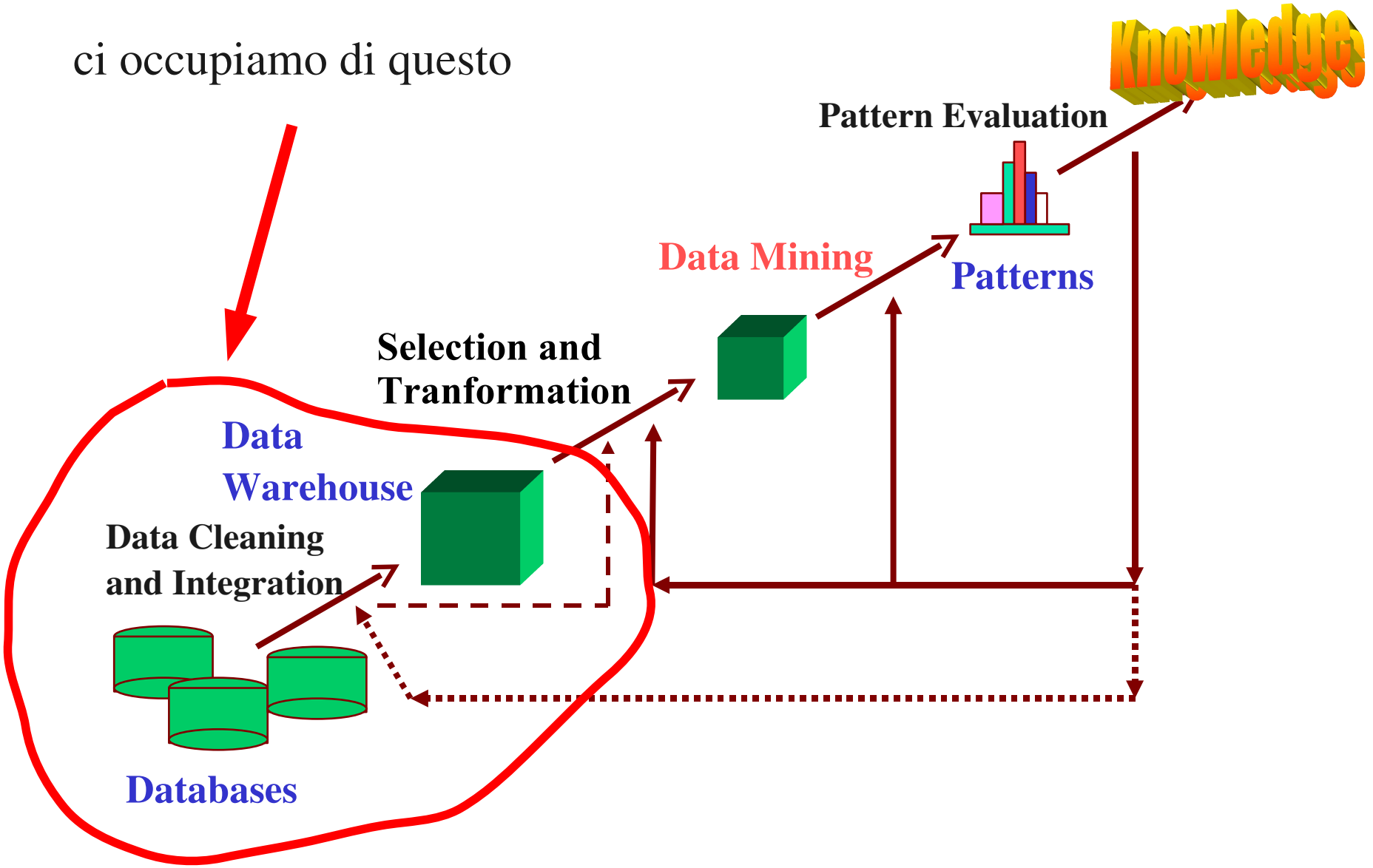
Data Warehouse e OLAP

Gianluca Amato

Corso di Laurea Specialistica in Economia Informatica
Università “G. D'Annunzio” di Chieti-Pescara
ultimo aggiornamto: 03/04/09

Knowledge Discovery in Databases



Cosa è un Data Warehouse

Cosa è un Data Warehouse (1)

- W.H.Immon, esperto progettista di data warehouse, li definisce come:
  - A subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making.
- Subject-oriented:
  - organizzato attorno a degli specifici aspetti dell'azienda (clienti, vendite, ordini, etc...)
  - focalizzato sui dati utili al processo decisionale, e non sulle operazioni giornaliere
  - contiene tipicamente dati aggregati

Cosa è un Data Warehouse (2)

- Integrated
  - integra dati da sorgenti diverse e di tipo eterogeneo (database relazionali, file di testo, database transazionali, etc...)
  - assicura la consistenza dei dati integrati utilizzando tecniche di data cleaning e data integration.
    - i dati vengono convertiti per assicurarne la consistenza e solo successivamente inseriti nel Data Warehouse
- Time-variant
  - i dati non forniscono solo informazioni attuali ma hanno una prospettiva storica (per esempio, dati sugli ultimi 5-10 anni)

Cosa è un Data Warehouse (3)

- Nonvolatile
  - è un archivio fisicamente separato dalle basi di dati usate per le operazioni quotidiane.
    - non è possibile dunque che si tratti di una “vista” all'interno del database operativo.
  - non richiede operazioni di aggiornamento continuo e dunque non necessità di supporto per la gestione delle transazioni e della concorrenza.
  - le uniche operazioni effettuabili su un data warehouse sono il caricamento iniziale dei dati e l'accesso in lettura.

Esempio di Data Warehouse

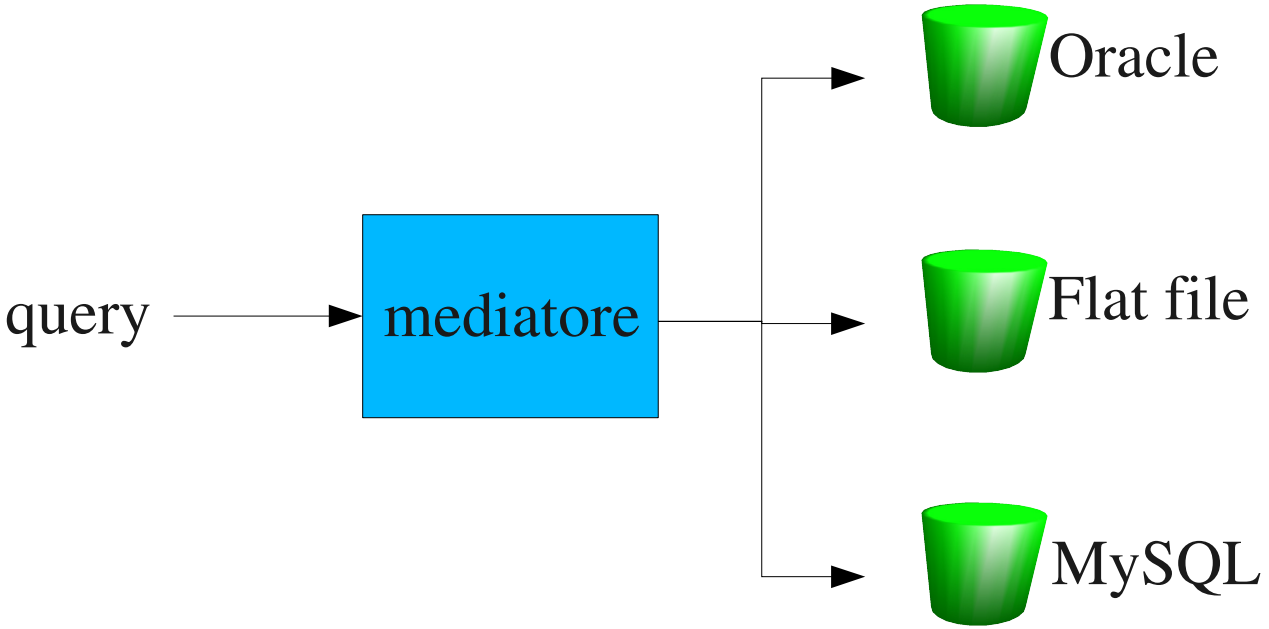
- Una catena di supermercati può avere database diversi, uno per ogni punto vendita
- Occorre metterli assieme per ottenere il data warehouse di tutte le vendite della catena
  - una possibile incongruenza: il campo “settore merceologico” contiene “alimentari” per un supermercato e “generi alimentari” per un altro

Una definizione in italiano

- Volendo una definizione compatta di Data Warehouse in italiano, si potrebbe usare la seguente:
- Un data warehouse è 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.

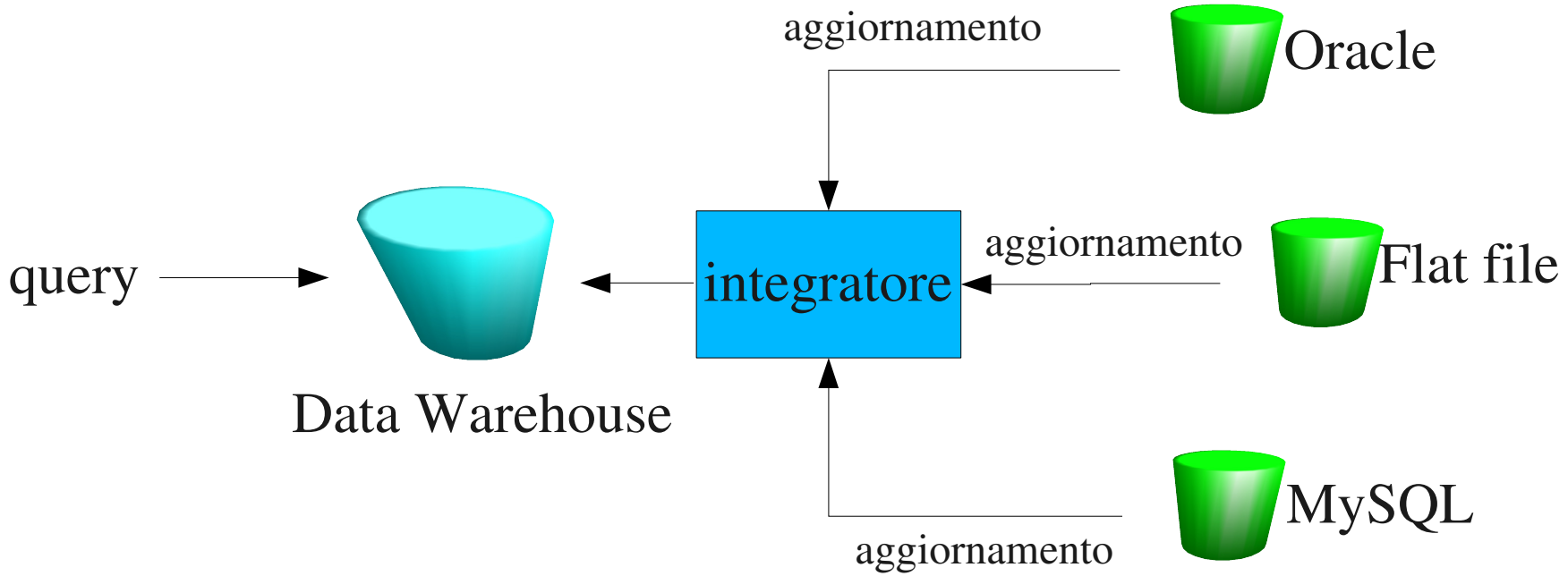
DBMS eterogenei vs Data Warehouse (1)

- 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 vs Data Warehouse (2)

- Approccio dei data warehouse "update-driven"



DBMS eterogenei vs Data Warehouse (3)

- 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.
  - I dati non sono però aggiornato fino all'ultima transazione.

Data Warehouse, OLAP, OLTP

- I sistemi informativi che si poggiano su un database tradizionale 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.

Differenze tra OLTP e OLAP (1)

- OLTP
  - orientati al cliente (adoperati da impiegati o da clienti stessi dell'organizzazione).
  - dati dettagliati, spesso eccessivamente per essere utili a fini decisionali.
  - sviluppato partendo da un diagramma ER.
  - dati correnti.
  - accessi corti e da trattare in maniera atomica, che richiedono controllo della concorrenza.

Differenze tra OLTP e OLAP (2)

- OLAP
  - orientati al marketing e utilizzati dai manager, analisti dei dati, etc..
  - dati riassunti ed aggregati.
  - sviluppato partendo da diagrammi a stella o a fiocco di neve.
  - dati storici.
  - interrogazioni in sola lettura ma molto complesse.

Differenze tra OLTP e OLAP (3)

	OLTP	***OLAP***
utente	commesso, professionista informatico	espero di analisi dei dati
funzione	operazioni giornaliere	supporto alle decisioni
sviluppo del database	orientato alla applicazione	orientato all'argomento di analisi
dati	correnti, aggiornati, dettagliati	storici, aggregati, multidimensionali
uso	ripetitivo	ad-hoc
accesso	lettura/scrutta	principalmente sin sola lettura, in particolare scansioni complete
tipo di oprtazioni	transazioni semplici	interrogazioni in lettura complesse
#record acceduti per operazione	decine	milioni
#utenti	migliaia	centinaia
dimensione del database	100MB-GB	100GB-TB
metrica per le prestazioni	transazioni al secondo	interrogazioni al secondo

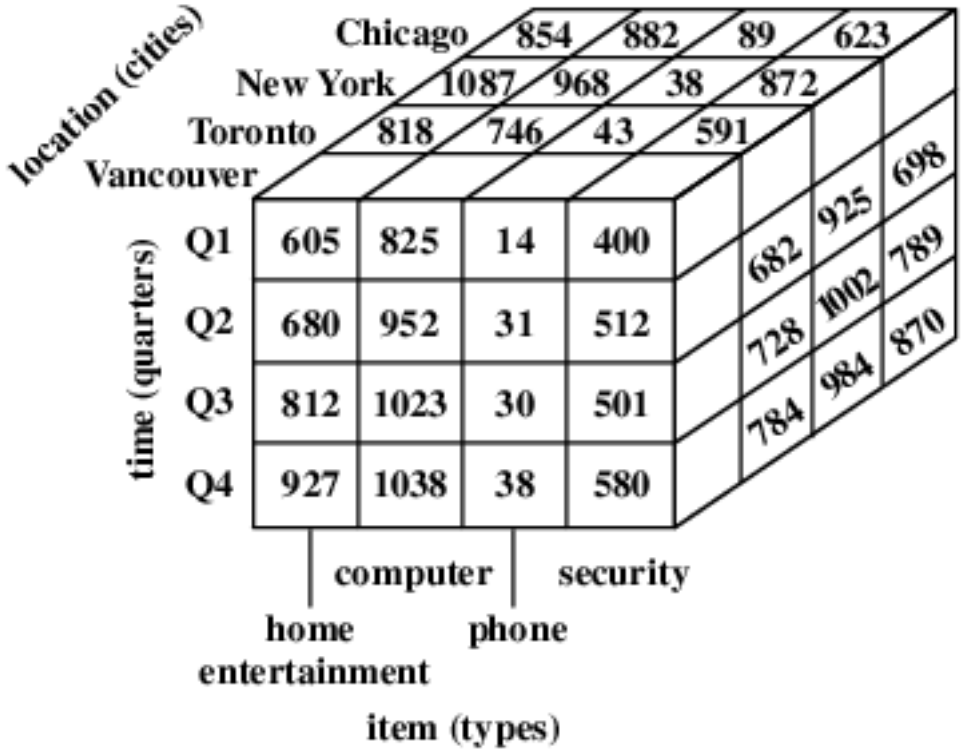
Modello dei dati multi-dimensionale

Modello multidimensionale

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

Esempio di cubo di dati (1)

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



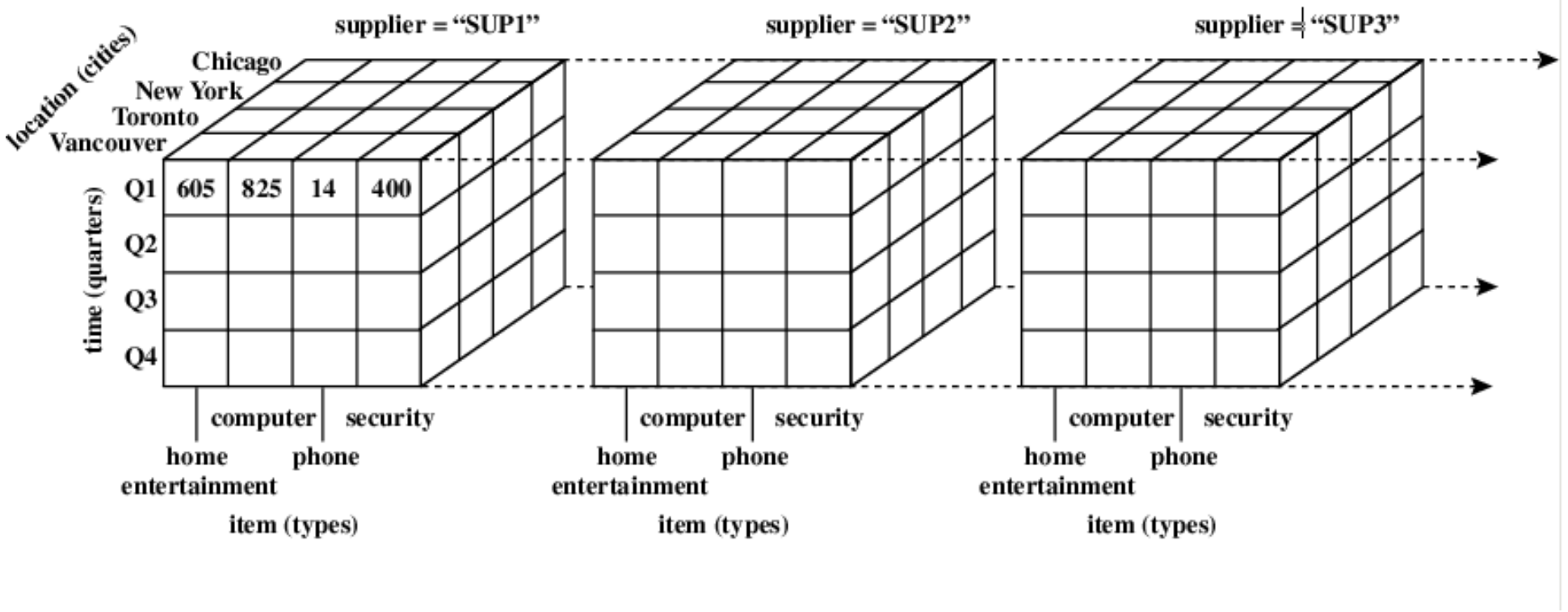
Esempio di cubo di dati (2)

- Lo stesso cubo del lucido precedente, nel classico modello relazionale, corrisponde alla tabella:

location	***time***	***item***	***units_sold***
Vancouver	Q1	home entertainment	605
Vancouver	Q1	computer	825
.....	.....	.....	.....
.....	.....	.....	.....
.....	.....	.....	.....
Chicago	Q4	Security	
.....	.....	.....	.....

Esempio di cubo di dati (3)

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

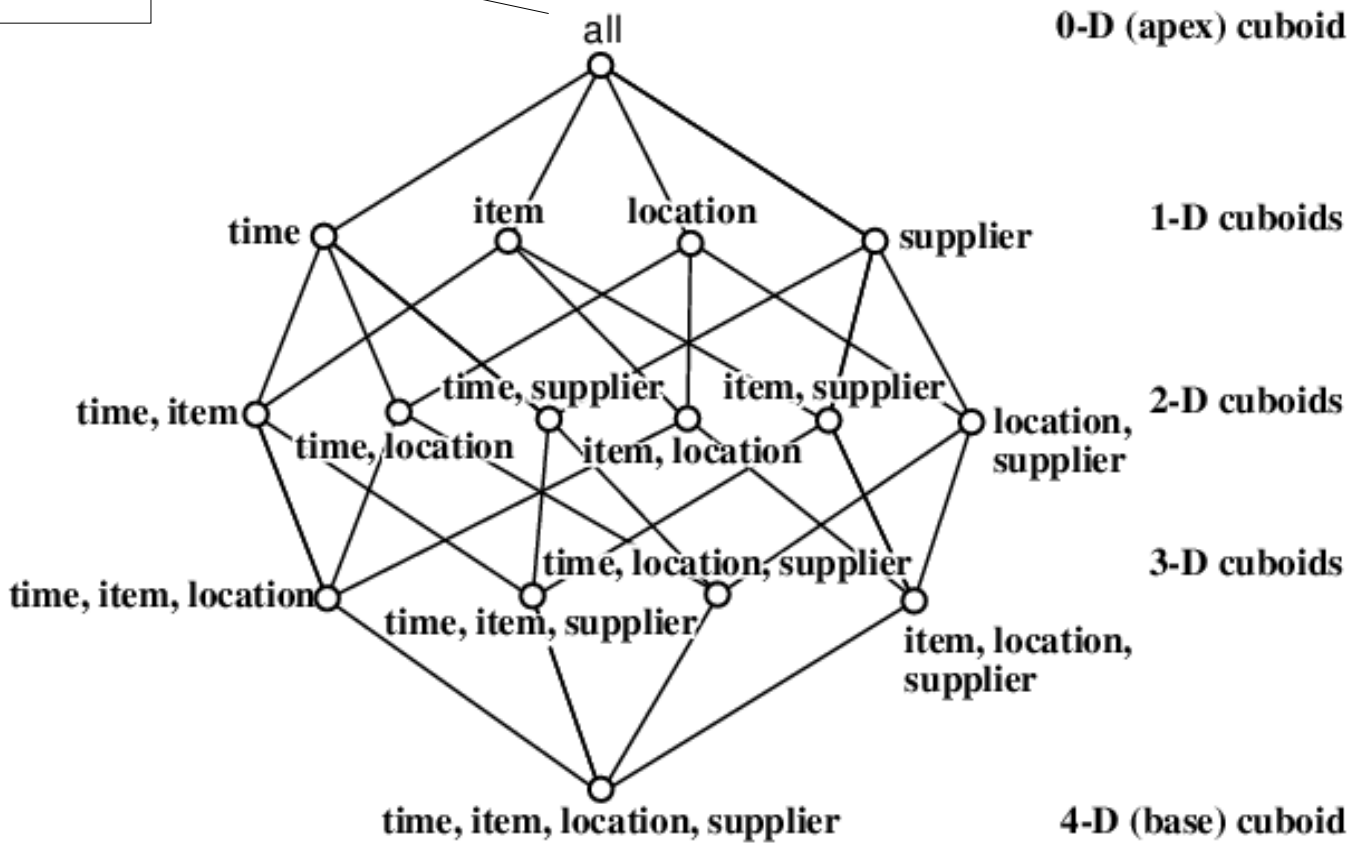


Cuboidi e data cube

- 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

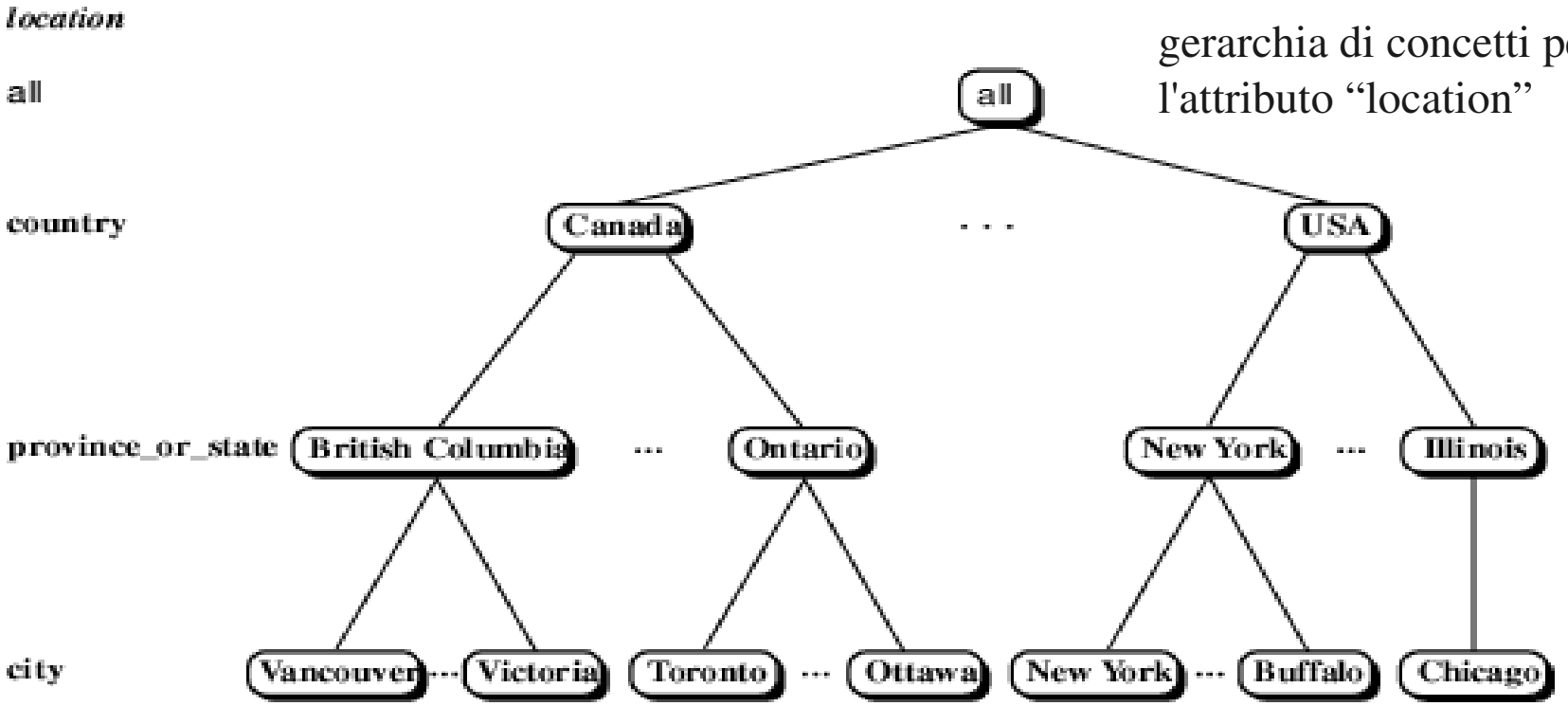
cuboide apice



cuboide base

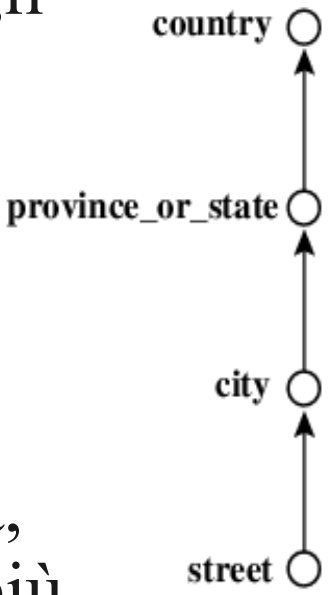
Gerarchie di Concetti (1)

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

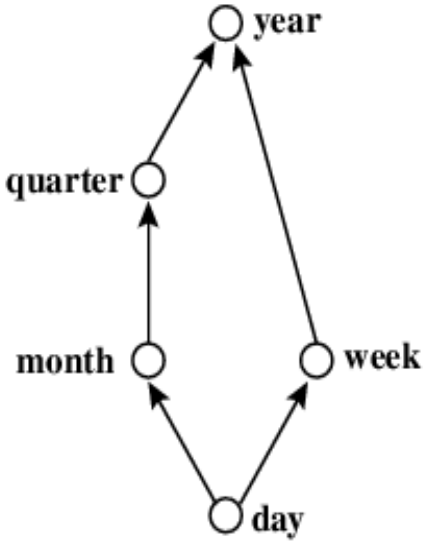


Gerarchie di Concetti (2)

- Molte gerarchie di concetti si ottengono implicitamente dagli attributi che costituiscono il database.
  - la dimensione location è descritta dagli attributi street, city, province, country.
- Nella rappresentazione grafica, gli attributi sono ordinati dal più concreto al più generale.
- Si ottiene una schema hierarchy



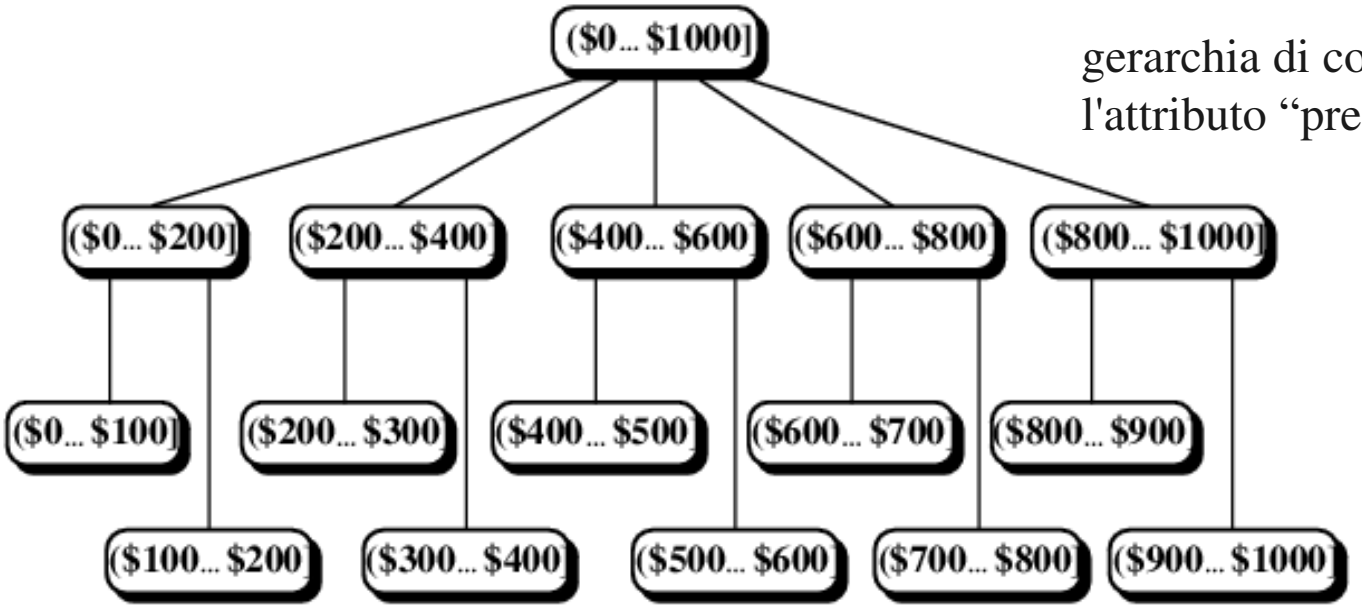
(a)



(b)

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.



gerarchia di concetti per l'attributo “prezzo”

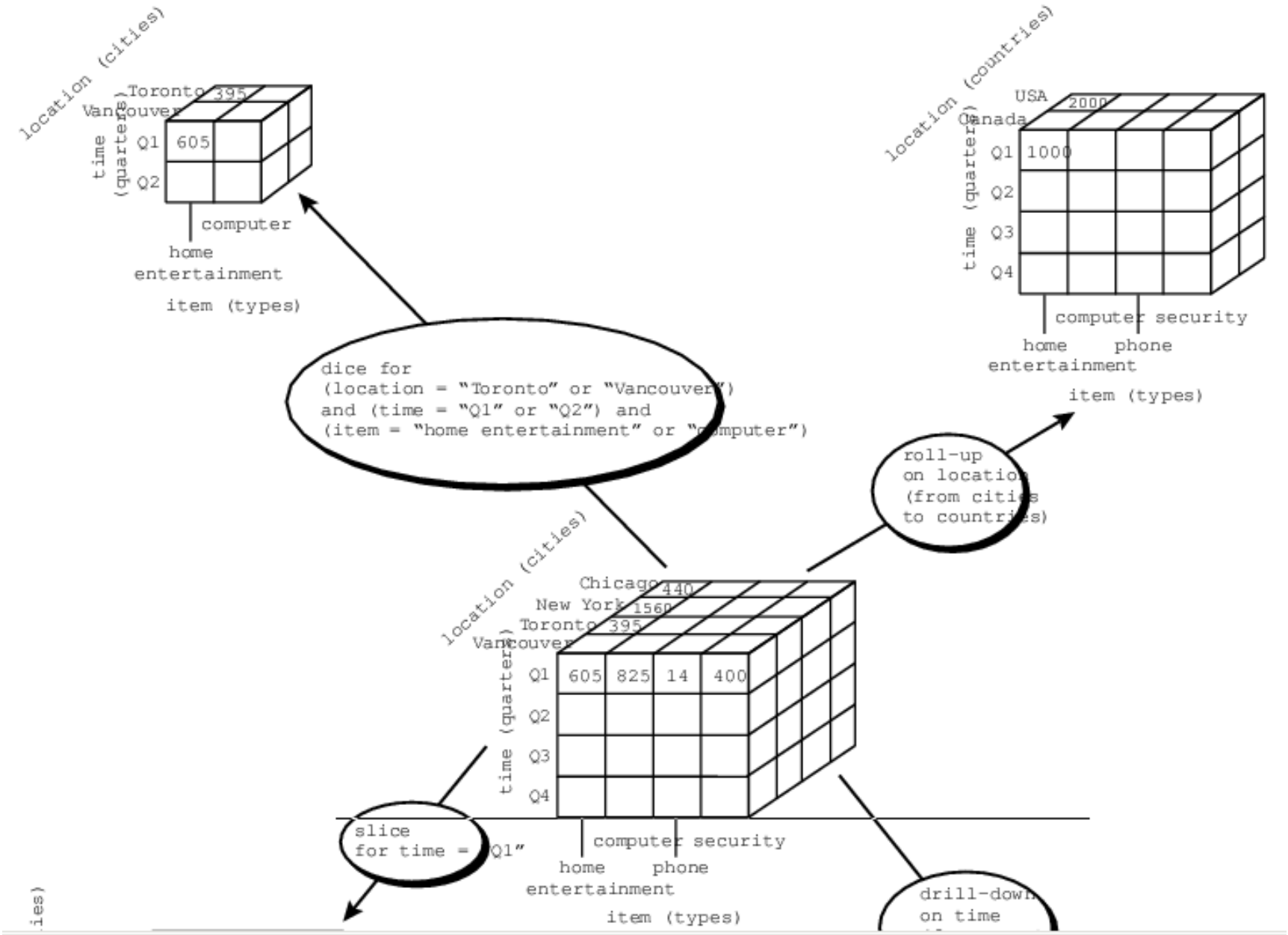
Gerarchie di Concetti (4)

- Le gerarchie di concetti possono:
  - essere predefinite dal sistema di data warehouse
    - ad esempio la gerarchia di concetti per l'attributo “time”;
    - il sistema consente di adattare la gerarchia predefinita alle esigenze dell'utente: ad esempio, far iniziare una settimana con la domenica o il lunedì.
  - essere fornite manualmente dall'utente o da un esperto del dominio di applicazione;
  - essere generate automaticamente sulla base di analisi statistiche
    - parleremo in lezioni future della generazione automatica di gerarchie.

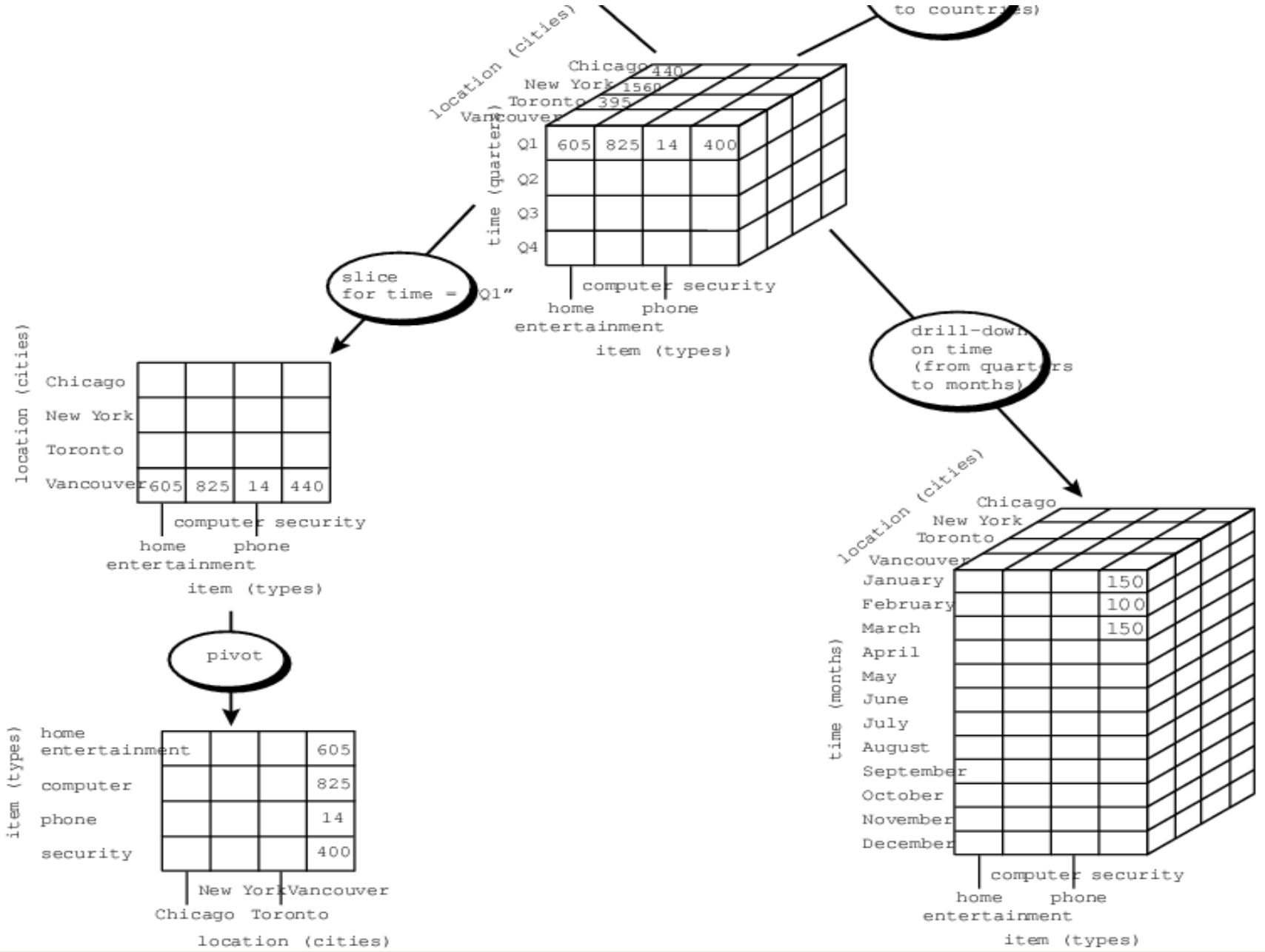
Operazioni sui cuboidi (1)

- I sistemi OLAP basati sul modello multi-dimensionale dei dati mettono a disposizione una serie di operazioni standard sui cuboidi.
  - Roll-up (drill-up): esegue delle aggregazioni risalendo sulle gerarchie dei concetti o eliminando alcune dimensioni.
  - Drill-down: è l'inverso del roll-up, si sposta da dati meno dettagliati a dati più dettagliati introducendo nuove dimensioni o scendendo nelle gerarchie dei concetti.
  - Slice: esegue una selezione su una dimensione, ottenendo un sottocubo di quello di partenza.
  - 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)

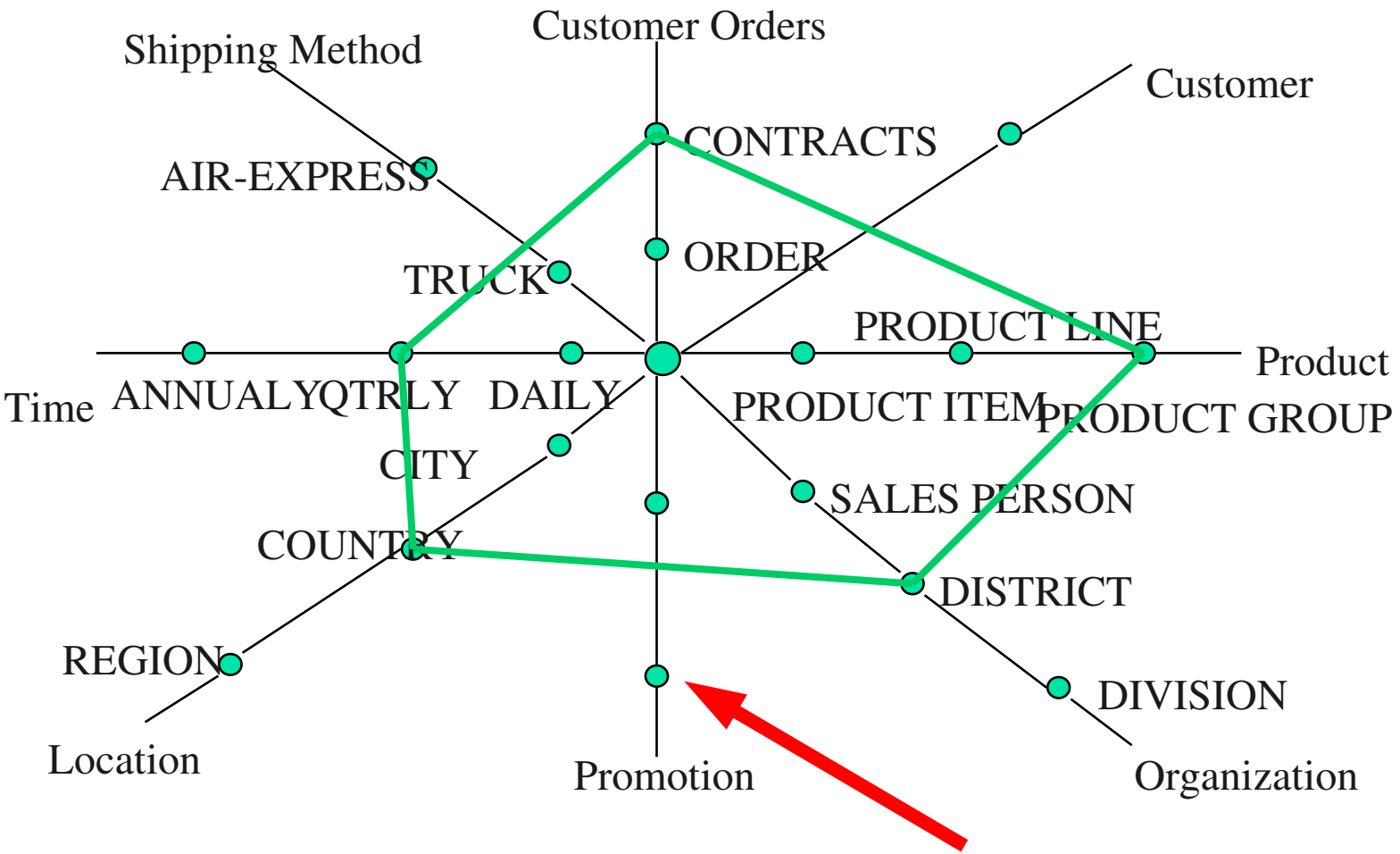


Operazioni sui cuboidi (4)

- Altre operazioni sui cuboidi:
  - drill-trough: quando il data ware-house è ottenuto a partire da dati in un database relazionale, scende sotto il livello di dettaglio del cuboide di base, accedendo direttamente ai dati di partenza.
  - operazioni statistiche: calcolo di valori medie, varianze, etc..
  - operazioni di matematica finanziaria
  - ....

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.

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 coordinate del punto.
- Ci sono tre tipi di funzioni di aggregazione, e quindi di misure, che è possibile utilizzare:
  - distributive: se S è un insieme di dati ed S_1, ..., S_n è una sua partizione, allora il valore della funzione per S è ricavabile dai valori della stessa funzione per S_1, ..., S_n.
    - Esempio: sum(), count(), min(), max().
      - $sum(S_1, ..., S_n)=sum(sum(S_1), ... sum(S_n))$
      - $count(S_1, ..., S_n)=sum(count(S_1), ... count(S_n))$

Tipi di misura (2)

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

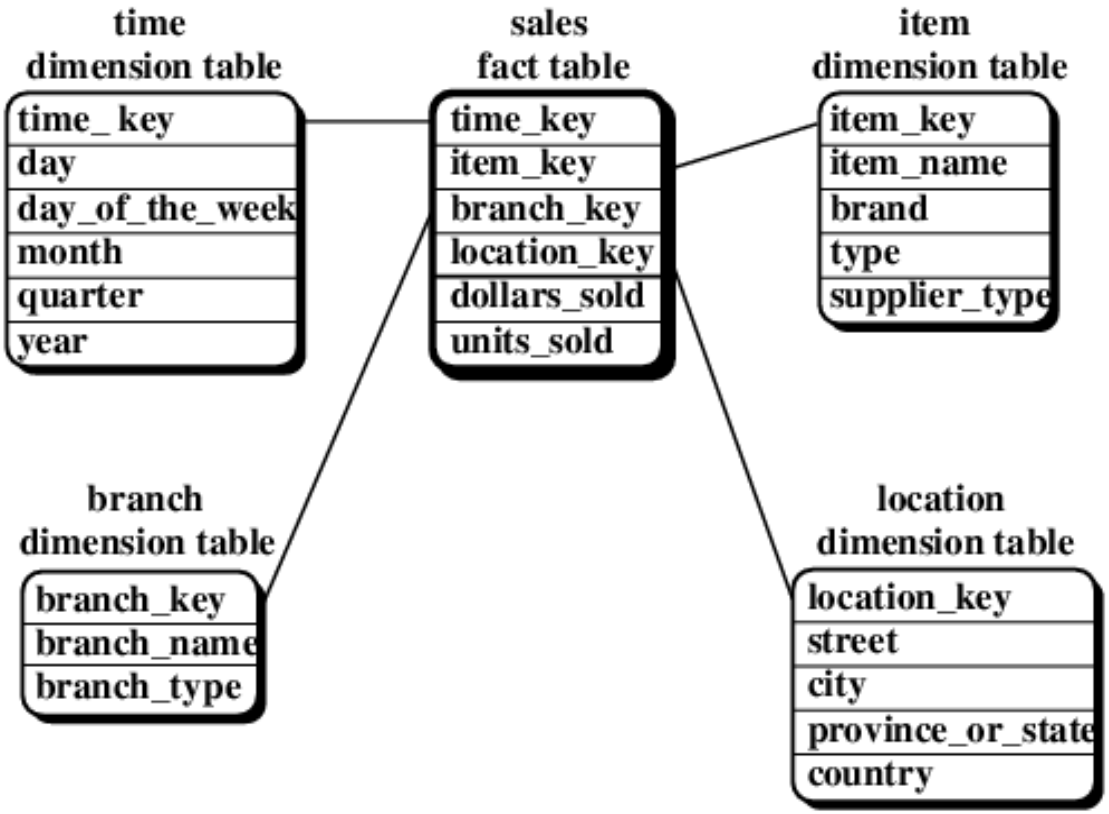
- Le funzioni olistiche sono difficoltose da calcolare. Esistono metodi di approssimazione.

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 ha una tabella delle dimensioni associata, che descrive gli attributi di cui è composta.
  - La dimensione oggetto può contenere gli attributi nome, marca, tipo.
- Il nucleo del data warehouse è memorizzato in una tabella dei fatti
  - "Unità di prodotto vendute" e "Ricavato dalla vendita" sono esempi di fatti.

Schema a stella (1)

- 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.

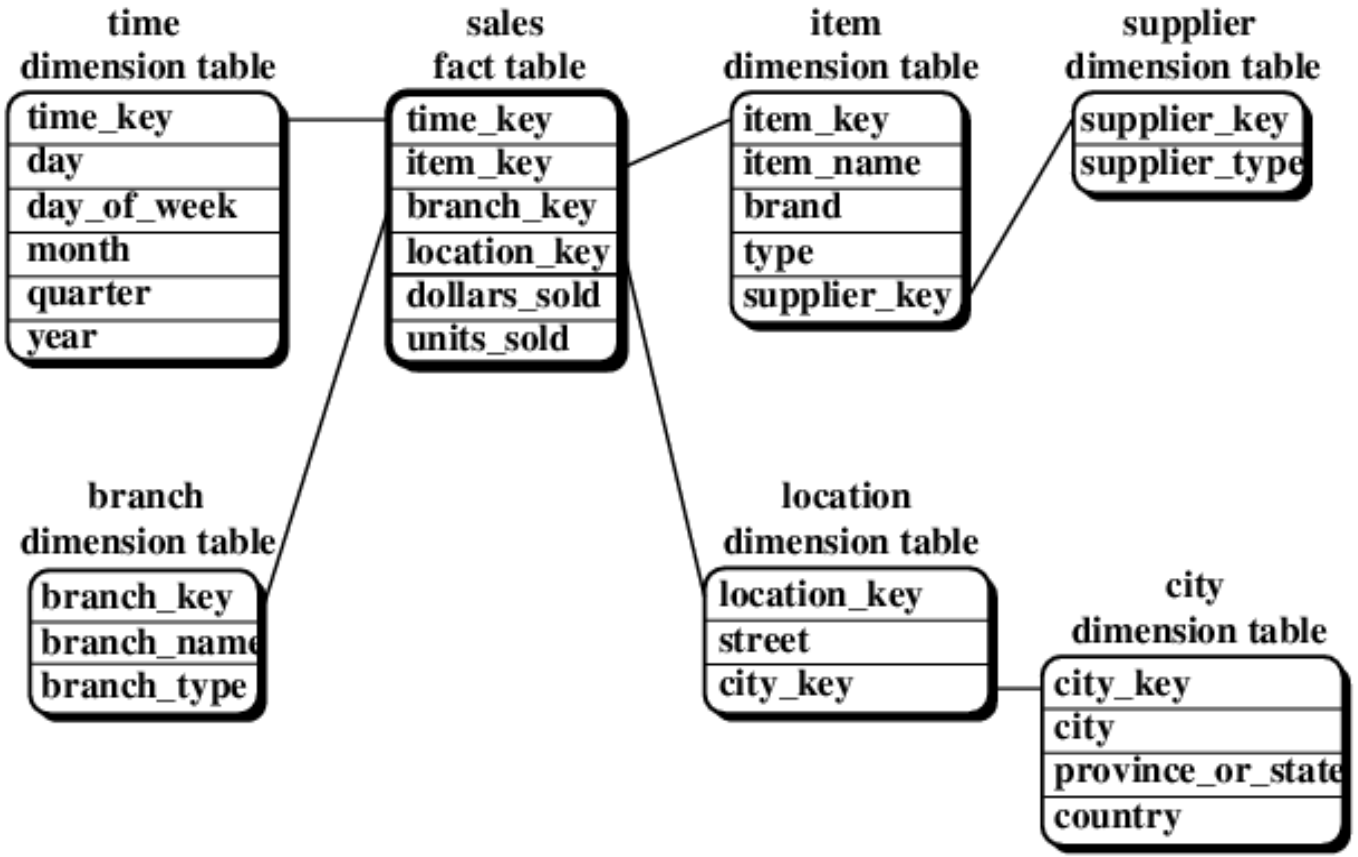


Schema a stella (2)

- Dato una schema a stella, un cuboide viene determinato scegliendo:
  - un fatto dalla tabella dei fatti
  - un insieme di dimensioni
  - per ogni dimensioni scelta, un attributo nella relativa tabella.
- Ad esempio, il cuboide al lucido 18 corrisponde alla scelta del fatto units_sold e degli attributi type, quarter e city.
- Il cuboide corrisponde al risultato della query SQL:
  - `select sum(units_sold) from sales natural join item natural join location natural join time group by type, quarter, city.`

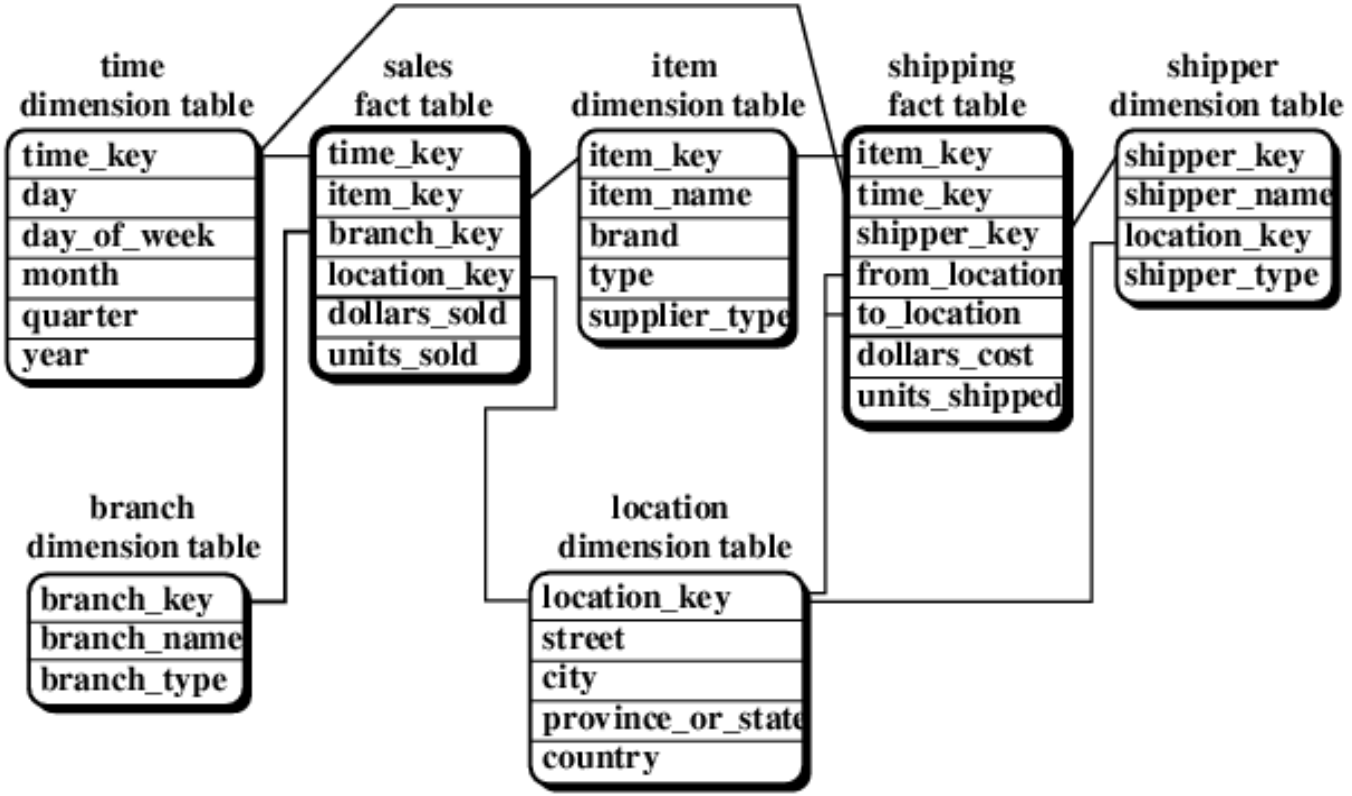
Schema a fiocco di neve

- 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 warehouse => fact constellation schema
- data mart => star schema

Data Warehouse e OLAP

Cosa è un data warehouse

Un modello dei dati multidimensionale

Architettura dei data warehouse

Dai data warehouse al data mining

Architettura di un Data Warehouse

Lo sviluppo di un data warehouse

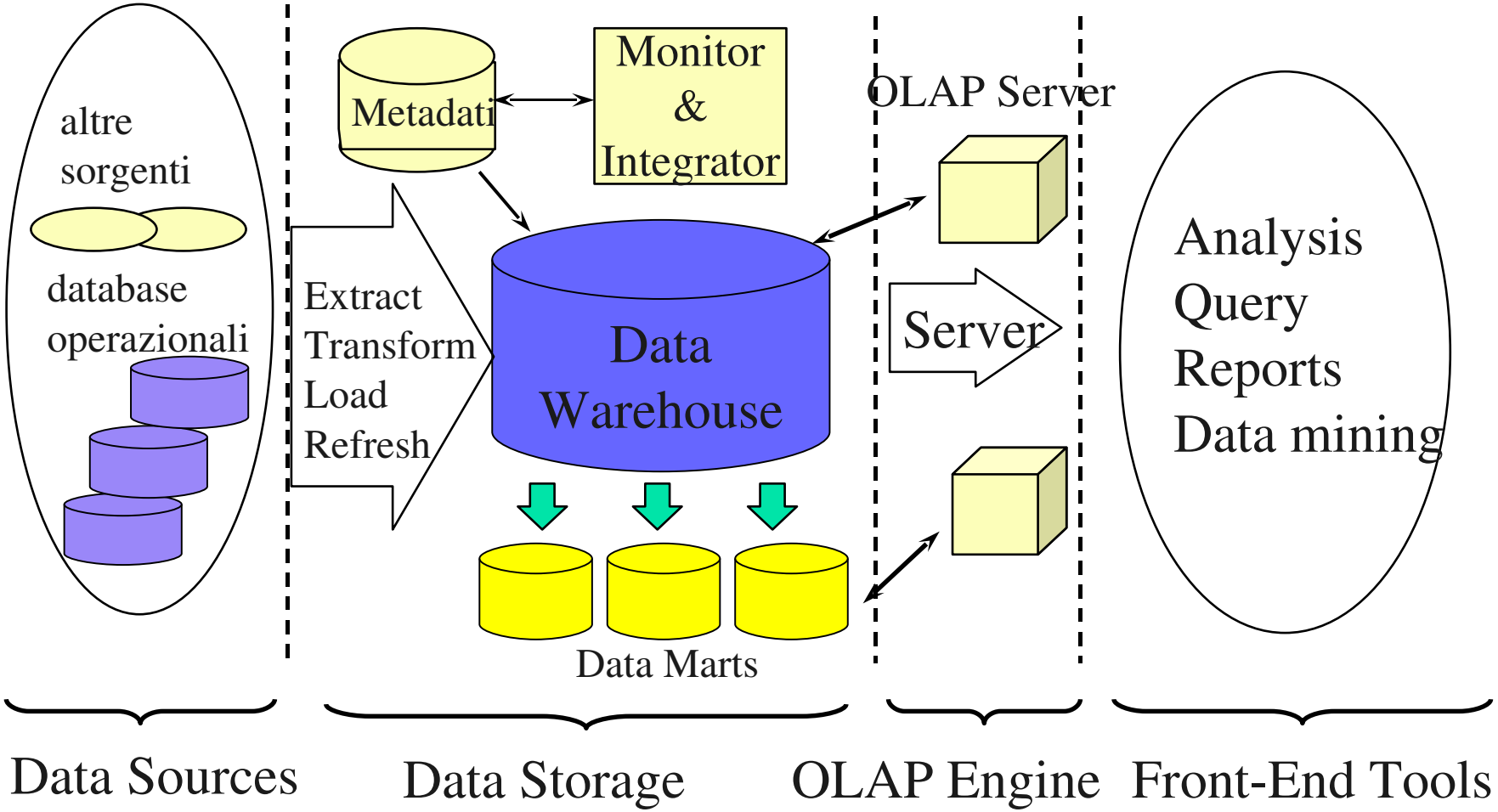
- Si può seguire un approccio top-down, bottom-up o misto
  - top-down: si inizia con la pianificazione della struttura generale e poi si passa alla implementazione di tutto il data warehouse. Utile se il problema è chiaro e se la tecnologia è matura
  - bottom-up: si inizia con esperimenti e prototipi 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 massima (singole transazioni, riassunti giornalieri, etc..)
  - scelta delle dimensioni e delle gerarchie di concetti
  - scelta delle misure che popoleranno la tabella dei fatti
- Altri aspetti dell'uso del DW che vanno presi in considerazione sono:
  - installazione iniziale e addestramento del personale
  - aggiornamento dei dati, piani per “disaster recovery”, controllo degli accessi, controllo della crescita dei dati

Architettura di un data warehouse

- L'architettura tipica è a 3 livelli



Tecnologie in un sistema OLAP

- Data Warehouse: 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.

Server OLAP (2)

- 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
- 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 Data Mining

OLAP vs Data Mining

- Con i sistemi OLAP è possibile scoprire regolarità nei dati
  - in particolare, l'attività di data mining che abbiamo chiamato "Descrizione di Concetti" 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 (1)

- 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:
    - JDBC, 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.

OLAP e Data Mining (2)

- L'integrazione di sistemi OLAP con data mining prende il nome di OLAM (on-line analytical mining).
  - sono di solito tool interattivi;
  - permettono di manipolare i cuboidi con le operazioni standard dei sistemi OLAP;
  - consentono di richiamare funzioni di data mining su richiesta;
  - permettono di applicare funzioni OLAP ai risultati delle analisi.