IL DATA WAREHOUSE

Vedremo

- Il cubo multidimensionale
- Eventi secondari e aggregazione
- Schemi a stella e a fiocco di neve
- Progettazione di un Data Warehouse

Il modello multidimensionale

Modello alla base della rappresentazione e per l'interrogazione delle informazioni in un Data Warehouse.

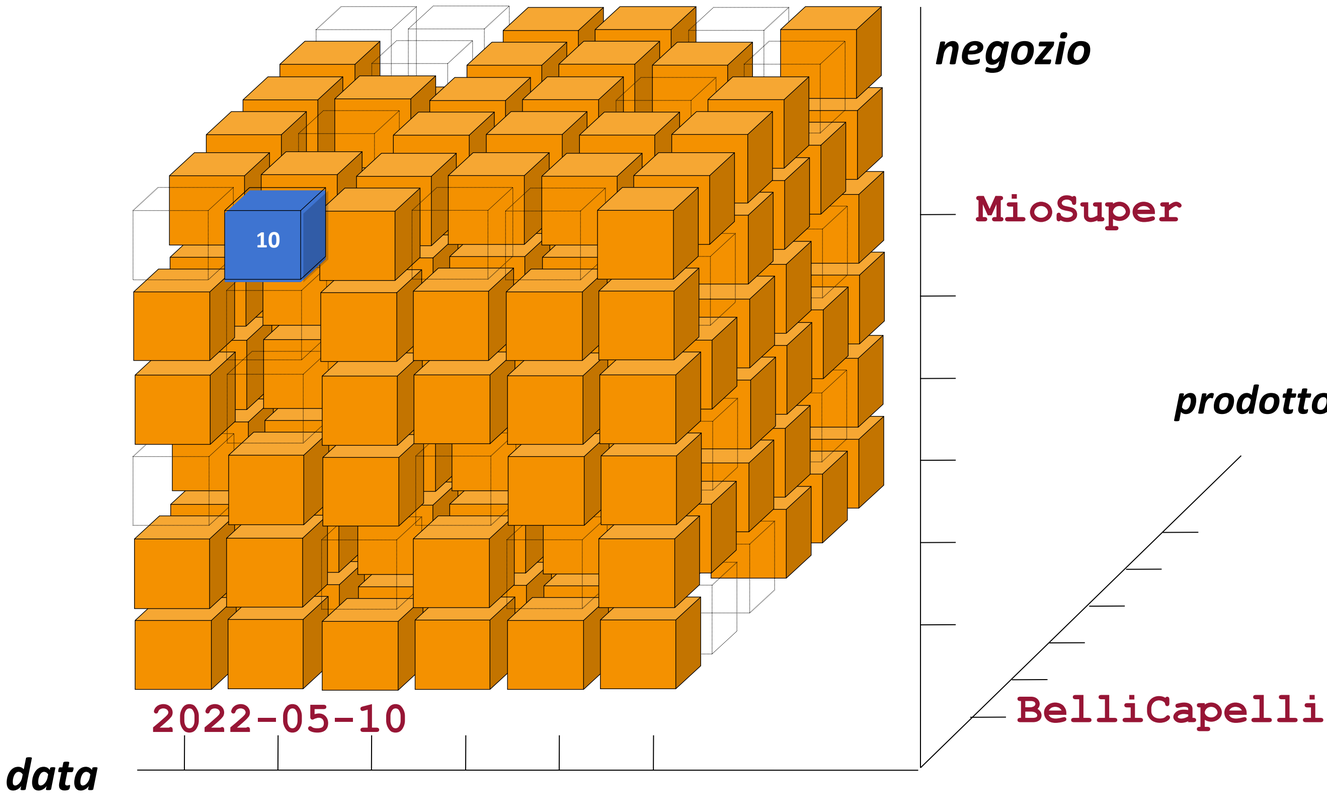
I fatti di interesse per l'analisi sono rappresentati in cubi:

- ogni cella del cubo memorizza misure numeriche;
- ogni asse è una dimensione per l'analisi;
- ogni dimensione può essere la radice di una gerarchia (diversa granularità).

Il cubo delle vendite

- Ogni cella rappresenta le vendite di un prodotto, in una data, in un punto vendita.

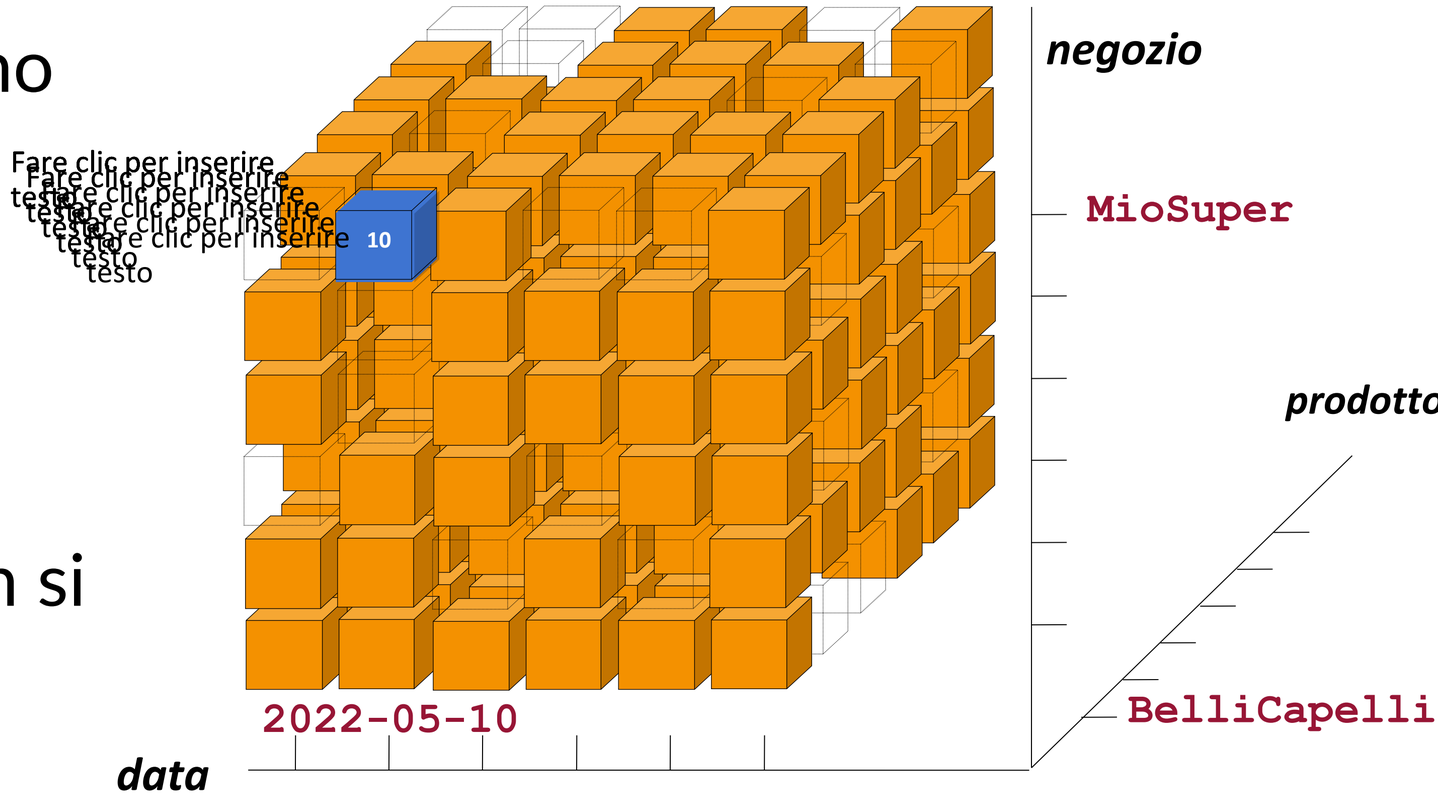
- La cella blu corrisponde al fatto che 10 confezioni di shampoo BelliCapelli siano state vendute nel punto vendita MioSuper il 10 maggio 2022.



Fatti

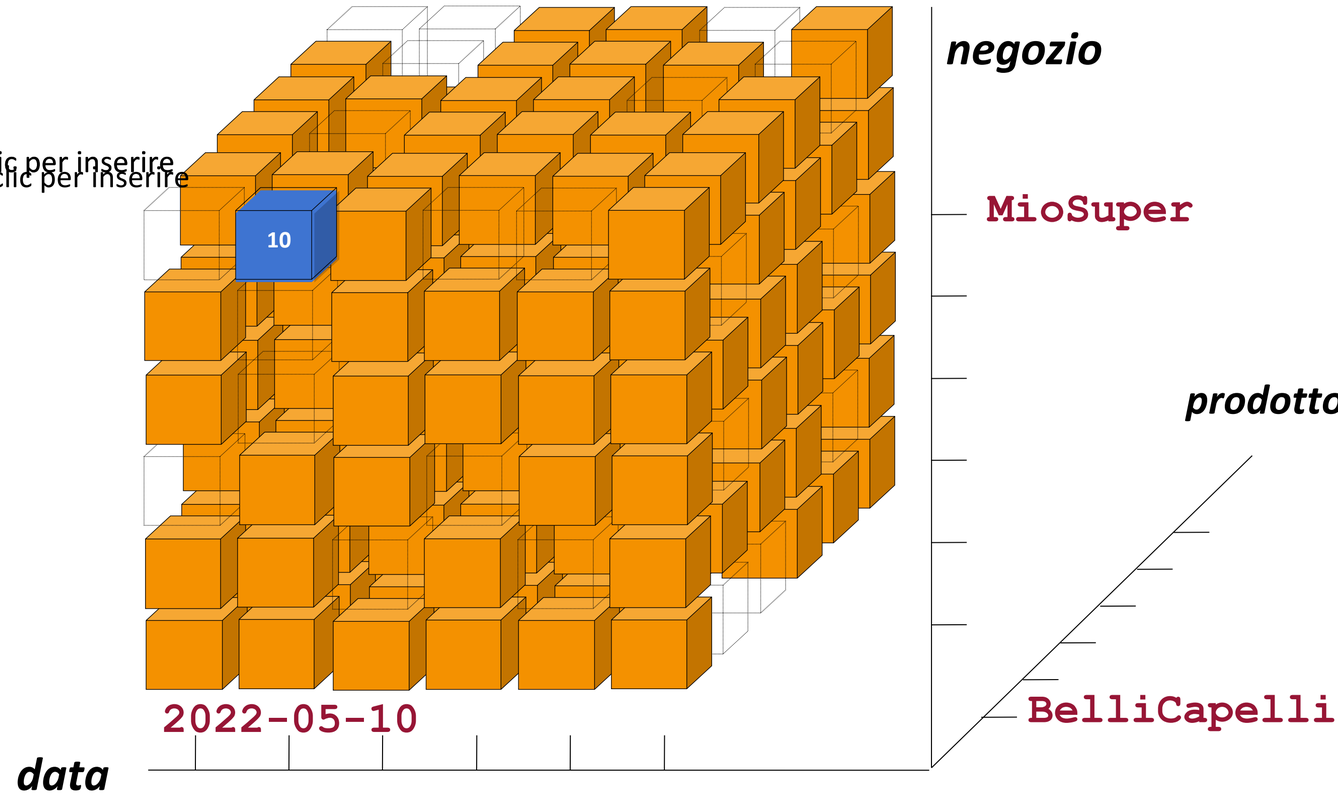
Un fatto è un concetto rilevante per i processi decisionali:

- eventi che si svolgono all'interno di un'organizzazione;
- vendite (di un prodotto, in un punto vendita, in una data), ma anche spedizioni, ordini, ecc;
- una cella del cubo;
- il cubo è **sparso**: eventi che non si sono realizzati sono celle tratteggiate.



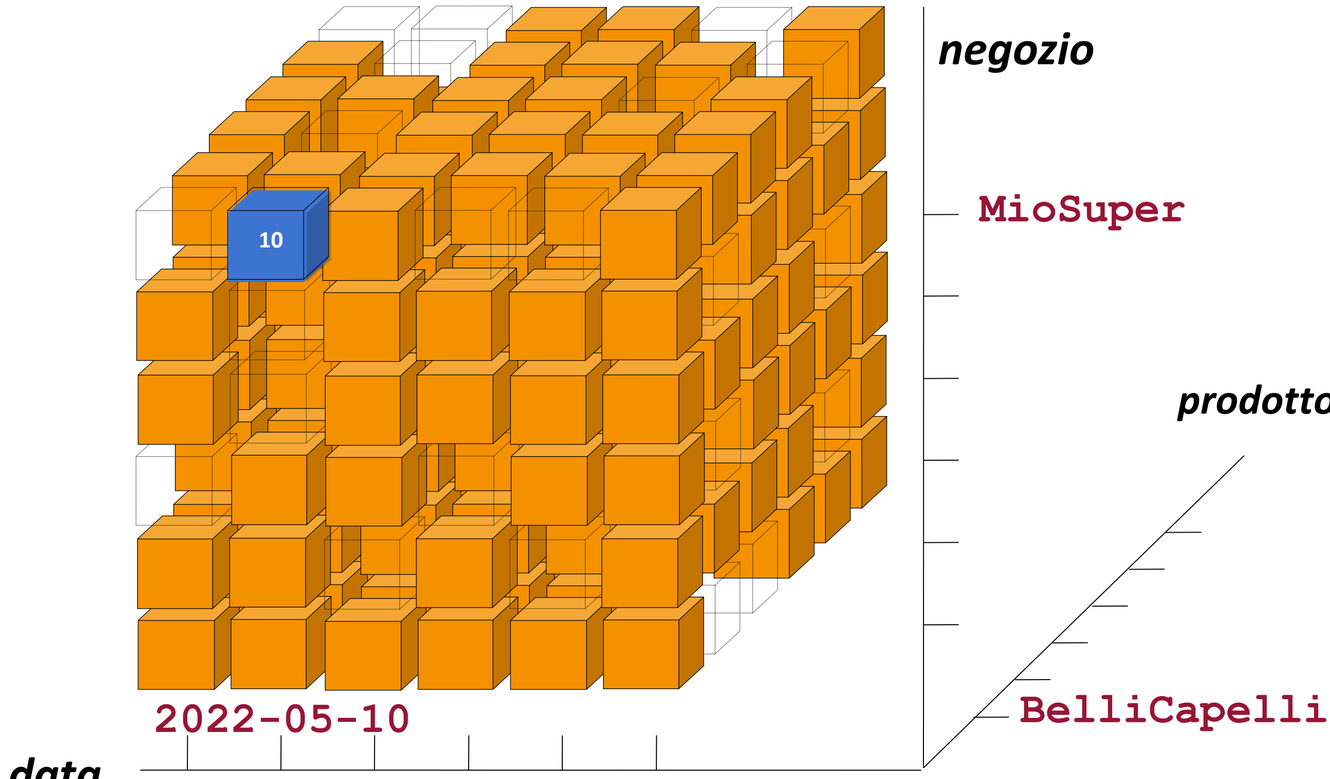
Misure

- Una misura è una proprietà numerica di un fatto:
- descrive una proprietà quantitativa rilevante per l'analisi;
- ogni vendita è quantificata dal numero di prodotti venduti e dal ricavo corrispondente;
- la cella blu riporta il valore della misura quantità (10).



Dimensioni

- Una dimensione è una proprietà di fatto con un dominio finito;
- descrive una prospettiva, un punto di vista, da cui analizzare il fatto;
- le dimensioni tipiche per il fatto di vendita sono: i prodotti (cosa?), negozi (dove?) e date (quando?);
- corrispondono agli assi dello spazio (non necessariamente 3) e quindi individuano le coordinate delle celle.



Attributi dimensionali

Gli attributi dimensionali sono:

- le dimensioni;
- altri possibili attributi a valori discreti, che le descrivono.

- Un prodotto è descritto dalla sua tipologia, dalla categoria di appartenenza;
- gli attributi dimensionali in una stessa dimensione sono legati da associazioni uno-a-molti e da dipendenze funzionali.

prodotto —— tipo ——— categoria

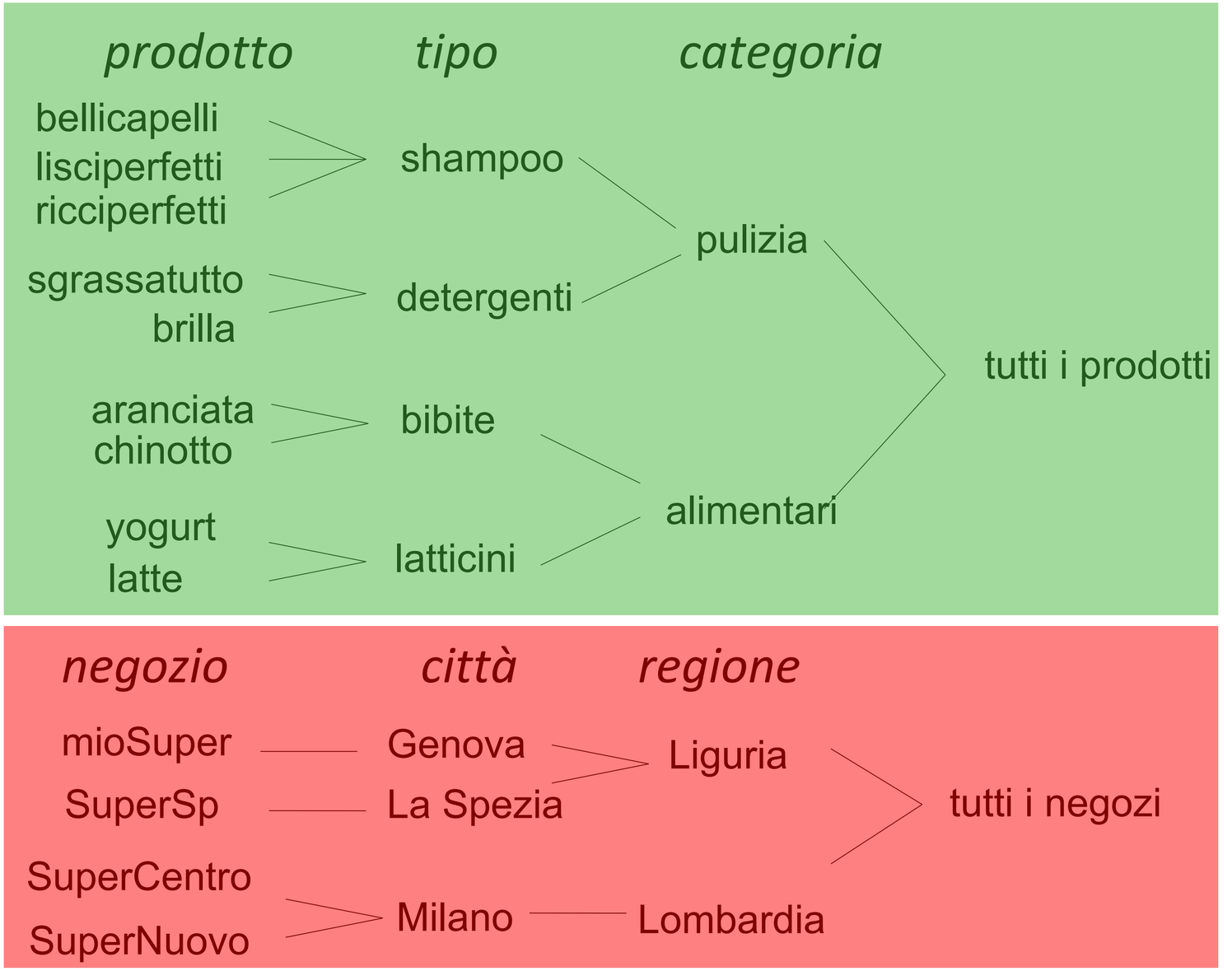
prodotto → tipo
tipo → categoria

negozio——— città ——regione

negozio → città
città → regione

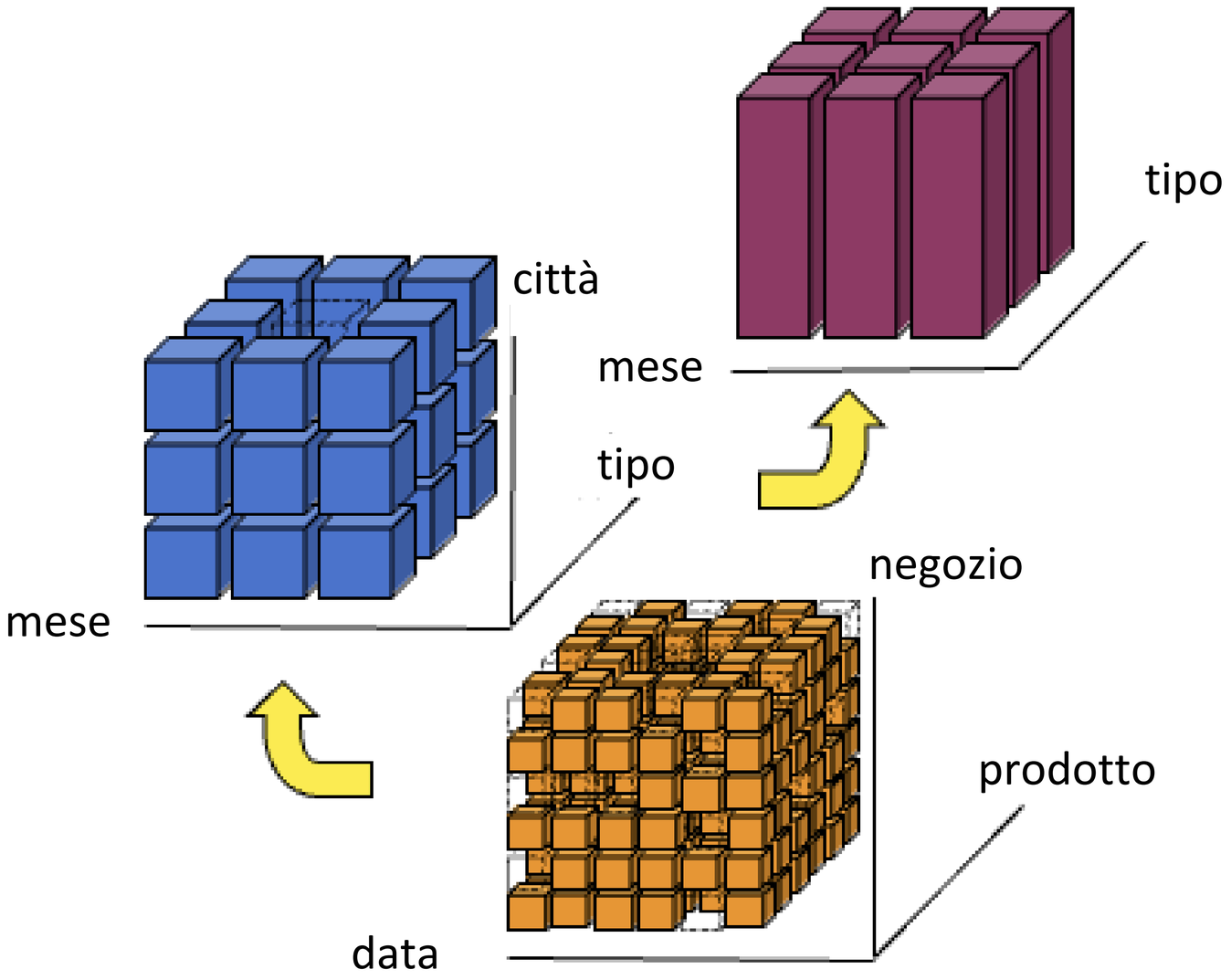
Gerarchie

- diversi livelli di dettaglio rispetto a cui osservare i fatti rispetto allo stesso punto di vista;
- corrispondono a un «cambio di scala» o a diversi livelli di zoom nello spazio multidimensionale.



Gerarchie e aggregazione

- Le gerarchie definiscono il modo in cui è possibile aggregare gli eventi primari;
- la dimensione alla radice di una gerarchia definisce la sua granularità di aggregazione più fine;
- gli altri attributi dimensionali corrispondono a una granularità gradualmente più grossolana.



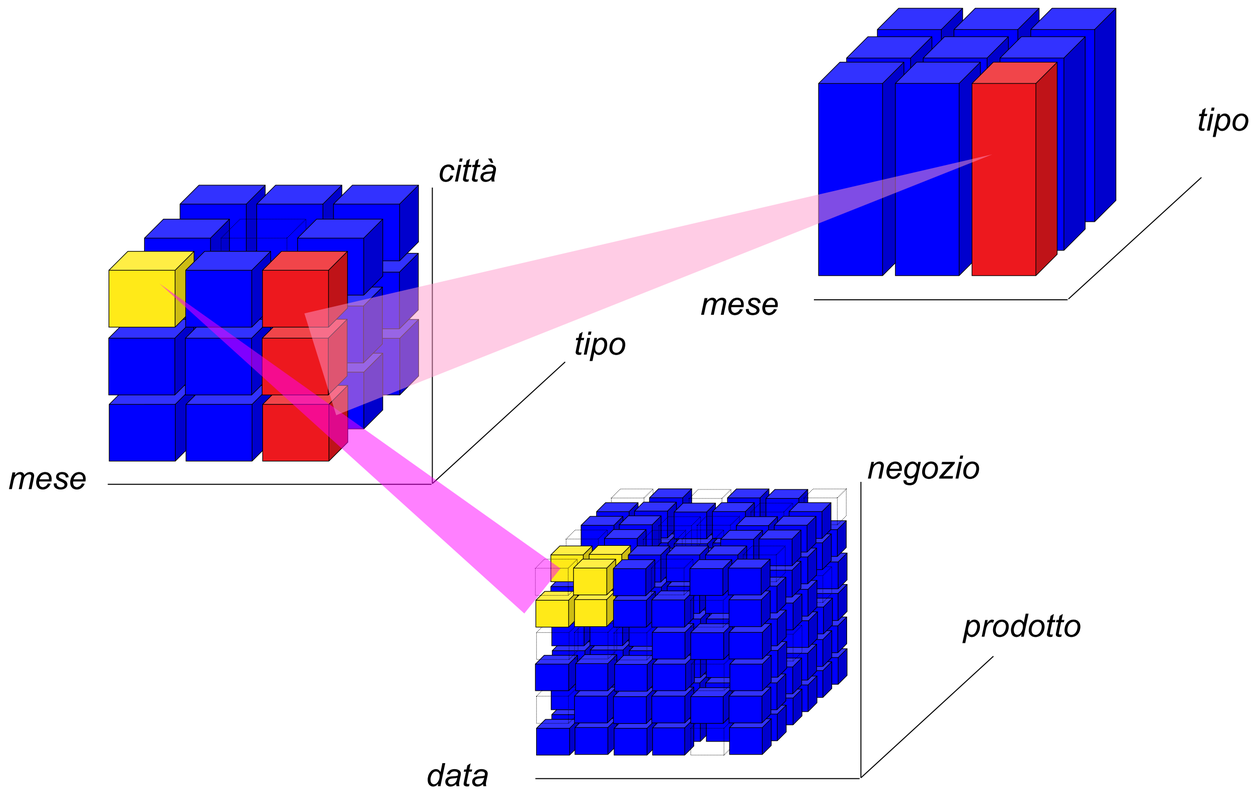
Eventi primari e secondari

- Eventi primari:
  - celle le cui coordinate si riferiscono alla dimensione (radice della gerarchia);
  - massimo livello di disaggregazione a cui possiamo analizzare i dati.
- Eventi secondari:
  - celle ottenute aggregando rispetto ad altri attributi dimensionali;
  - possono essere derivati a partire dagli eventi primari.

Gli insiemi di attributi dimensionali rispetto alla relazione di ordine parziale «più dettagliato di» costituiscono un lattice.

Misure e eventi secondari

- I valori delle misure di un evento secondario sono calcolati a partire da quelli delle misure degli eventi primari corrispondenti;
- le vendite mensili per tipo di prodotto in una città sono calcolate in base a quelle dei prodotti di quel tipo venduti nei punti vendita di quella città nelle date di quel mese.



Misure e aggregazione

L’aggregazione richiede la definizione di un **operatore idoneo** a comporre i valori.

- Le misure possono essere classificate in:
  - **misure di flusso** valutate cumulativamente al termine di un arco temporale (numero di prodotti venduti in un giorno);
  - **misure di livello** valutate in momenti particolari (il numero di prodotti in inventario);
  - **misure unitarie** valutate in momenti particolari e espresse in termini relativi (prezzo unitario del prodotto).

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

- A seconda degli operatori utilizzabili le misure sono anche dette **additive, semi-additive, aggregabili, non aggregabili.**

Il modello multidimensionale dei dati – altri attributi

- Attributi **descrittivi**: memorizzano informazioni aggiuntive su un attributo dimensionale;
- attributi **cross-dimensionali**: i valori sono definiti dalla combinazione di due o più attributi dimensionali;
- attributi **condivisi** da diverse gerarchie o porzioni di gerarchie condivise tra più dimensioni;
- attributi **opzionali**, gerarchie incomplete;
- associazioni **molti-a-molti** tra attributi dimensionali.

ROLAP, MOLAP, HOLAP

I sistemi di Data Warehouse utilizzano diversi modelli:

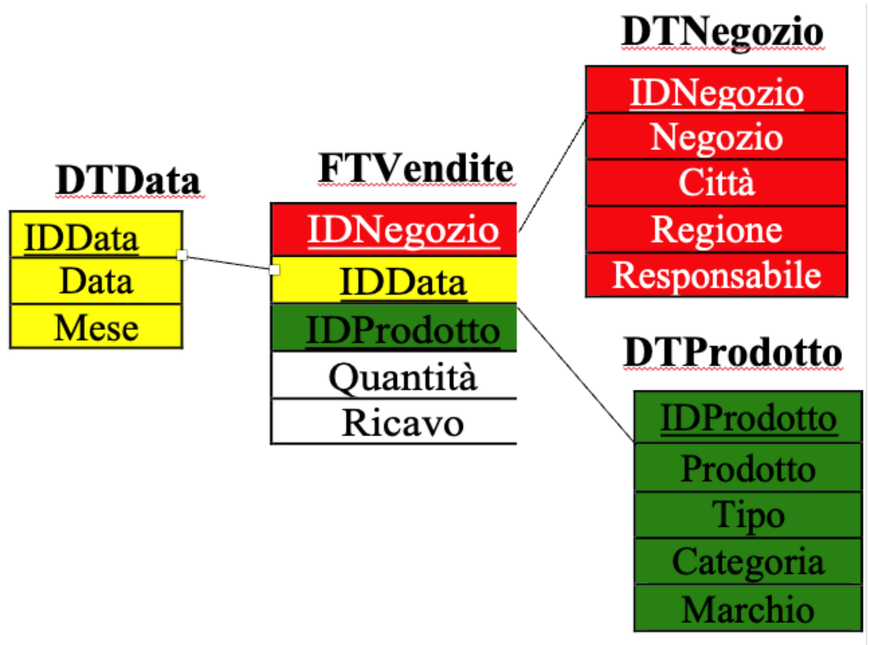
- **ROLAP**: modello relazionale (modello molto conosciuto e tecnologia consolidata).
- **MOLAP**: modello multidimensionale (celle del cubo, con problema della sparsità).
- **HOLAP**: soluzioni ibride, ad es. per cubi primari e secondari.

Schema a stella

Base per la modellazione multidimensionale nei sistemi relazionali.

Uno schema a stella è costituito da:

- una tabella delle dimensioni per ogni dimensione con:
  - una chiave primaria (tipicamente surrogata);
  - gli attributi dimensionali (diversi livelli di aggregazione).
- Una tabella dei fatti contenente:
  - le misure;
  - chiavi esterne che riferiscono le tabelle delle dimensioni;
  - la cui chiave è l'unione delle chiavi esterne.



Schema a stella (segue)

- Le tabelle delle dimensioni sono de-normalizzate, a causa della transitività delle dipendenze funzionali.
- Per ogni negozio in una certa città, viene ripetuto in che regione si trova quella città.
- Ogni informazione è recuperabile con un'unica operazione di join a partire dalla tabella dei fatti.
- Il nome schema a stella è proprio dovuto al fatto che le tabelle delle dimensioni sono disposte (come «raggi») intorno alla tabella dei fatti.

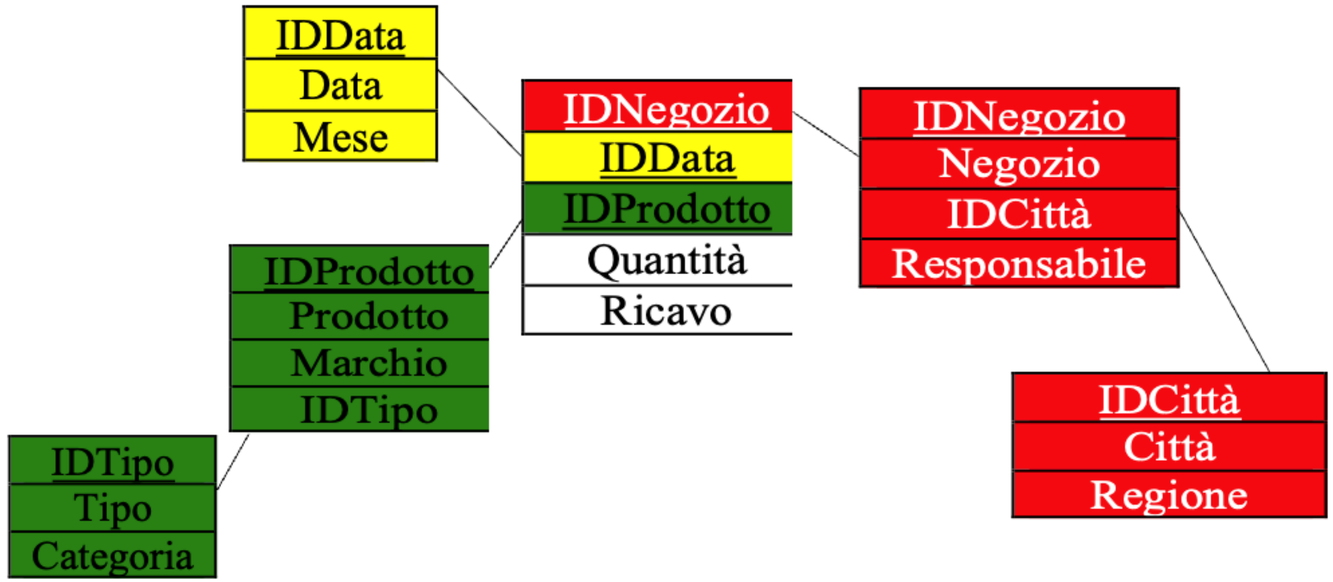
Schema a fiocco di neve

Riduce la de-normalizzazione delle tabelle delle dimensioni rimuovendo alcune delle dipendenze transitive.

Le tabelle delle dimensioni contengono:

- una chiave primaria (tipicamente surrogata);
- gli attributi dimensionali determinati funzionalmente da tale chiave;
- zero o più riferimenti di chiave esterna ad altre tabelle delle dimensioni (secondarie);

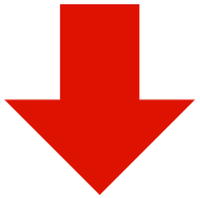
- le tabelle delle dimensioni le cui chiavi sono importate nella tabella dei fatti sono dette primarie.



Schema a fiocco di neve

- maggior normalizzazione;
- interrogazioni che coinvolgono gli attributi delle tabelle delle dimensioni secondarie sono meno efficienti (richiedono un join in più);
- interrogazioni che coinvolgono solo gli attributi contenuti nella tabella dei fatti e nelle tabelle delle dimensioni primarie più efficienti (tabelle più piccole);
- scelta degli attributi su cui effettuare snowflaking;
- particolarmente utile in presenza di eventi secondari memorizzati.

Eventi secondari e viste

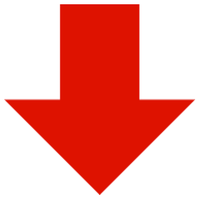


Gli eventi secondari possono essere precalcolati utilizzando viste materializzate.

Questo permette di eseguire interrogazioni OLAP accedendo a cubi secondari di dimensioni molto minori.

La scelta delle viste da materializzare viene effettuata bilanciando lo spazio richiesto (e i relativi tempi di aggiornamento al caricamento di nuovi dati) e i vantaggi in termini di velocizzazione delle interrogazioni OLAP.

Altri schemi relazionali per Data Warehouse

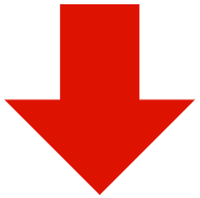


Schema a costellazione:

- più tabelle dei fatti che possono memorizzare fatti primari (es. vendite, spedizioni, promozioni, ecc.) o secondari (es. vendite pre-aggregate per mese).

- Bridge table per gestire associazioni molti a molti, attributi cross-dimensionali, ecc.
- Junk table per dimensioni degeneri.

Progettazione di un Data Warehouse



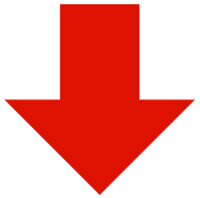
Costruire un Data Warehouse è un'attività molto complessa:

- problematiche e rischi sia di natura organizzativa che architetturale.

Per minimizzare i rischi sono state sviluppate metodologie e sono possibili diversi approcci:

  - Top down vs bottom up;
  - Supply driven vs demand driven;
- qualità dei dati di partenza;
- coinvolgimento di chi ha conoscenza delle sorgenti operazionali.

Riepilogo e conclusioni finali



Abbiamo visto:

- **Il cubo multidimensionale**
- **Schemi a stella e a fiocco di neve**
- **Progettazione di un Data Warehouse**