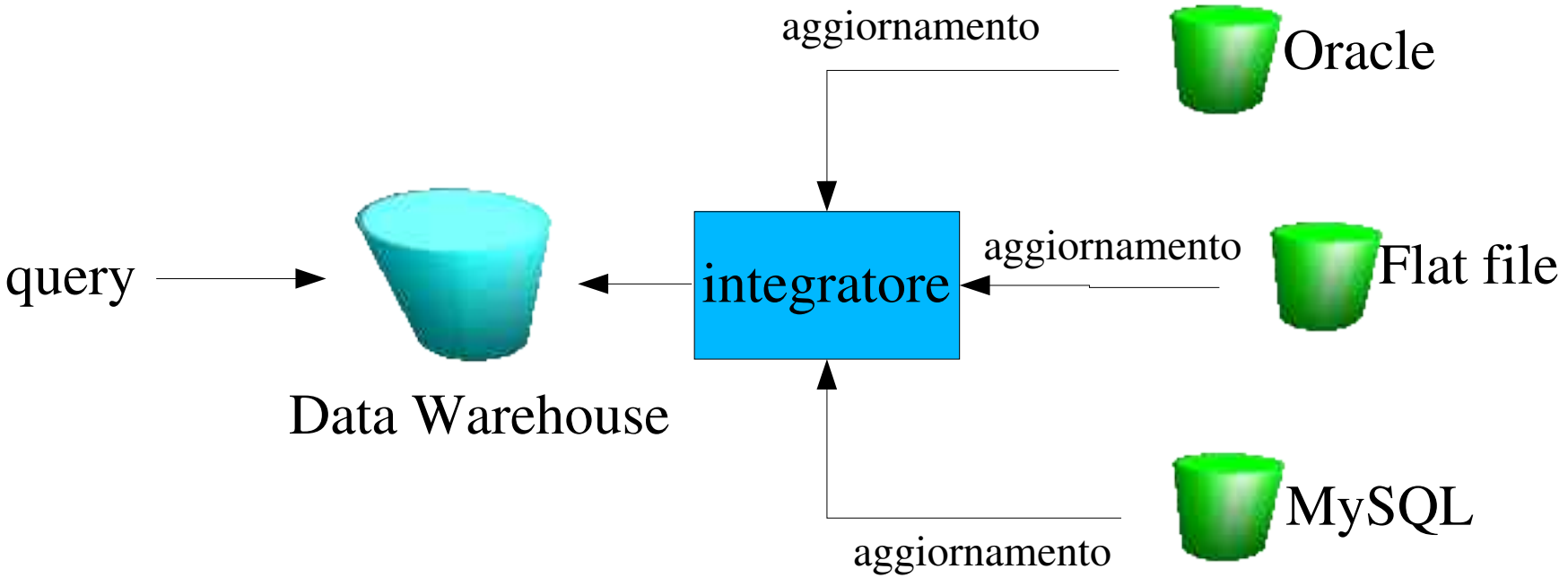
DBMS eterogenei

- A parte il problema dell'analisi dei dati a scopi decisionali, i data warehouse sono anche utilizzati semplicemente per integrare diverse basi di dati.
- Approccio tradizionale "query-driven"



DBMS eterogenei

- Approccio "update-driven"



DBMS eterogenei

- Nell'approccio query-driven, quando una query arriva al sistema intregrato, un mediatore genera delle sottoquery per i vari DBMS eterogenei, mette insieme i risultati e risponde alla query originale.
  - Il compito del mediatore può essere molto complesso
  - Le query del mediatore interferiscono con le query dirette ai singoli database.
- Nell'approccio update-driven, l'informazione è integrata in anticipo.
  - Non c'è interferenza tra query al data-warehouse e query ai singoli database.

Data Warehouse, OLAP, OLTP

- I database tradizionali vengono spesso chiamati sistemi OLTP (on-line transaction processing).
  - La loro funzione è eseguire le operazioni giornaliere: modifica dei dati e semplici operazioni di lettura.
- Un data-warehouse, invece, è il cuore di un sistema OLAP (on-line analytical processing).
  - La loro funzione è fornire supporto a operazioni di analisi dei dati e a processi decisionali.

Differenza tra OLTP e OLAP

- OLTP: orientati al cliente e adoperati da impiegati o da clienti stessi dell'organizzazione. OLAP: orientati al marketing e utilizzati dai manager, analisisti dei dati, etc..
- OLTP: contengono dati che sono spesso troppo dettagliati per essere utili a fini decisionali. OLAP: i dati sono spesso riassunti ed aggregati.
- OLTP: viene sviluppato partendo da un diagramma ER. OLAP: diagrammi a stella o a fiocco di neve.
- OLTP: dati correnti. OLAP: dati storici.
- OLTP: accessi corti e da trattare in maniera atomica, che richiedono controllo della concorrenza. OLAP: operazioni di query in sola lettura ma molto complesse.

Sommario differenze tra OLTP e OLAP

	OLTP	**OLAP**
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimension al integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

Perchè i Data Warehouse ?

- Per sfruttare il meglio del mondo DBMS e deti Data Warehouse:
  - DBMS: orientati a sistemi OLTP, transazioni, concorrenza, indicizzazione, recupero da danni.
  - Warehouse: orientati a sistemi OLAP, query complesse, vista multidimensionale, dati integrati e consolidati.

Data Warehouse e OLAP

Cosa è un data warehouse

Un modello dei dati multidimensionale

Architettura dei data warehouse

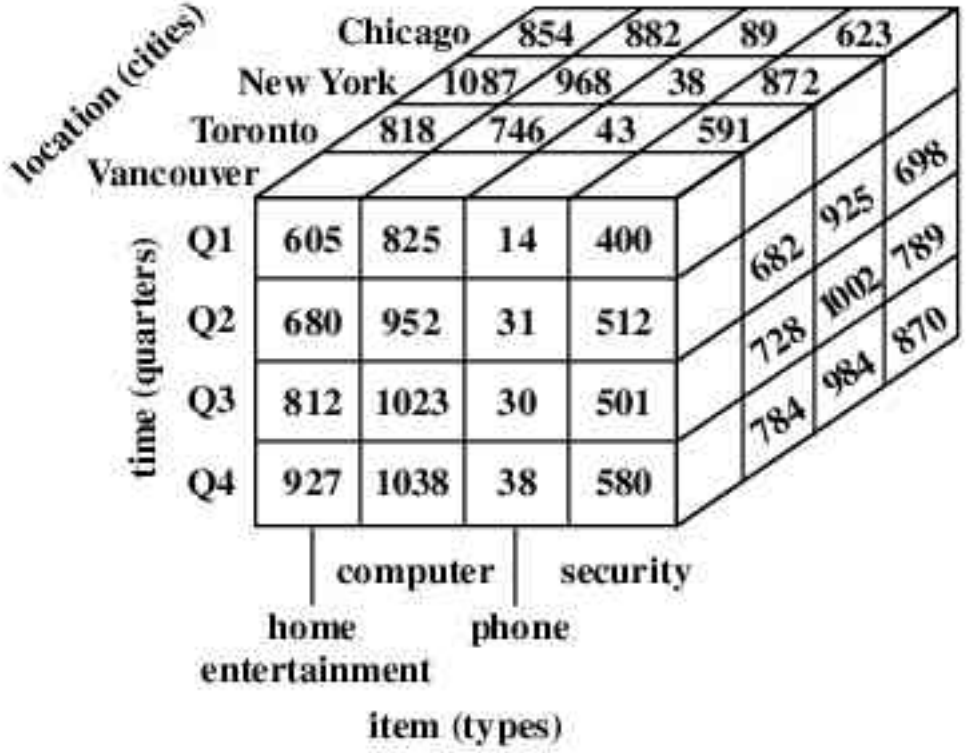
Dai data warehouse al data mining

Modello multidimensionale

- Un data warehouse è basato su un modello di dati multidimensionale. I dati sono visti sotto forma di ipercubi.
    - La AllElectronics può creare un warehouse "vendite" per registrare le vendite dell'azienda in base alle dimensioni tempo, oggetto, filiale e località.
- Le dimensioni del cubo sono le entità rispetto alle quali una organizzazione vuole mantenere i propri dati.
- In ogni posizione del cubo viene inserito un fatto, ovvero la misura numerica della quantità che si vuole analizzare.
  - "Unità di prodotto vendute" e "Euro venduti" sono esempi di fatti.

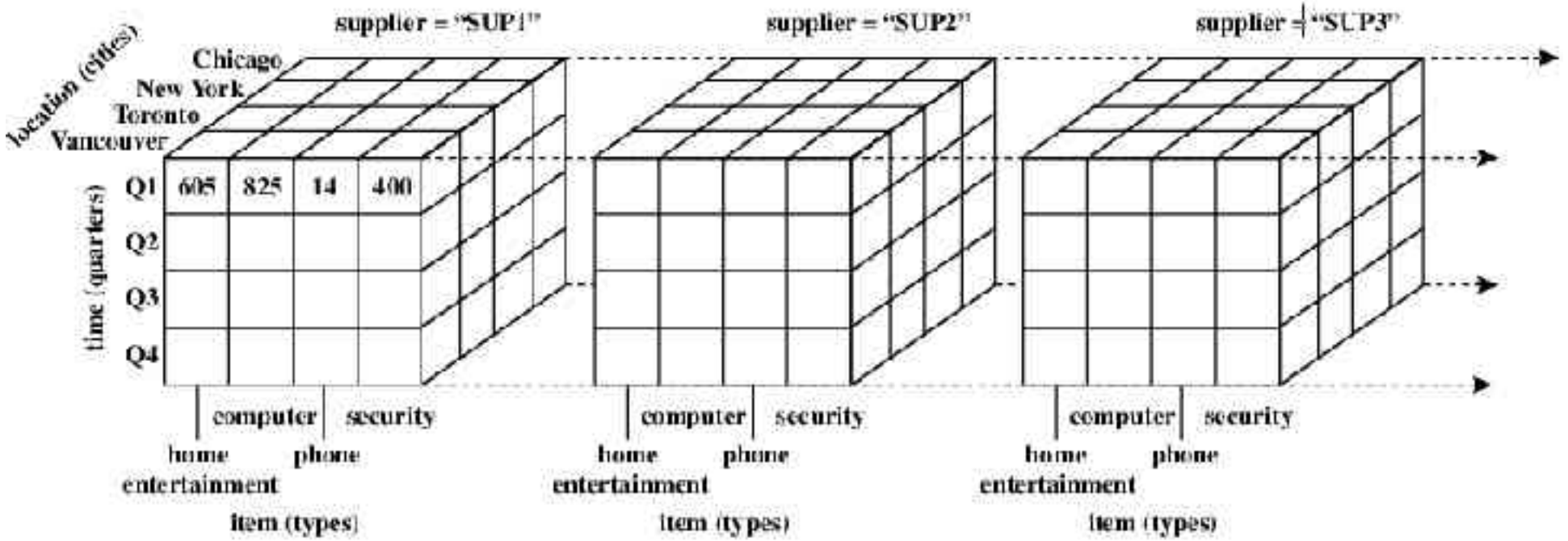
Esempio di cubo di dati (1)

Una rappresentazione 3D delle vendite della AllElectronics, sulla base delle dimensioni time, item, location



Esempio di cubo di dati (2)

Una rappresentazione 4D delle vendite della AllElectronics, sulla base delle dimensioni time, item, location, supplier

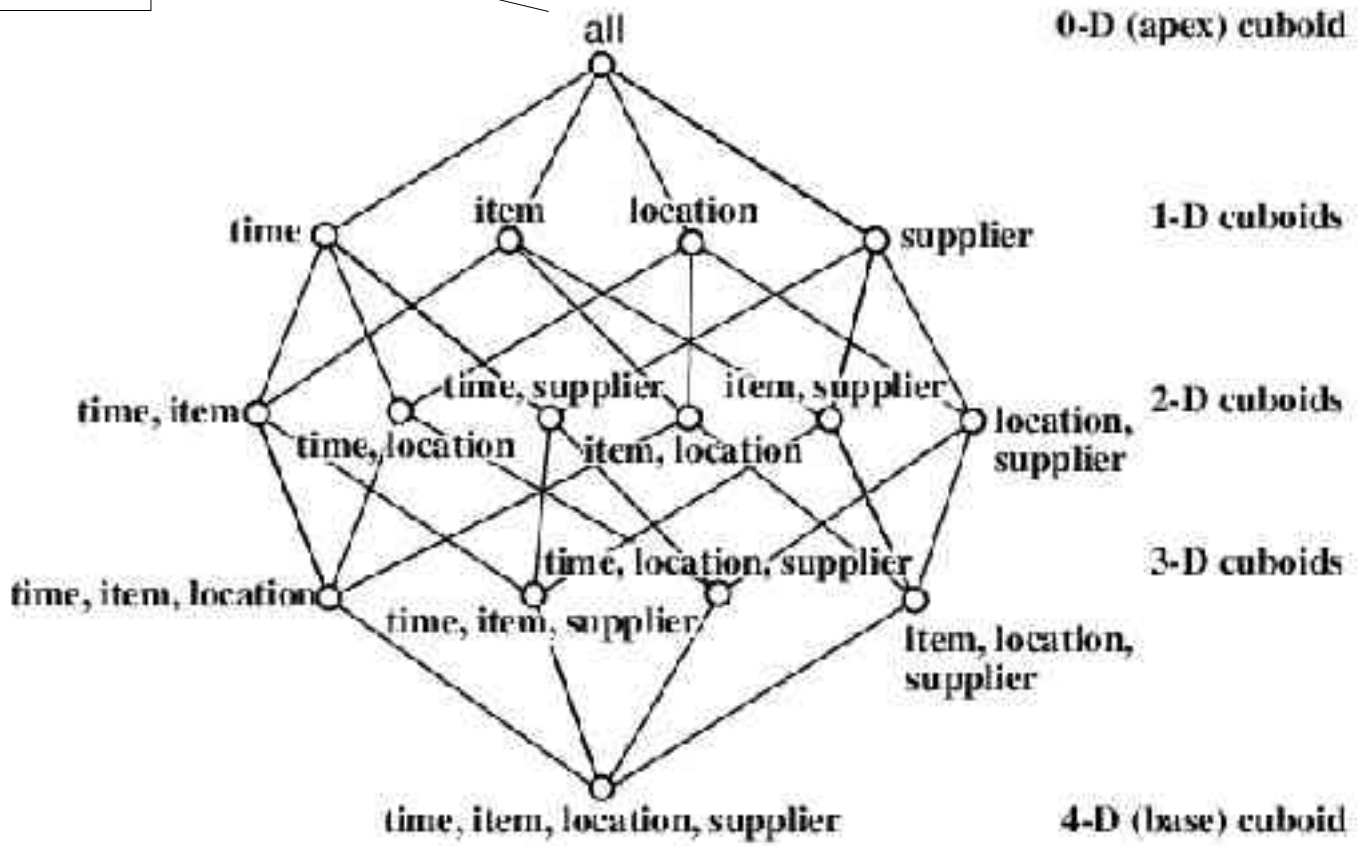


Cuboidi e data cube

- Nella letteratura sulle data warehouse, ognuno dei cubi n-dimensionali è chiamato cuboide.
- Si hanno cuboidi diversi a seconda delle dimensioni che vengono scelte e del livello di dettaglio di ogni dimensione
  - per la dimensione time si può scegliere come livello di dettaglio un quadrimestre (come fatto nei lucidi precedenti), ma anche un singolo mese, o un semestre.
- L'insieme di tutti i cuboidi viene chiamato data cube.

Reticolo dei cuboidi

cuboide apice



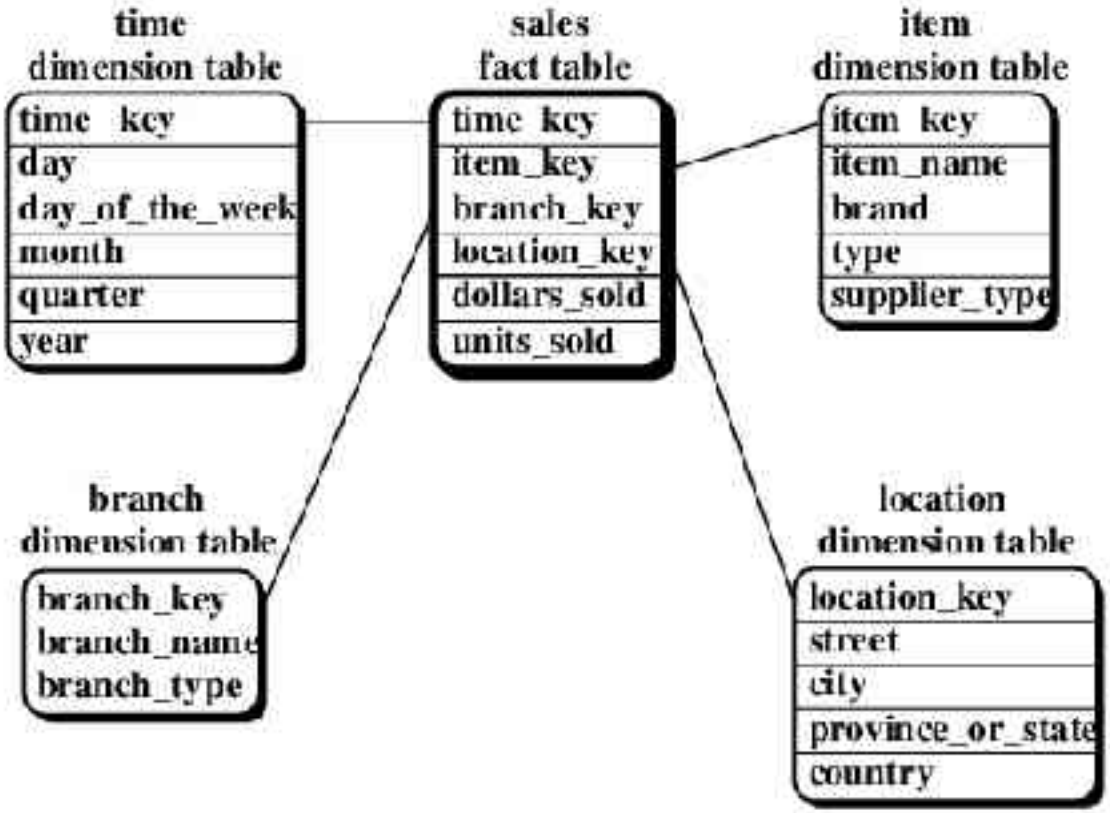
cuboide base

Schemi per database multidimensionali

- Un database per applicazioni OLTP è sviluppato a partire da una diagramma ER
- Per i data warehouse si utilizzano modelli alternativi: schemi a stella, a fiocco di neve e a galassia.
- Ogni dimensione può avere una tabella delle dimensioni associata, che descrive gli attributi di cui è composta.
  - La dimensione oggetto può contenere gli attributi nome, marca, tipo.
- L'argomento centrale del warehouse è rappresentato da una tabella dei fatti: un fatto è una misura numerica.
  - "Unità di prodotto vendute" e "Euro venduti" sono esempi di fatti.

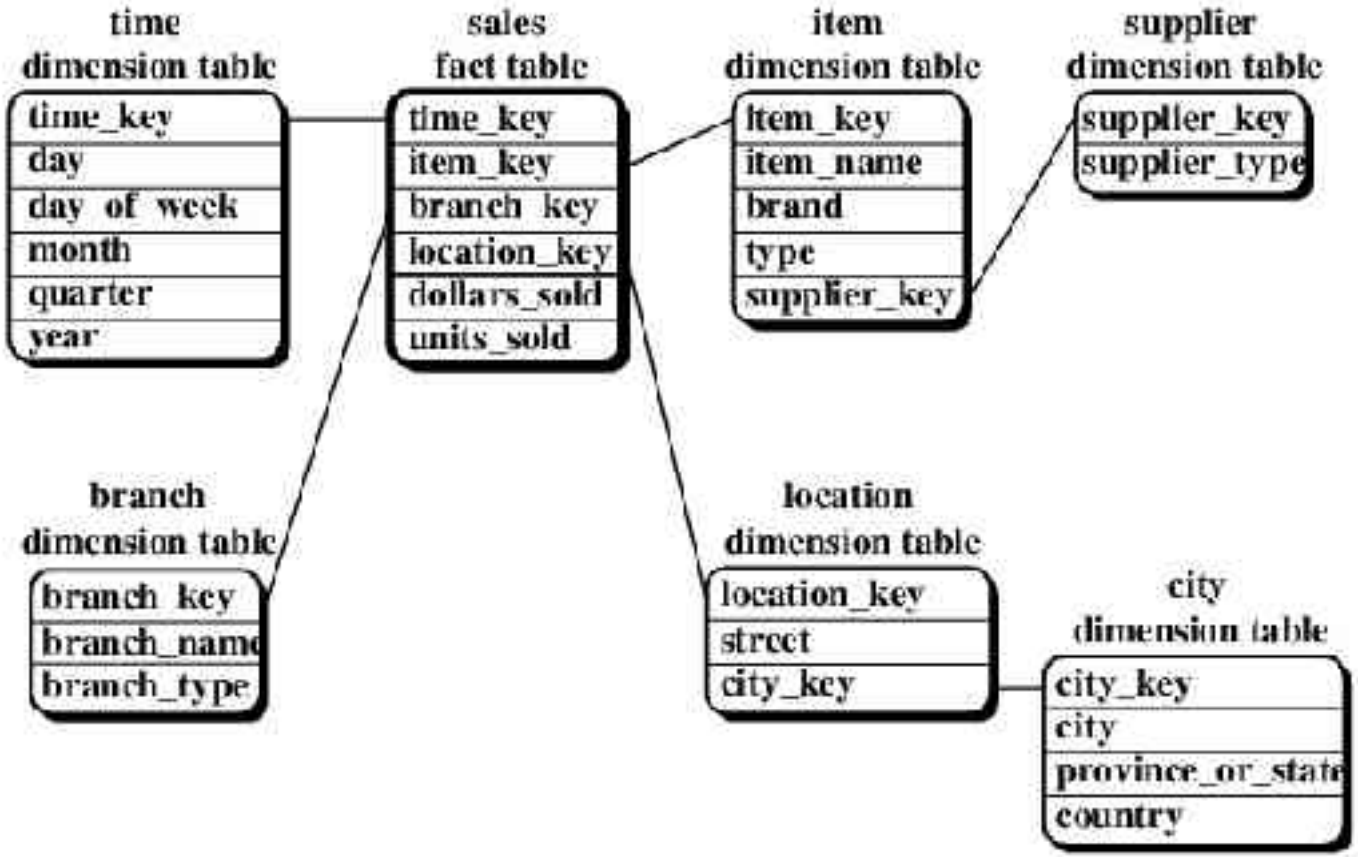
Schema a stella

- Nello schema a stella abbiamo una tabella dei fatti e varie tabelle delle dimensioni.
- La tabella dei fatti contiene le chiavi esterne per le tabelle delle dimensioni.
- Le tabelle delle dimensioni non sono normalizzate.



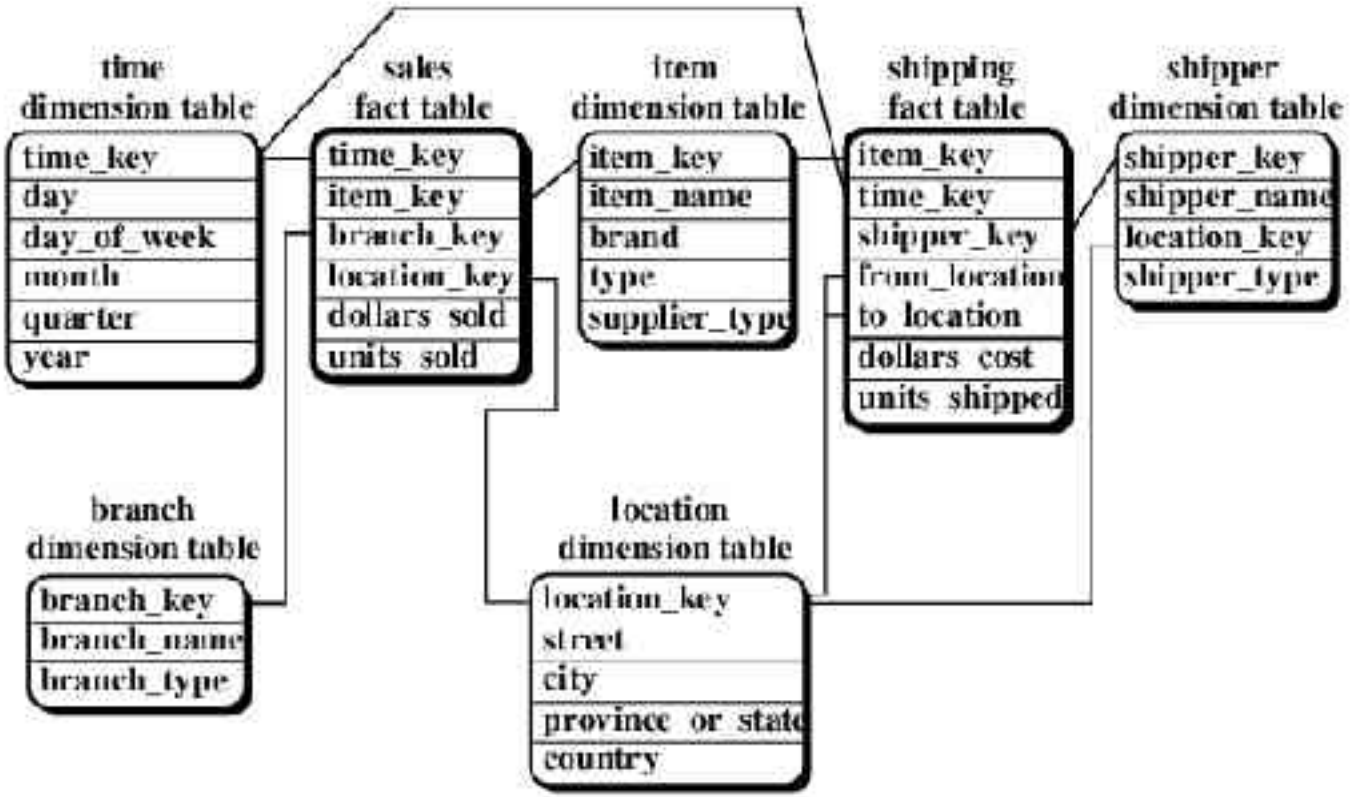
Schema a fiocco di neve

- Le tabelle delle dimensioni sono normalizzate
- Si risparmia spazio ma è meno efficiente perché le query richiedono più join per essere eseguite.



Schema a galassia

- Detto anche a costellazione di fatti (fact constellation schema)
- Caratterizzato da varie tabelle di fatti che condividono le tabelle delle dimensioni.



Quale schema scegliere?

- Si fa spesso distinzione tra data warehouse e data mart.
  - un data warehouse raccoglie informazioni su tutti gli aspetti di una organizzazione: clienti, vendite, personale, etc..
  - un data mart è un sottoinsieme del data warehouse focalizzato su un singolo aspetto (ad esempio le vendite) e gestito da un singolo dipartimento.
- data warehose => fact constellation schema
- data mart => star schema

Tipi di misura (1)

- Una misura è una funzione numerica che può essere calcolata per ogni punto di un cuboide, aggregando i dati corrispondenti alle coppie dimensione-valore che definiscono quel punto.
- Ci sono tre tipi di misure:
  - distributive: quando il valore per un insieme di dati può essere ricavato partizionando l'insieme in n-insiemi più piccoli, applicando la funzione misura a questi ultimi, e ricostituendo poi il risultato finale.
    - sum(), count(), min(), max().
    - infatti sum(v_1,...v_n)=sum(sum(v_1,vn/2), sum($v_{n/2+1}$,v_n))

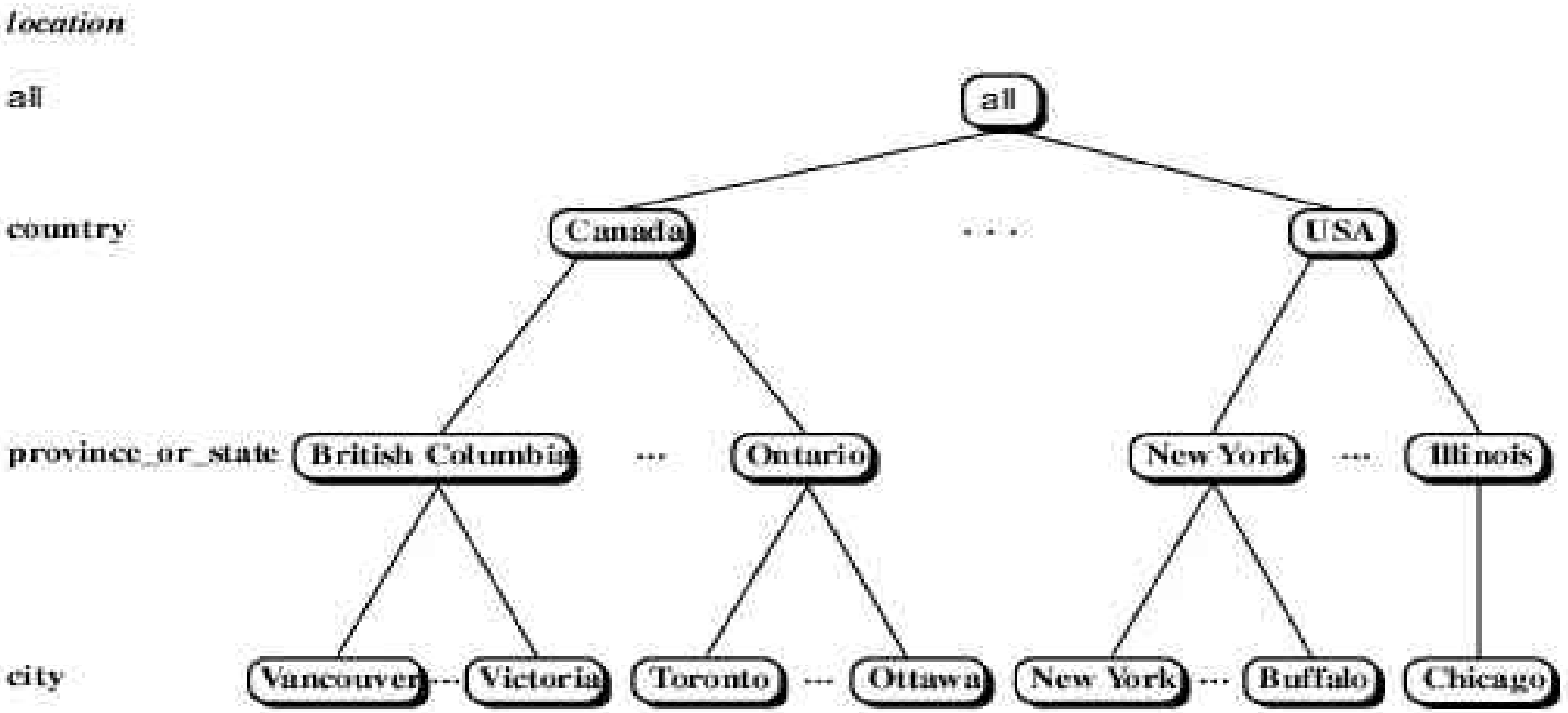
Tipi di misura (2)

- algebriche: se può essere calcolata con una funzione algebrica con M argomenti (M intero limitato), ognuno dei quali ottenuto applicando una misura distributiva.
  - media() è algebrica, in quanto media()=sum()/count() con sum() e count() entrambe distributive.
- olistiche: quando non esiste un limite costante alla dimensione di memoria necessaria per descrivere un sotto-aggregato.
  - media(), moda() sono funzioni olistiche

- Le funzioni olistiche sono difficoltose da calcolare. Esistono metodi di approssimazione.

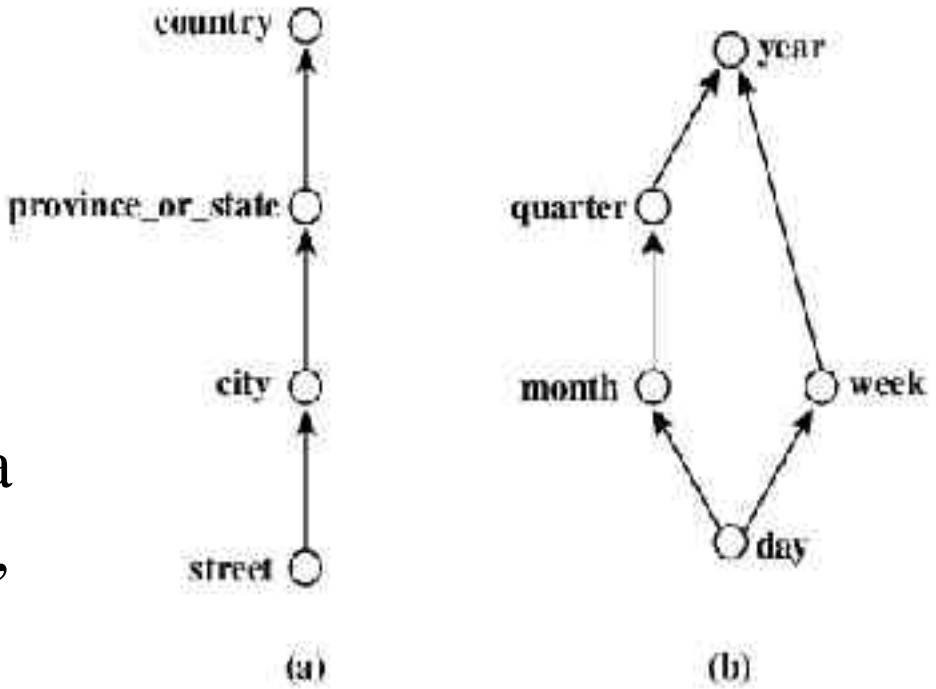
Gerarchie di Concetti (1)

- Una gerarchia di concetti (concept hierarchy) è un insieme di associazioni tra concetti concreti e concetti più astratti che viene associata ad una dimensione.



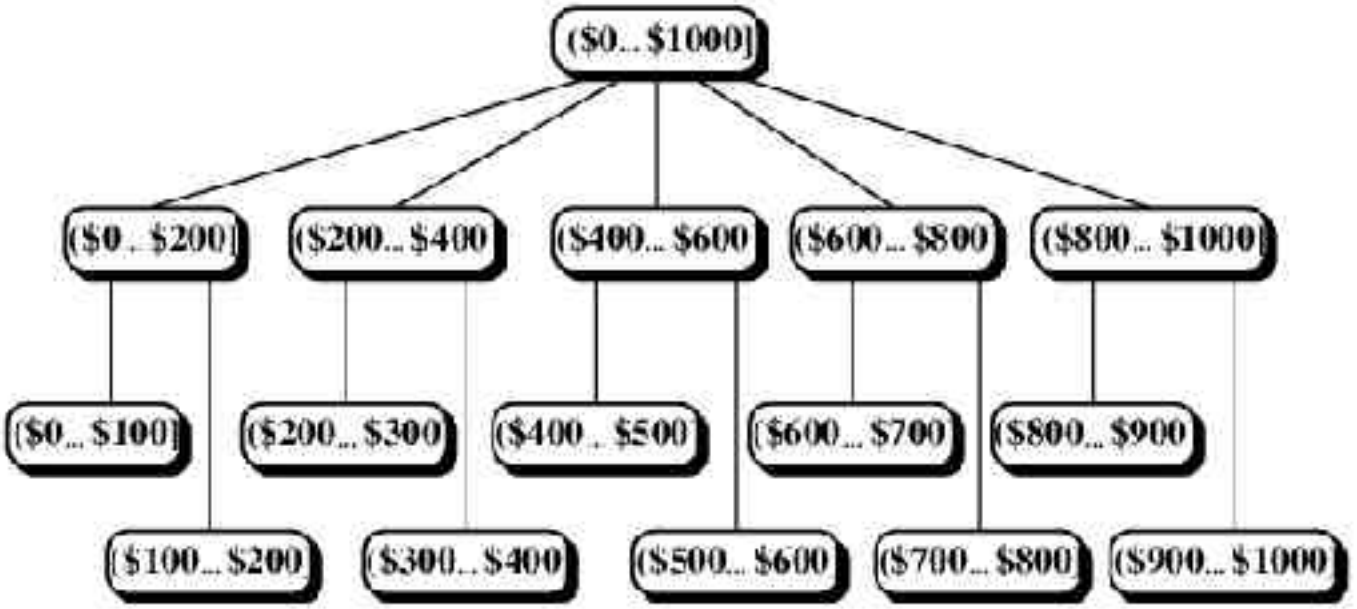
Gerarchie di Concetti (2)

- Molte gerarchie di concetti sono implicite dallo schema di database.
  - la dimensione location è descritta dagli attributi street, city, province, country.
  - la dimensione time è descritta da day, week, month, quarter, year.
- Gli attributi di una dimensione possono essere ordinati dal più concreto al più generale.
- Si ottiene una schema hierarchy



Gerarchie di Concetti (3)

- Le gerarchie di concetti possono anche essere ottenute discretizzando o raggruppando i valori di base di una data dimensione.
- Si parla di set-grouping hierarchy.



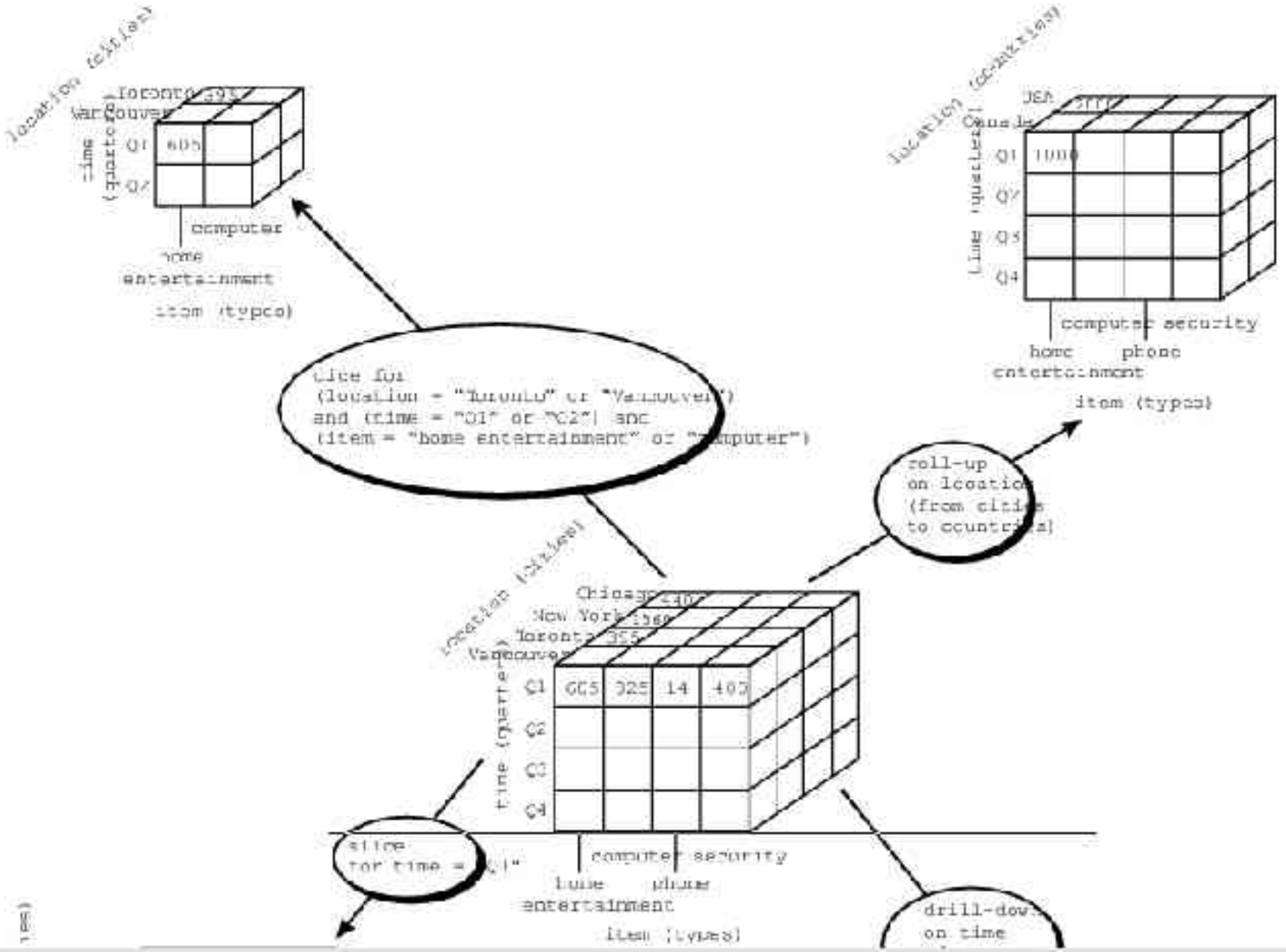
Gerarchie di Concetti (4)

- Le gerarchie di concetti possono:
  - essere fornite manualmente dall'utente o da un esperto del dominio di applicazione;
  - essere generate automaticamente sulla base di analisi statistiche.

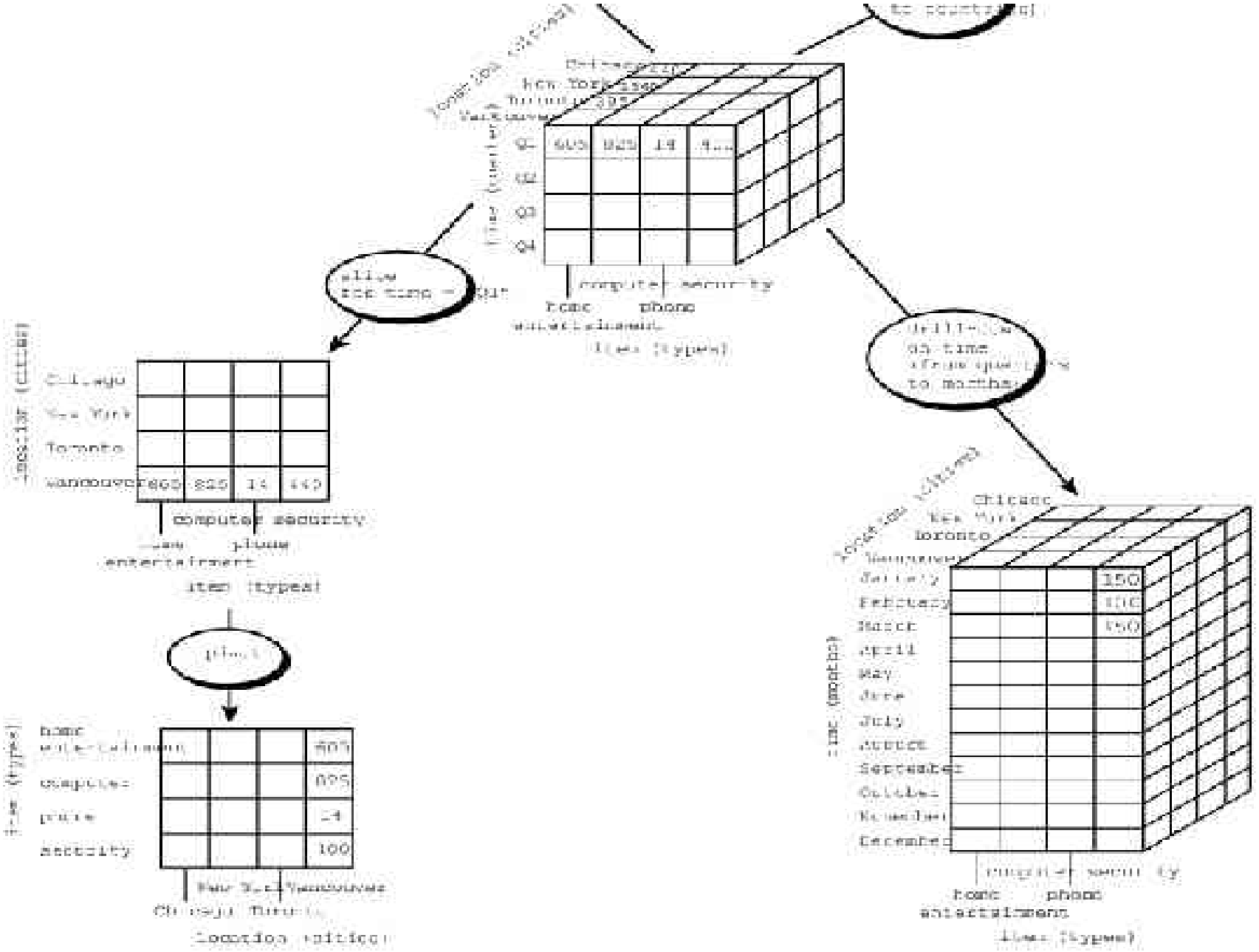
Operazioni sui cuboidi (1)

- I sistemi OLAP basati sul modello multidimensionale dei mettono a disposizione una serie di operazioni standard sui cuboidi.
- Roll-up (drill-up): esegue delle aggregazioni risalendo una gerarchia dei concetti o eliminando una dimensione.
- Drill-down: è l'inverso del roll-up, si sposta da dati meno dettagliati a dato più dettaglia introducendo una nuova dimensione o scendendo in una gerarchia dei concetti.
- Slice: esegue una selezione su una dimensione.
- Dice: esegue una selezione su una o più dimensioni.
- Pivot: ruota gli assi in un cuboide, lasciando inalterati i dati.

Operazioni sui cuboidi (2)

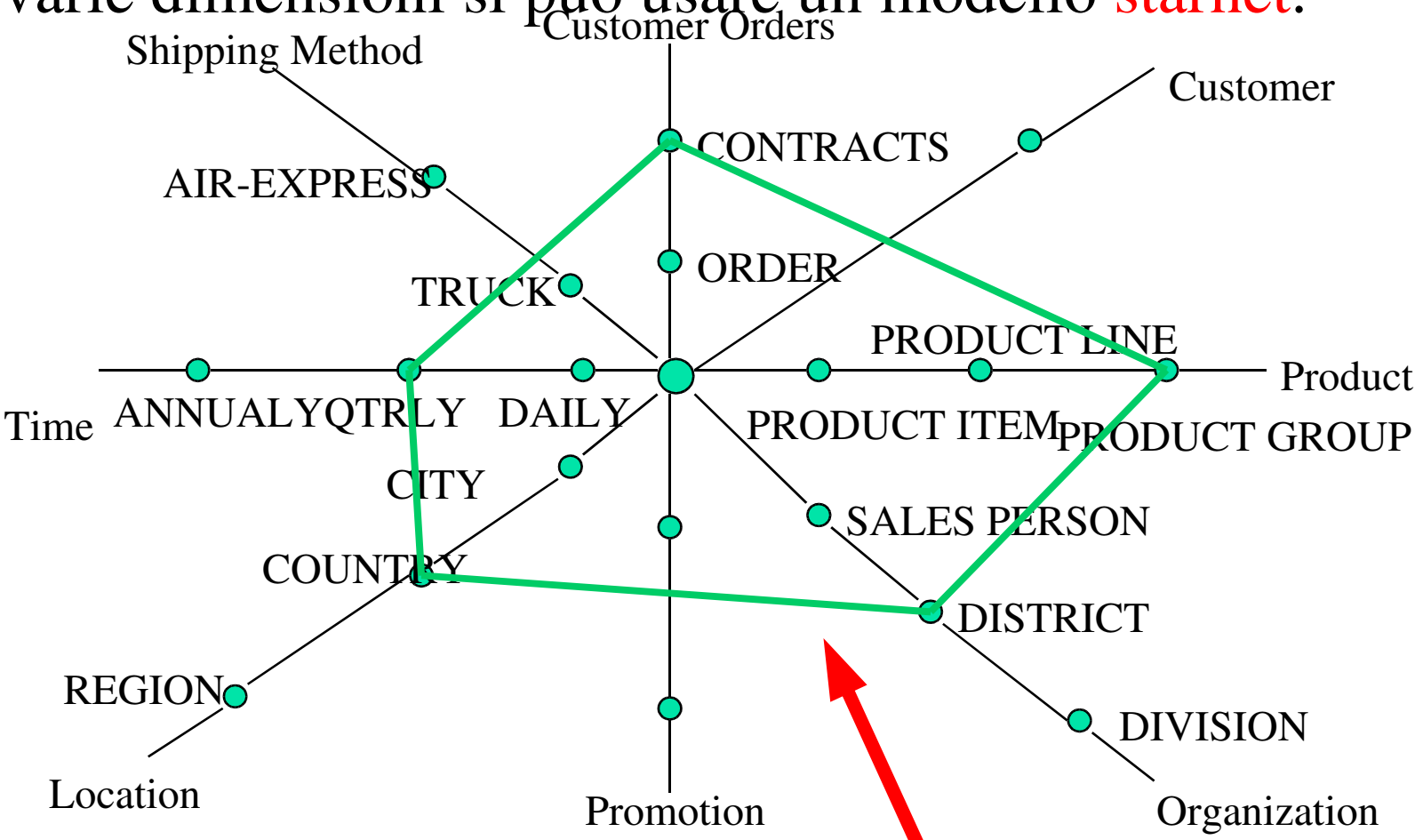


Operazioni sui cuboidi (3)



Modello starnet

- Per visualizzare i livelli di granularità disponibili nelle varie dimensioni si può usare un modello starnet.



- I cerchi nella starnet sono detti footprint.

Data Warehouse e OLAP

Cosa è un data warehouse
Un modello dei dati multidimensionale
Architettura dei data warehouse
Dai data warehouse al data mining

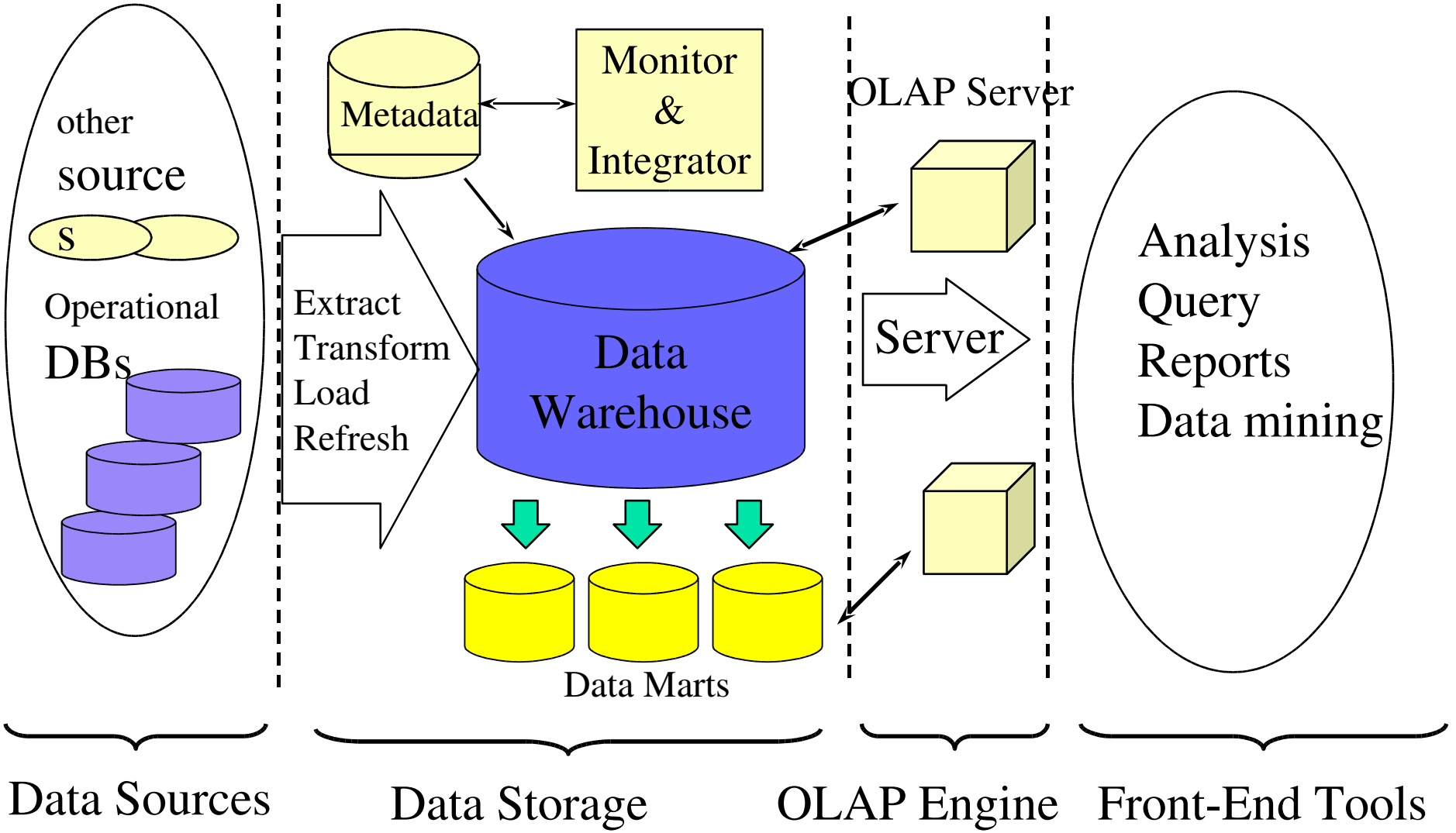
Lo sviluppo di un data warehouse

- Si può seguire un approccio top-down, bottom-up o misto
  - top-down: si inizia con la pianificazione della struttura generale e poi si passa alla implementazione di tutto il data warehouse. Utile se il problema è chiaro e se la tecnologia è matura
  - bottom-up: si inizia con esperimenti e prototipi che si possono mettere assieme per formare una struttura via via più complessa. Consente di avere qualcosa di funzionante da subito.
- In generale il processo di sviluppo si compone di varie fasi (le solite tipiche dell'ingegneria del software): pianificazione e studio dei requisiti, analisi del problema, progettazione del warehouse, caricamento dati e testing.

La progettazione di un data warehouse

- La progettazione si compone in generale di queste fasi:
  - scelta del processo da analizzare (vendite, ordini, ecc.)
  - scelta del livello di granularità (singole transazioni, riassunti giornalieri, etc..)
  - scelta delle dimensioni e delle gerarchie di concetti
  - scelta delle misure che popoleranno la tabella dei fatti

Architettura a più livelli



Tecnologie in un sistema OLAP

- Data Warehouse: tipicamente tecnologie tipiche di un database relazionale, ma ottimizzato per il tipo di operazioni tipiche.
- Server OLAP:
  - server ROLAP (relational OLAP): usano tecniche dei database relazionali;
  - server MOLAP (multidimensional OLAP): immagazzinano dati multidimensionali sotto forma di array. Eventualmente usano algoritmi di compressione in caso di matrici sparse.

Server OLAP (1)

- I server OLAP devono produrre cuboidi su richiesta dell'utente. Ci sono varie scelte:
  - nessuna materializzazione: i cuboidi vengono calcolati su richiesta
  - materializzazione totale: tutti i cuboidi del data cube (il reticolo dei cuboidi) sono pre-calcolati
  - materializzazione parziale: alcuni cuboidi vengono pre-calcolati, altri vengono calcolati su richiesta.
- La materializzazione totale sarebbe la più efficiente, ma spesso è impossibile perché richiede troppo memoria.
  - almeno 2^n cuboidi per n dimensioni, di più se abbiamo anche diversi livelli nella gerarchia dei concetti

Server OLAP (2)

- La materializzazione parziale è il metodo più usato:
  - quale cuboidi pre-calcolare?
    - ad esempio quelli più utilizzati
  - quando si calcola un nuovo cuboide, scegliere come cuboide di partenza quello pre-calcolato più adatto

Data Warehouse e OLAP

Cosa è un data warehouse
Un modello dei dati multidimensionale
Architettura dei data warehouse
Dai data warehouse al data mining

Applicazioni di un data warehouse

- Generazione di report
  - richiede supporto per interrogazioni standard in SQL, analisi statistiche di base, visualizzazione dei risultati sotto forma di grafici, tabelle, etc.
- OLAP
  - analisi multidimensionale dei dati
  - richiede supporto per la gestione dei data cube: drill-down, roll-up, slicing, etc..
- Data Mining
  - scoperta di conoscenza, ovvero di regolarità nascoste nei dati.

OLAP vs Data Mining

- Con i sistemi OLAP è possibile scoprire regolarità nei dati
  - in particolare, l'attività di data mining che abbiamo chiamato "Concept Description" è realizzabile con sistemi OLAP.
- Però:
  - i sistemi di data mining consentono altri tipi di analisi come classificazione, clustering, scoperta di regole associative
  - i sistemi OLAP aiutano l'analisi dei dati, mentre i sistemi di data mining hanno lo scopo di automatizzare l'analisi.
  - i sistemi di data mining non sono limitati ad operare su data warehouse.
    - analizzano anche dati geografici, testuali, transazionali, multimediali.

OLAP e Data Mining

- Sebbene i sitemi Data Mining non richiedano l'esistenza di un sistema OLAP sottostante, la loro integrazione è benefica:
  - Qualità dei dati
    - I data warehouse contengono dati integrati, puliti, consistenti.
  - Disponibilità di vari tool software ormai maturi che operano sui data warehouse:
    - ODBC, sistemi di reportisitca
  - Possibilità di effetturare analisi esplorative dei dati
    - Vista multidimensionale dei dati con operazioni di drilling, slicing, etc..
    - Consente di scegliere il miglior livello di granularità su cui applicare un algoritmo di data mining.

OLAM

- L'integrazione di sistemi OLAP con data mining prende il nome di OLAM (on-line analytical mining).
  - Sono di solito tool interattivi che permettono di manipolare i cuboidi con le operazioni standard dei sistemi OLAP e che consentono di richiamare funzioni di data mining su richiesta.