SAPIENZA
UNIVERSITÀ DI ROMA

DIPARTIMENTO DI INGEGNERIA INFORMATICA AUTOMATICA E GESTIONALE
ANTONIO RUBERTI

Sistemi Informativi Aziendali

Umberto Nanni

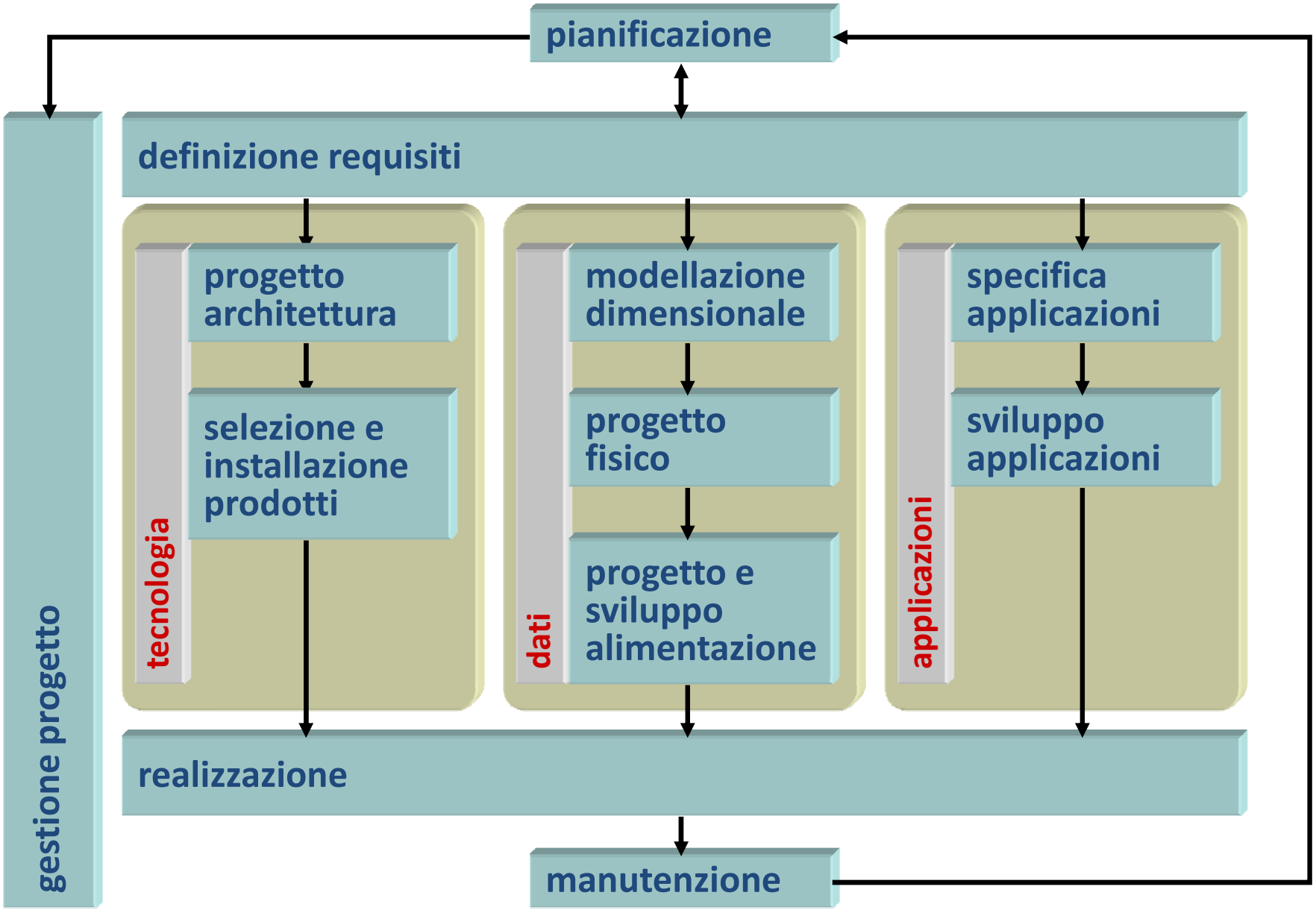
Introduzione al Data Warehousing per Sistemi Informativi Aziendali

b. Progetto di Datawarehouse

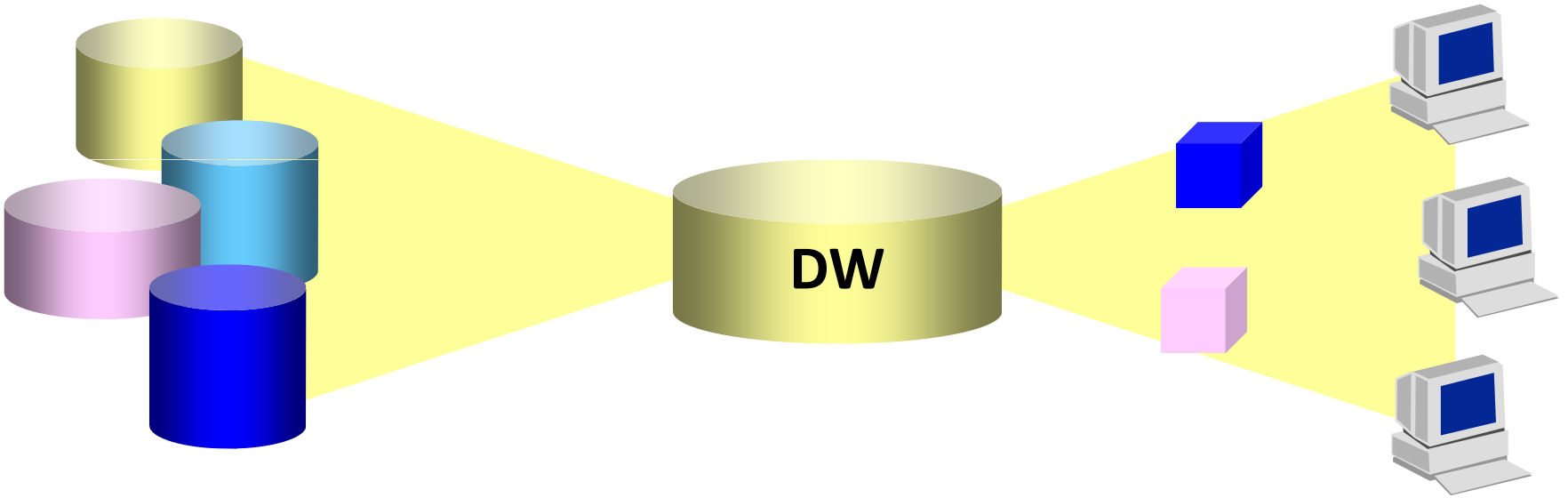
Progetto di Data Warehouse

- Definizione di obiettivi e pianificazione
  - fattibilità (confini, dimensione, sorgenti, …)
  - team
  - piano operativo
- Progetto dell'infrastruttura
  - alternative architetturali
  - alternative tecnologiche
- Progetto e sviluppo dei Data Mart
  - analisi con esperti del dominio

Ciclo di vita (Kimball, 1998)



Flussi dati & Evoluzione progettuale



Flusso dati

Logica di progettazione

Fasi di progetto di un Data Mart

1. Analisi e riconciliazione delle fonti dati **entità-relazione**

   schemi delle sorgenti → schema riconciliato
2. Analisi dei requisiti

   schema riconciliato → fatti, carico lavoro
3. Progetto Concettuale **schemi di fatto**

   schema riconciliato, fatti, carico lavoro → schemi di fatto
4. Progetto Logico **star-schema, snowflakes**

   schemi di fatto, carico lavoro → schema logico Data Mart
5. Progetto dell'Alimentazione

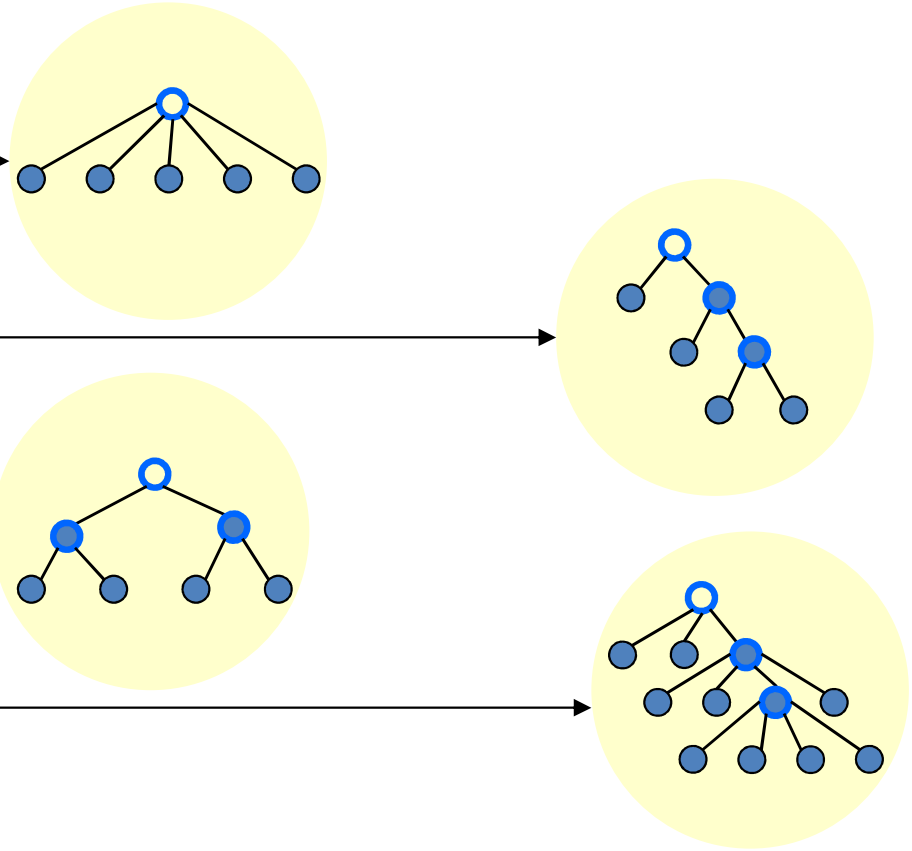
   schemi delle sorgenti, schema riconciliato, schema logico Data Mart → procedure alimentazione
6. Progetto Fisico

   schema logico Data Mart, carico lavoro, DBMS → schema fisico DM

Riconciliazione delle fonti dati

Integrazione di schemi:

- a un passo
- a scala
- bilanciato
- iterativo



Fatti – il punto della situazione

FATTO: categoria di eventi che si verificano nella realtà di interesse dell'organizzazione.

Per ciascun fatto:

- dimensioni: coordinate di analisi/classificazione
- misure: proprietà di un fatto, aspetti quantitativi

- gerarchia dimensionale: per ciascuna dimensione
- granularità di informazione: compromesso(*) tra quantità di informazione ed efficienza

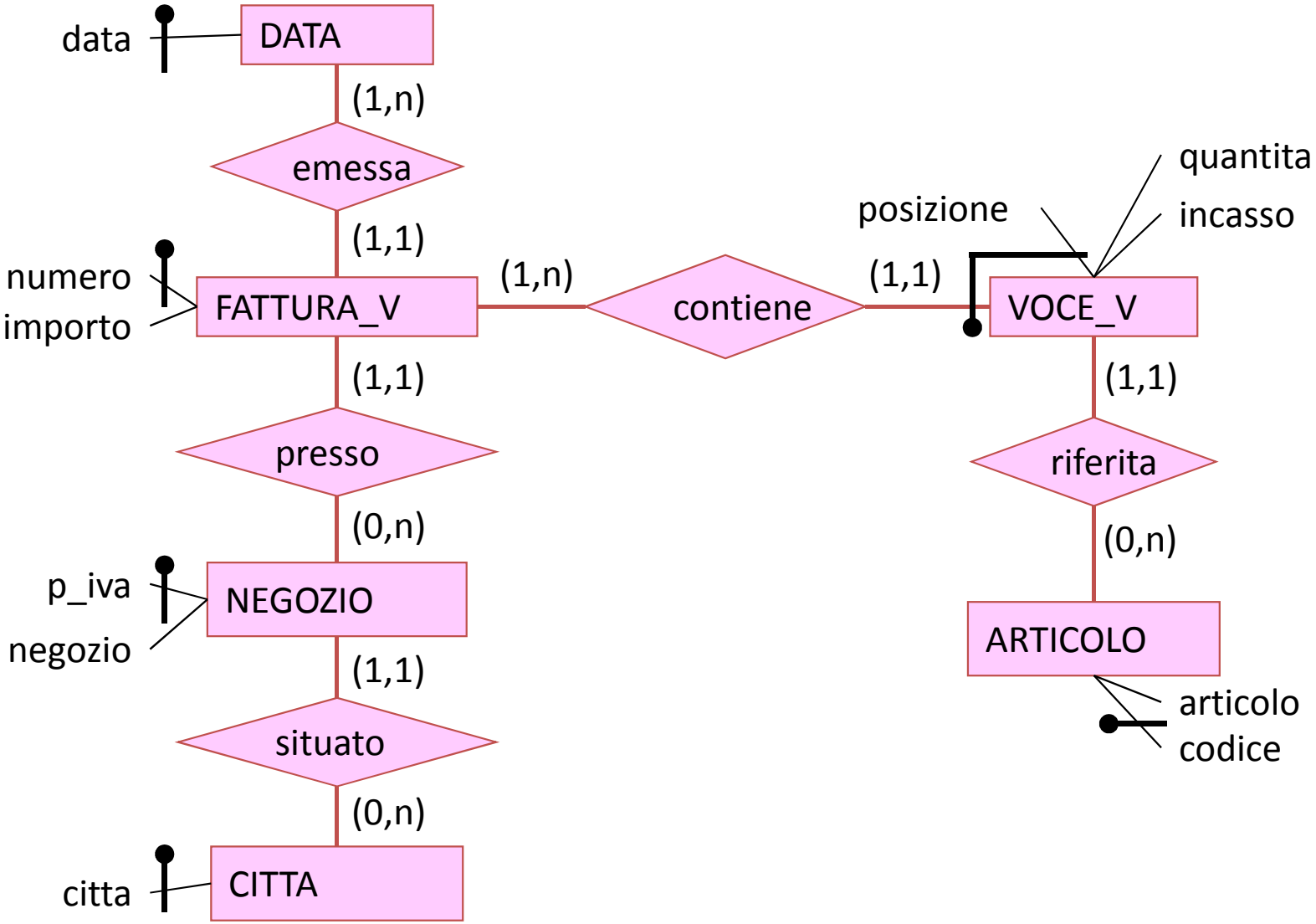
(*) Per compiti specifici esistono, in ogni caso:

- il DB operazionale / riconciliato
- il drill-through

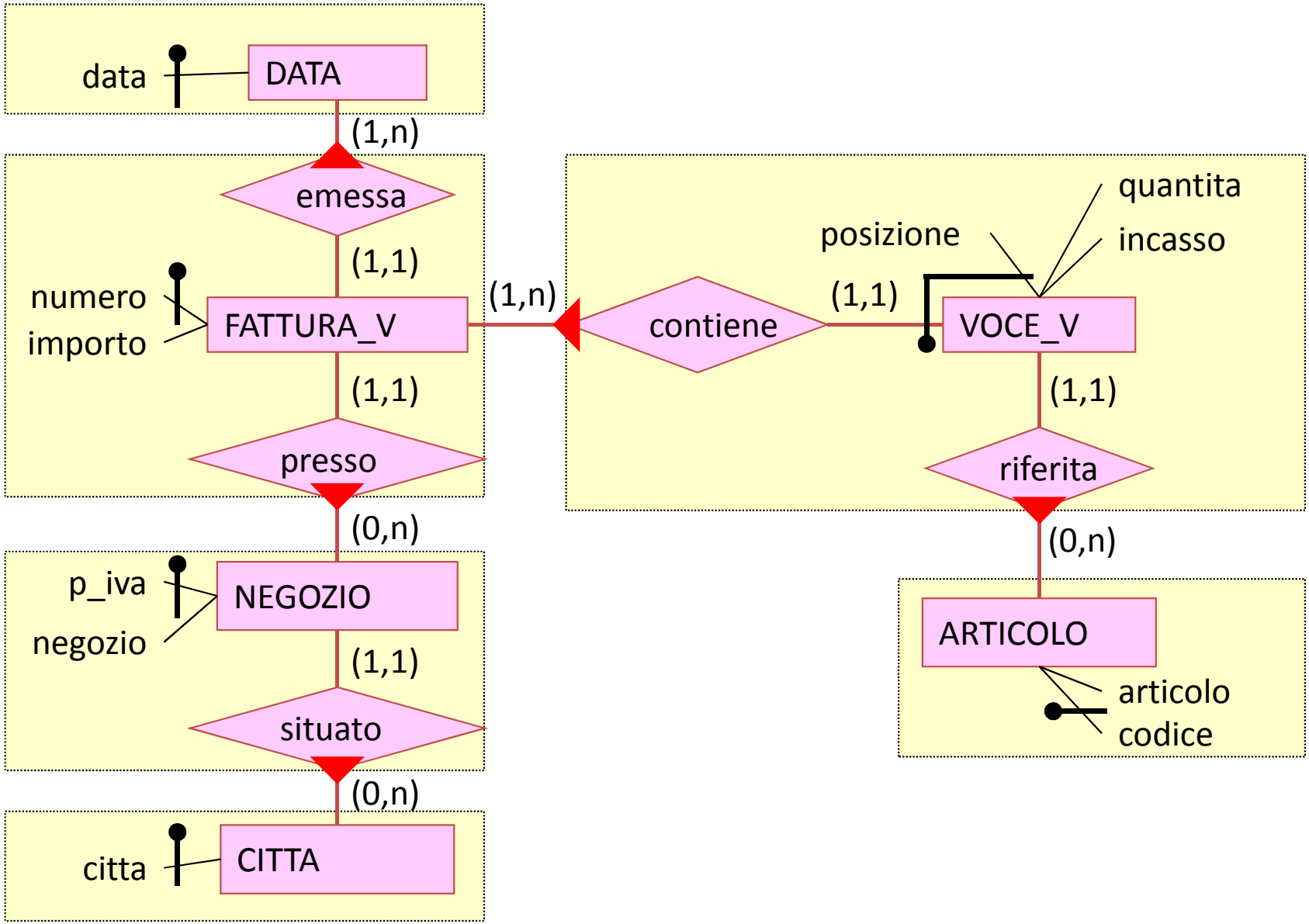
Progetto Concettuale

- Il modello ER non sembra adeguato (anche se resta un fondamentale supporto nella fase di progetto logico)
- Non esiste un consenso unanime sul modello da adottare
- Diverse proposte in letteratura:
  - Multidimensional Entity-Relationship Model
  - DFM - Dimensional Fact Model
  - ……

Schema ER del DB operazionale



Schema logico del DB operazionale

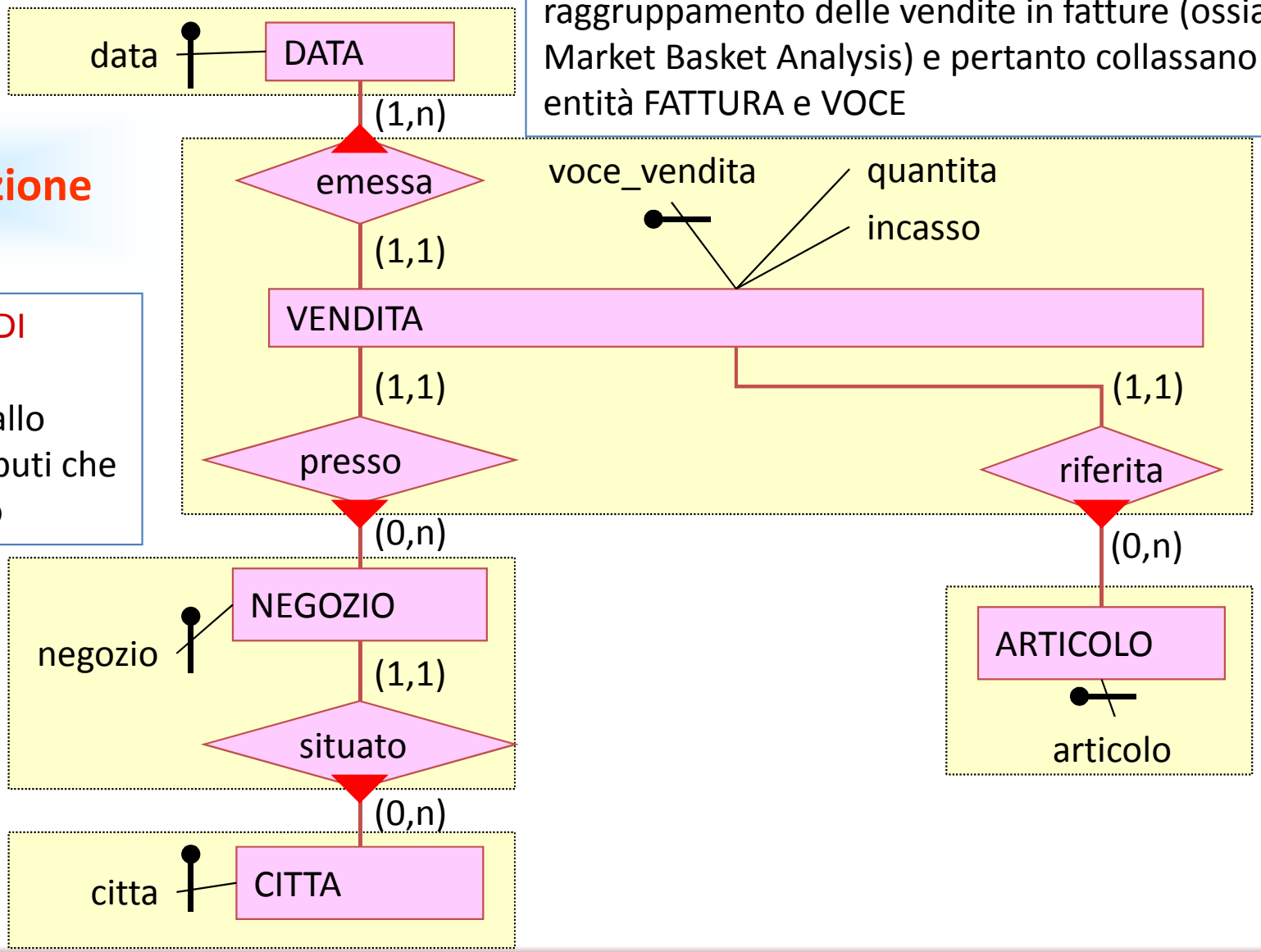


Schema logico DB operazionale - revisione

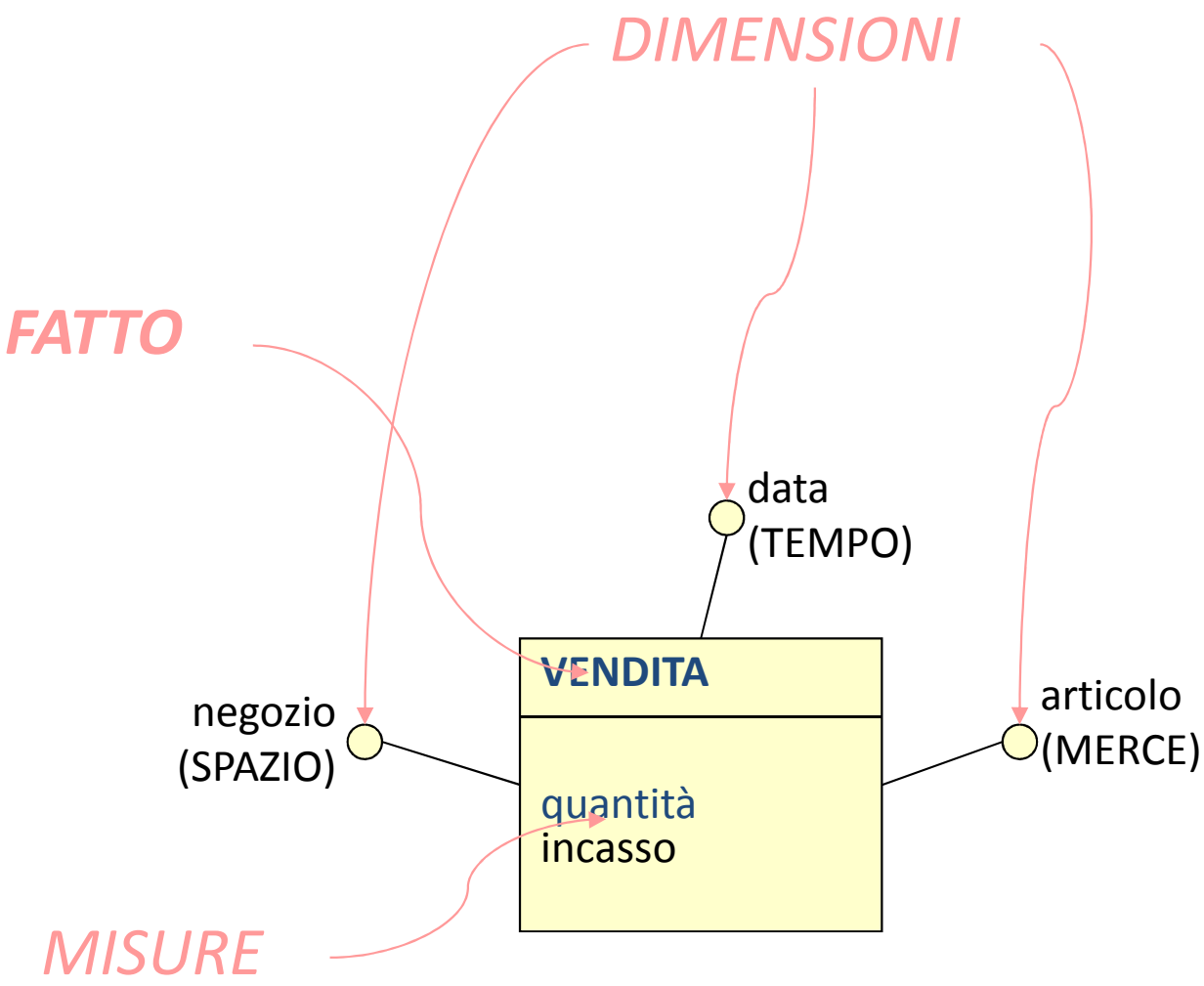
denormalizzazione

ACCORPAMENTO: si ipotizza che non interessi il raggruppamento delle vendite in fatture (ossia la Market Basket Analysis) e pertanto collassano le entità FATTURA e VOCE

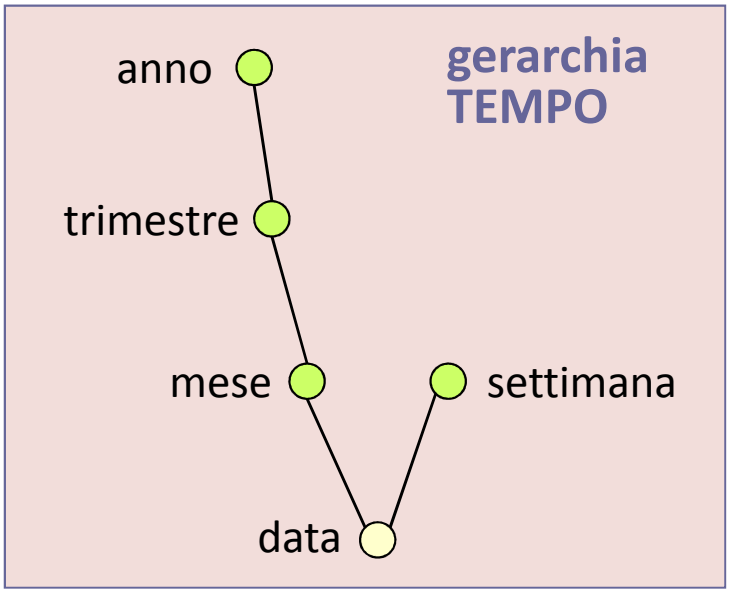
ELIMINAZIONE DI ATTRIBUTI:
vengono tolti dallo schema gli attributi che non interessano

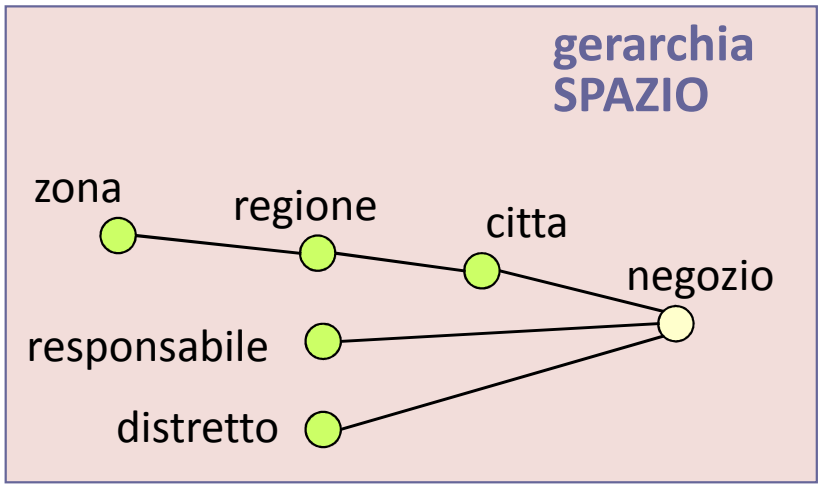


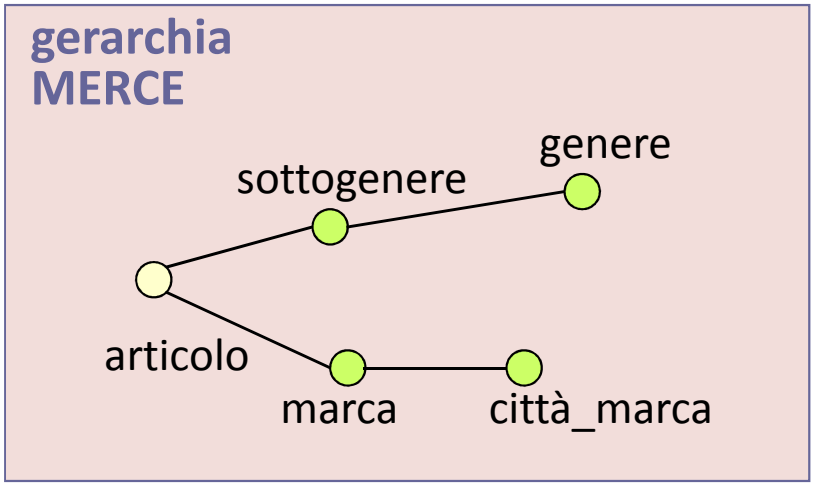
Schema di fatto (preliminare)



Gerarchie dimensionali

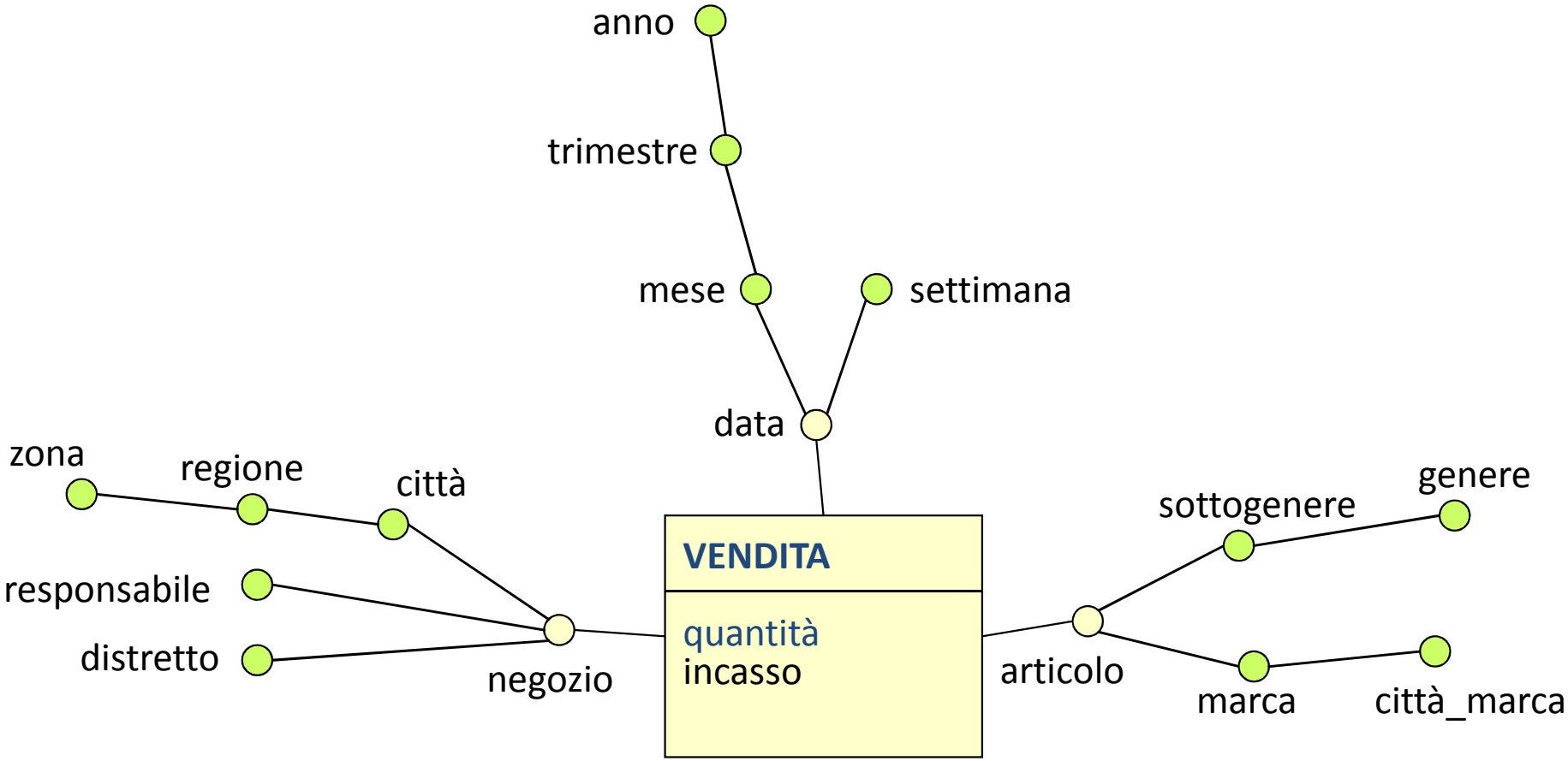






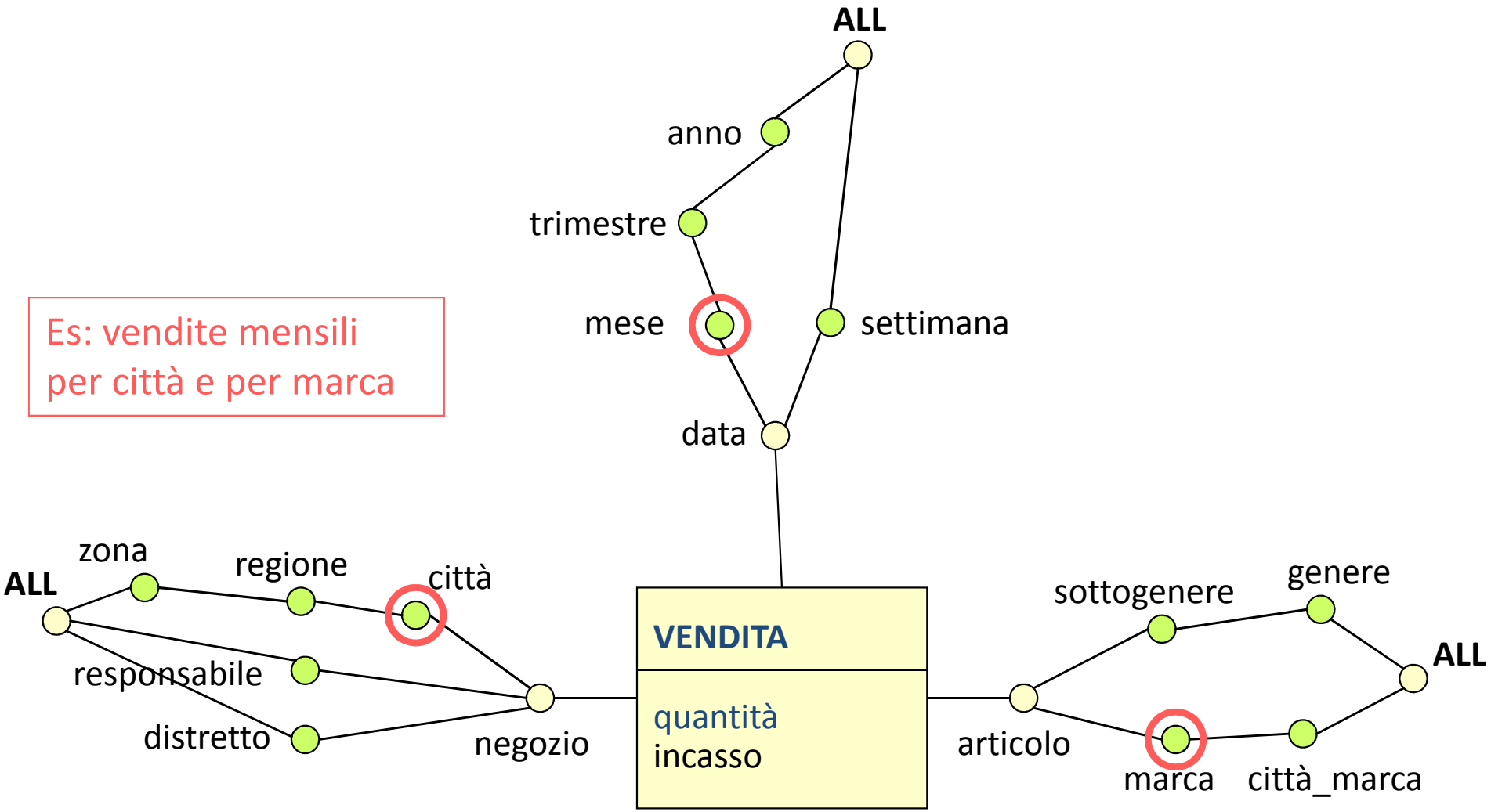
Dettagli (granularità)
In base a
specifiche di utente

Schema di fatto
DFM (Dimensional Fact Model)

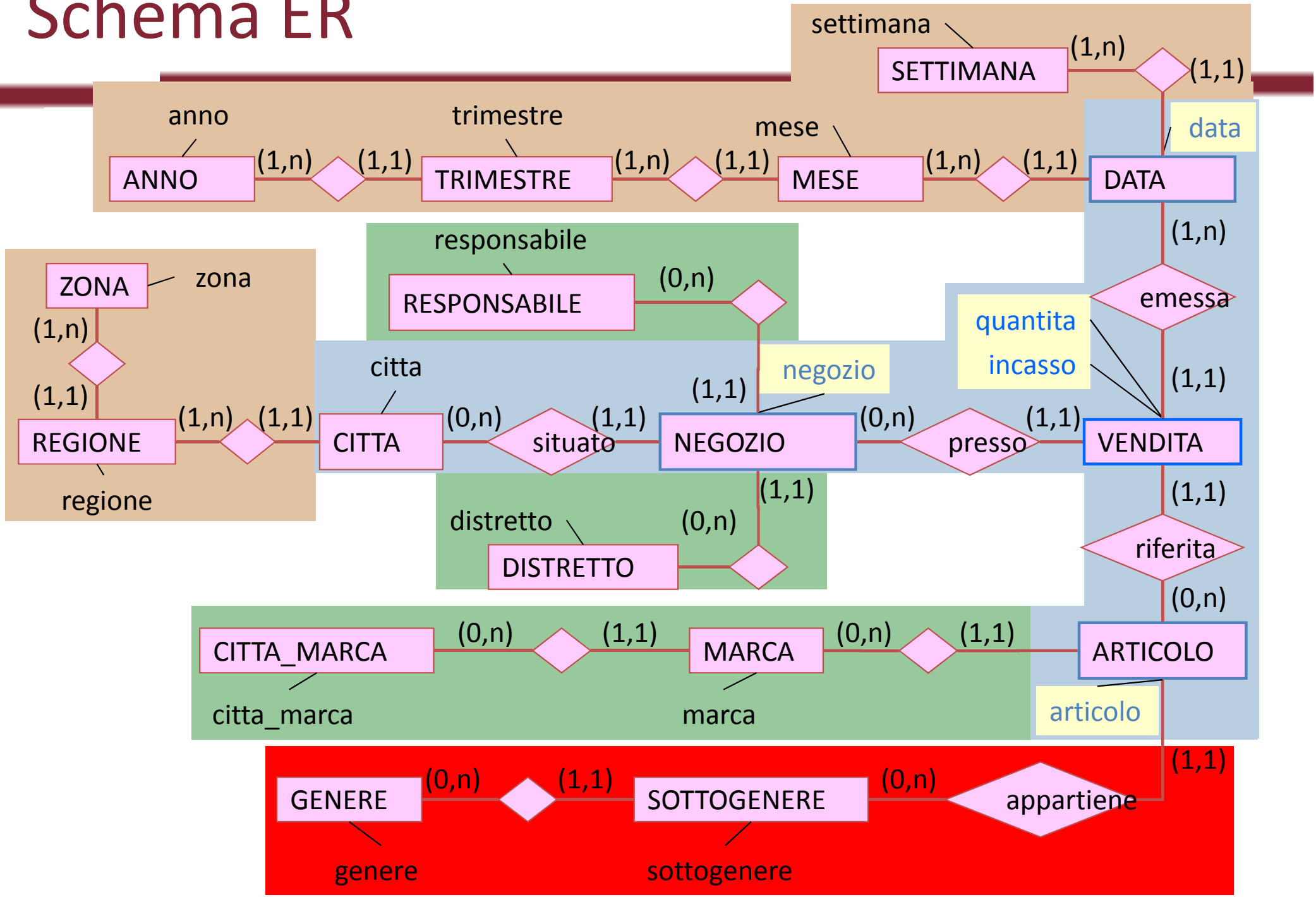


Schema di Fatto (esempio)

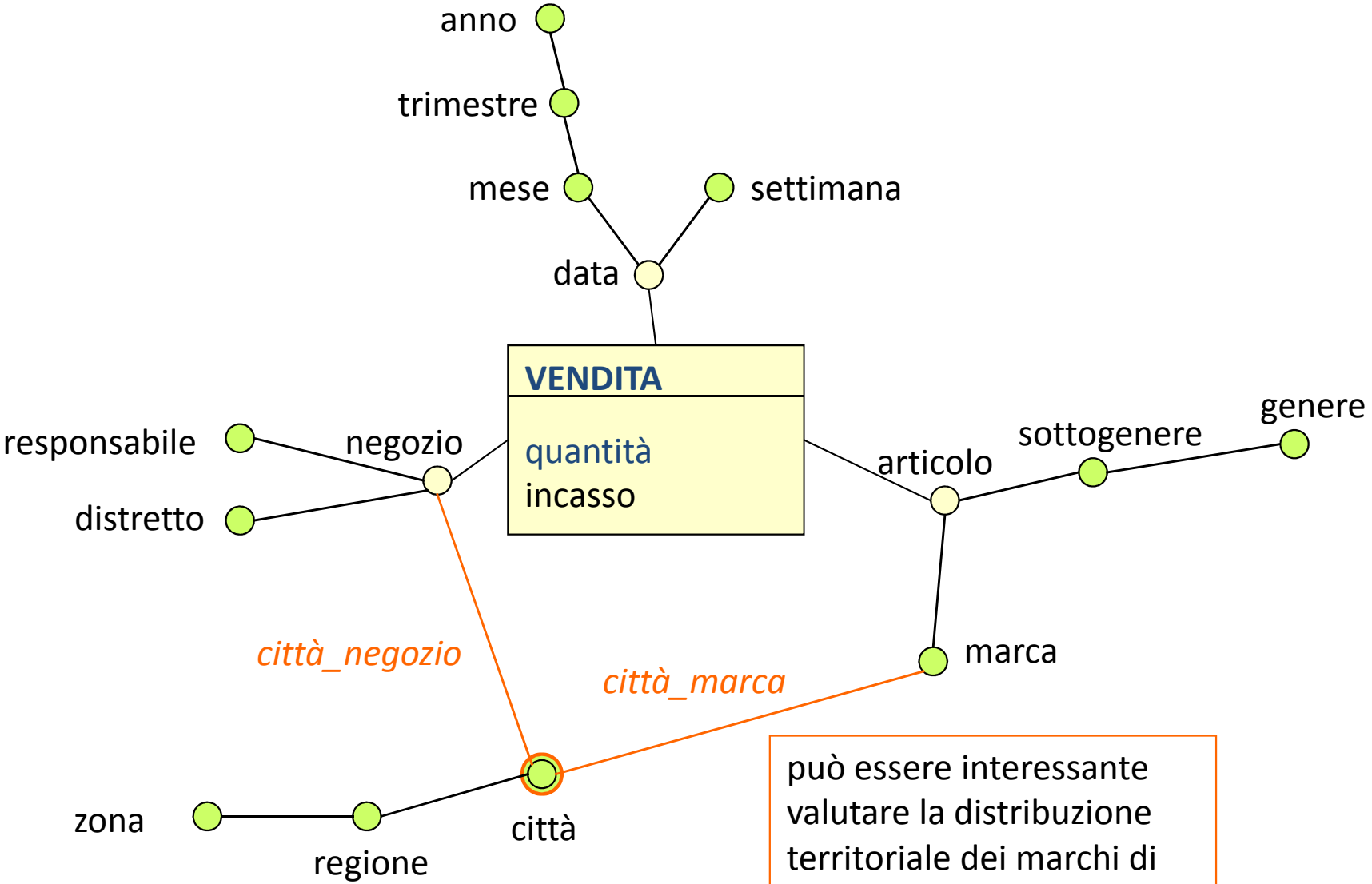
Es: vendite mensili
per città e per marca



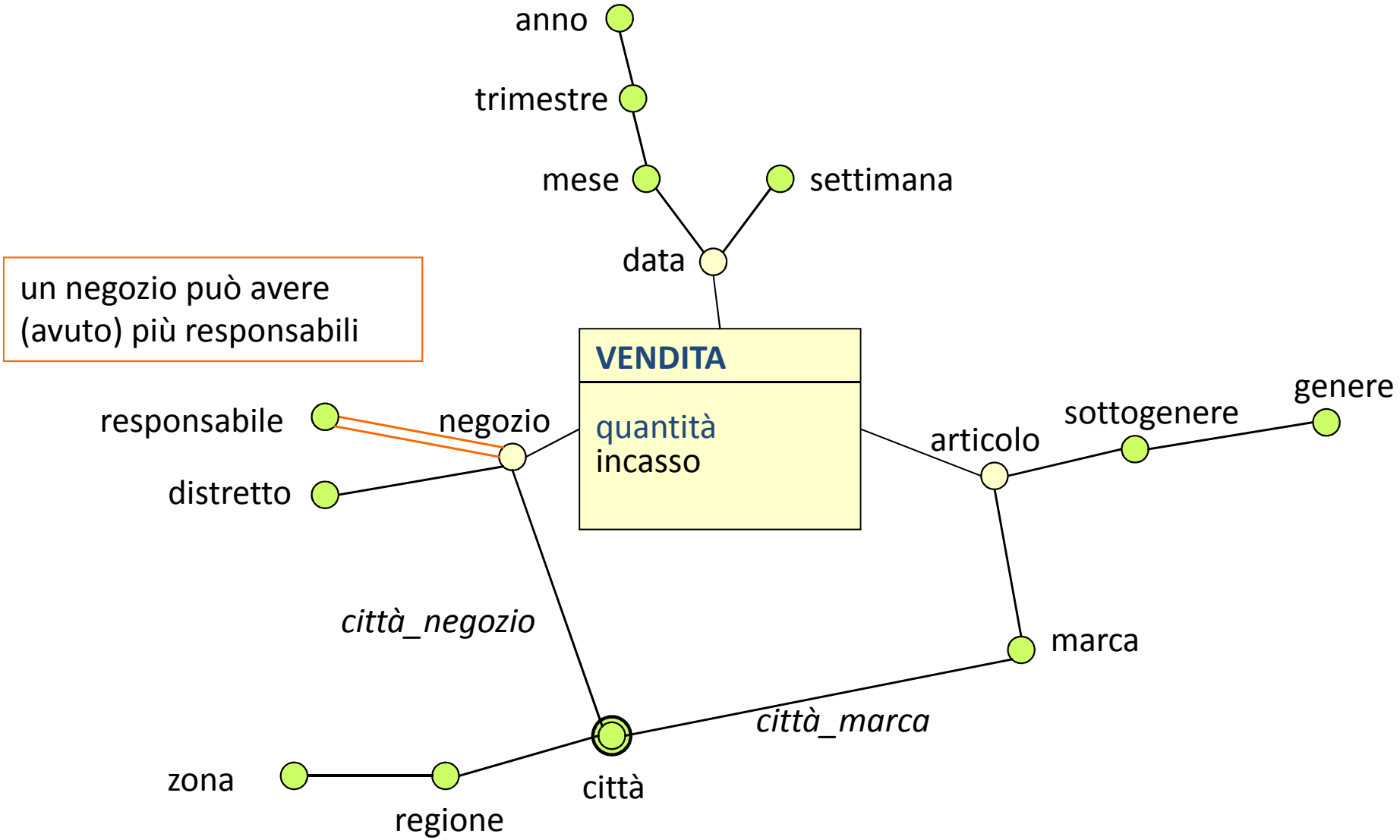
Schema ER


settimana
SETTIMANA
(1,n)
(1,1)
data
anno
trimestre
mese
ANNO
(1,n)
(1,1)
TRIMESTRE
(1,n)
(1,1)
MESE
(1,n)
(1,1)
DATA
(1,n)
responsabile
RESPONSABILE
(0,n)
emessa
ZONA
zona
(1,n)
quantita
incasso
negozio
citta
(1,1)
(1,1)
(1,1)
REGIONE
(1,n)
(1,1)
CITTA
(0,n)
situato
(1,1)
NEGOZIO
(0,n)
presso
(1,1)
VENDITA
regione
(1,1)
distretto
(0,n)
DISTRETTO
(1,1)
riferita
(0,n)
CITTA_MARCA
(0,n)
(1,1)
MARCA
(0,n)
(1,1)
ARTICOLO
citta_marca
marca
articolo
(1,1)
GENERE
(0,n)
(1,1)
SOTTOGENERE
(0,n)
appartiene
genere
sottogenere

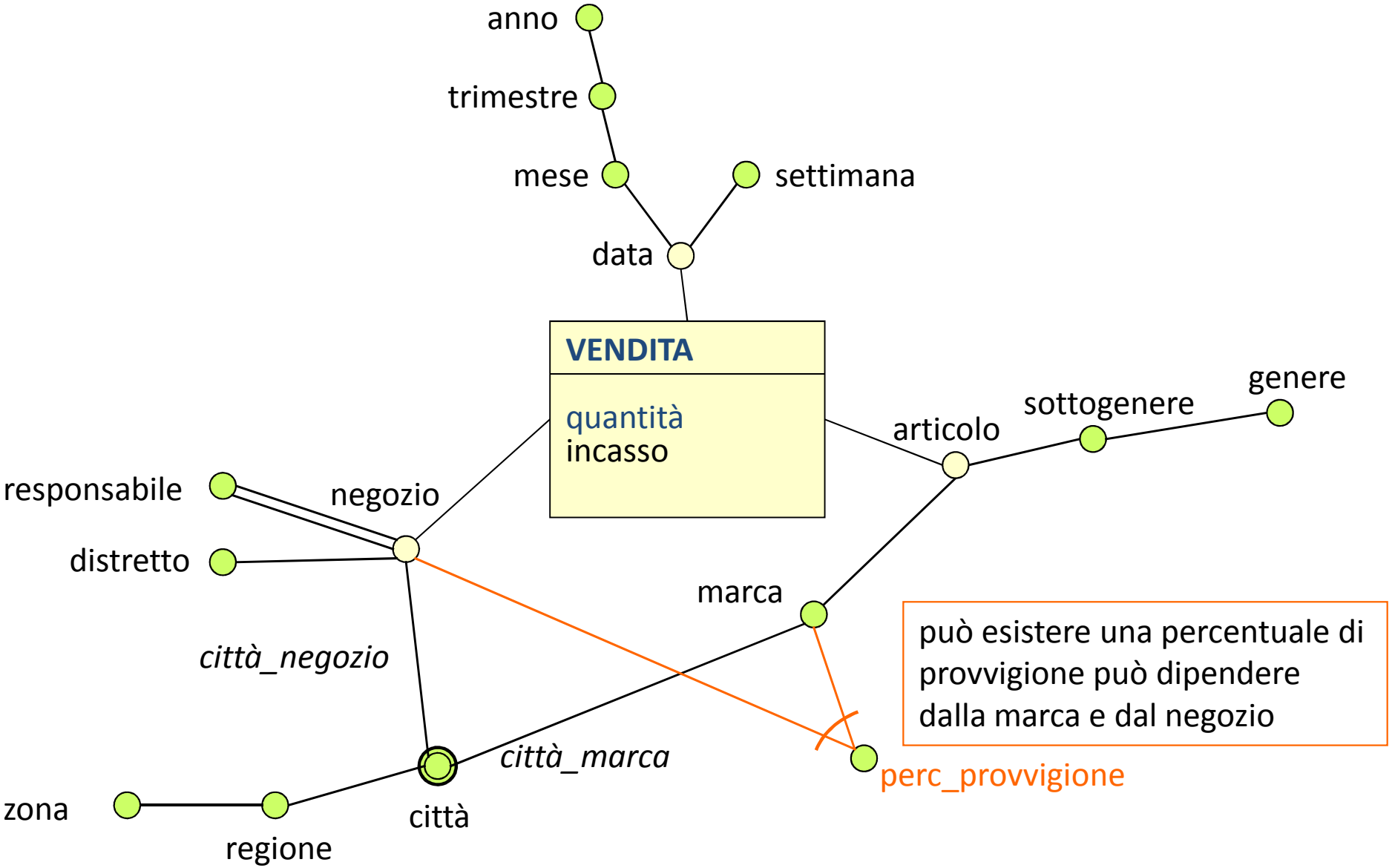
Gerarchie condivise e ruoli



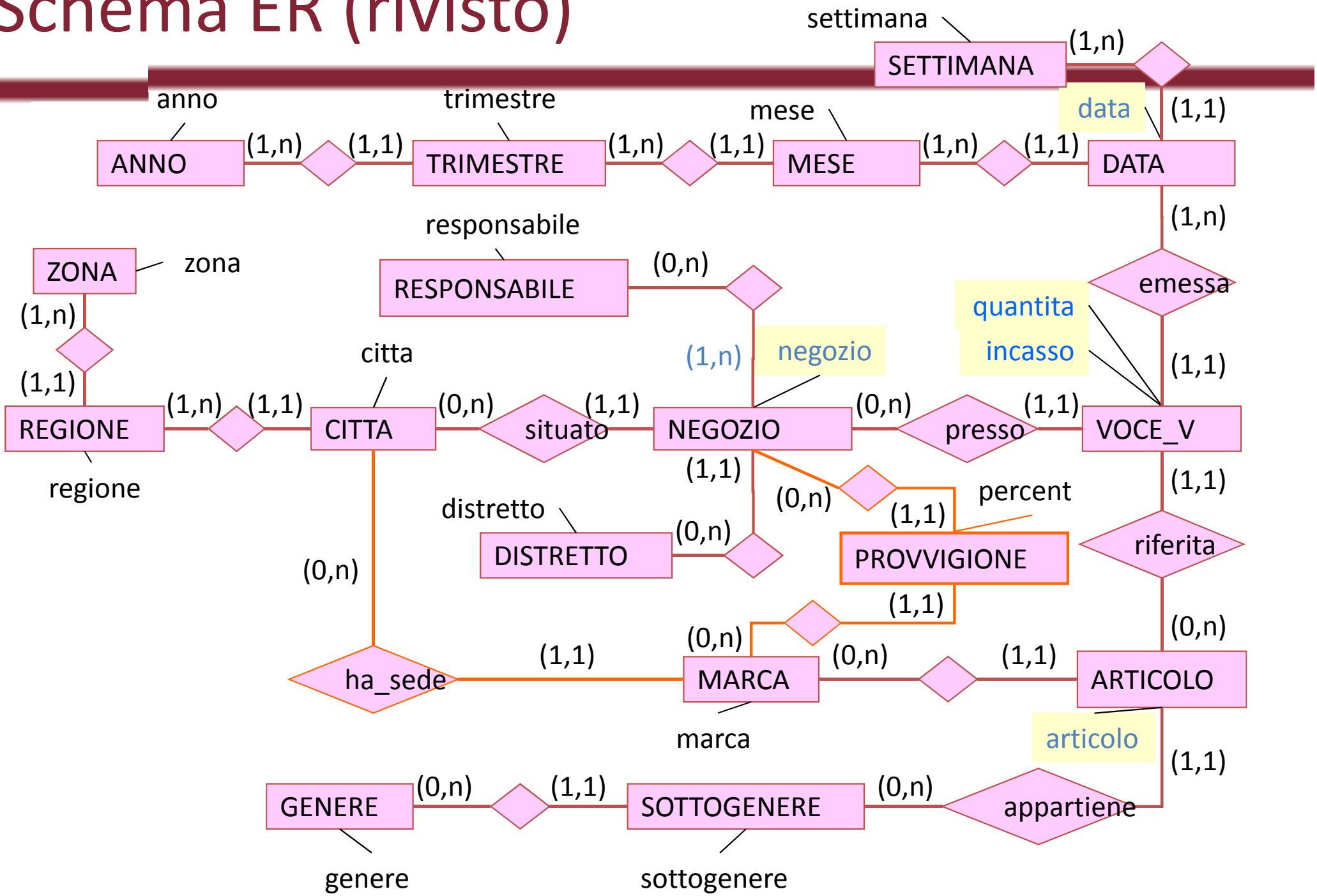
Archi multipli (relazioni n:n)



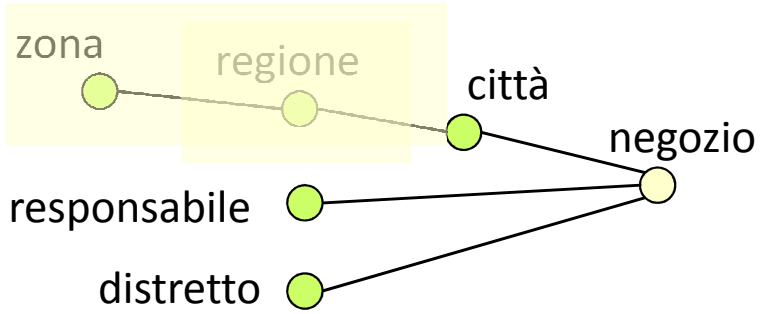
Attributi cross-dimensionali



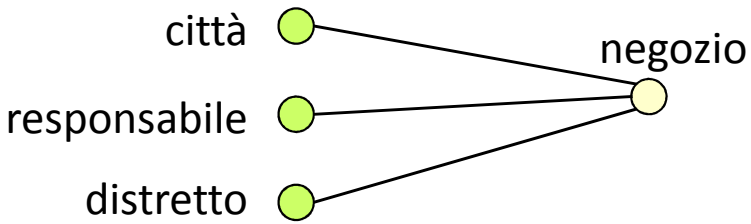
Schema ER (rivisto)



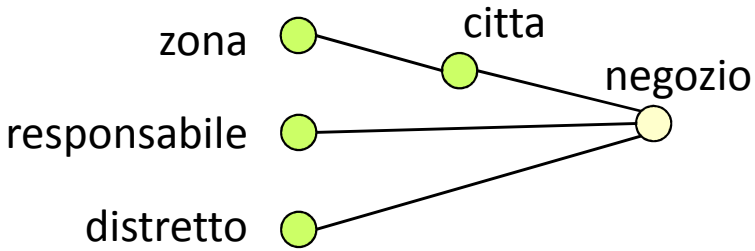
Manipolazione delle gerarchie



potatura



innesto



Alternative di rappresentazione per DW: *OLAP

ROLAP - Relational On-Line Analitical Processing

- dati su DBMS relazionale
- accesso indicizzato

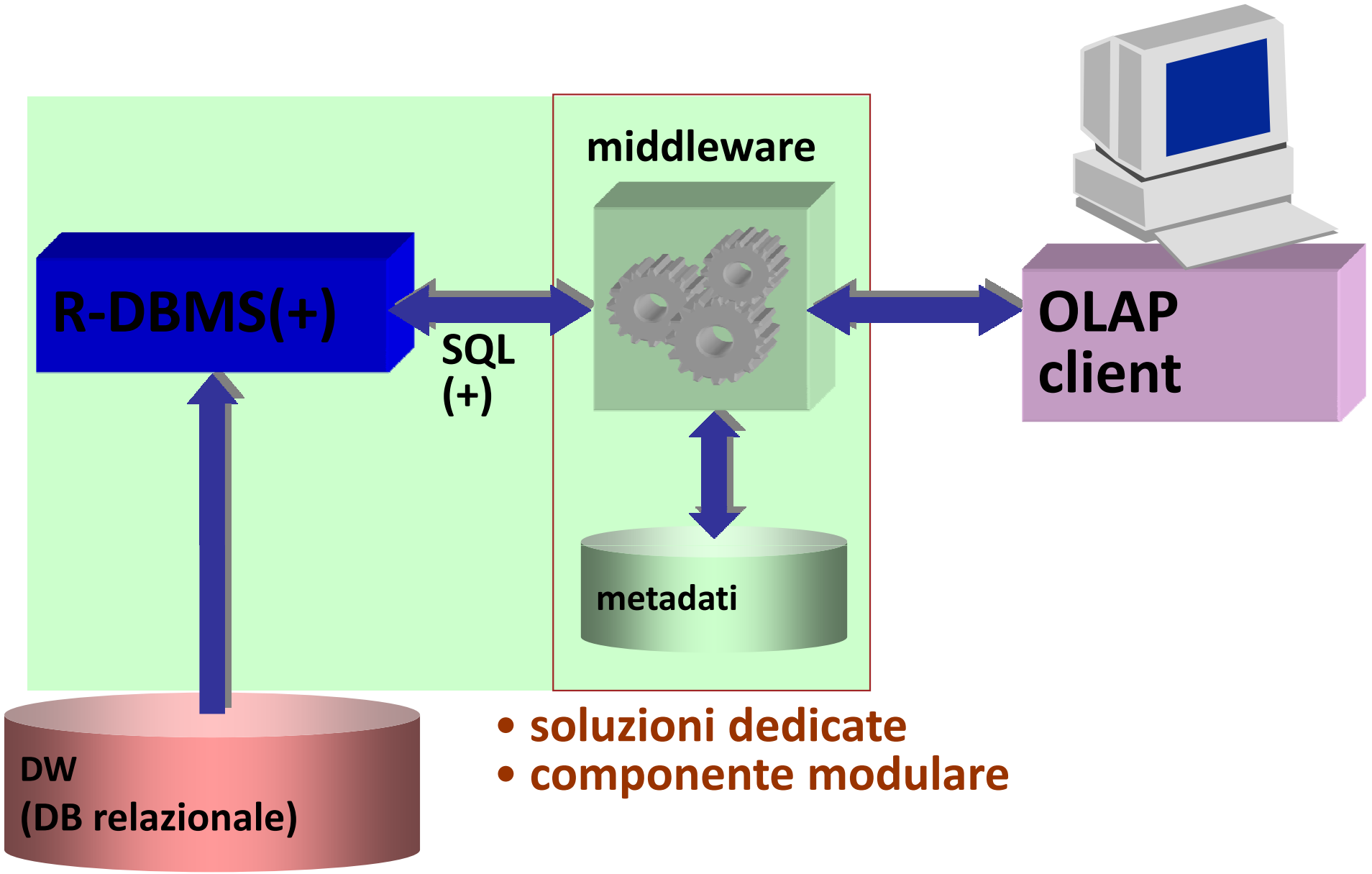
MOLAP - Multidimensional On-Line Analitical Processing

- dati su strutture multidimensionali
- accesso calcolato

HOLAP - Hybrid On-Line Analitical Processing

- dati su strutture di entrambe le tipologie
- introdotta da Oracle (Express Server, 2002)

Architettura ROLAP



- soluzioni dedicate
- componente modulare

Modello Logico MOLAP

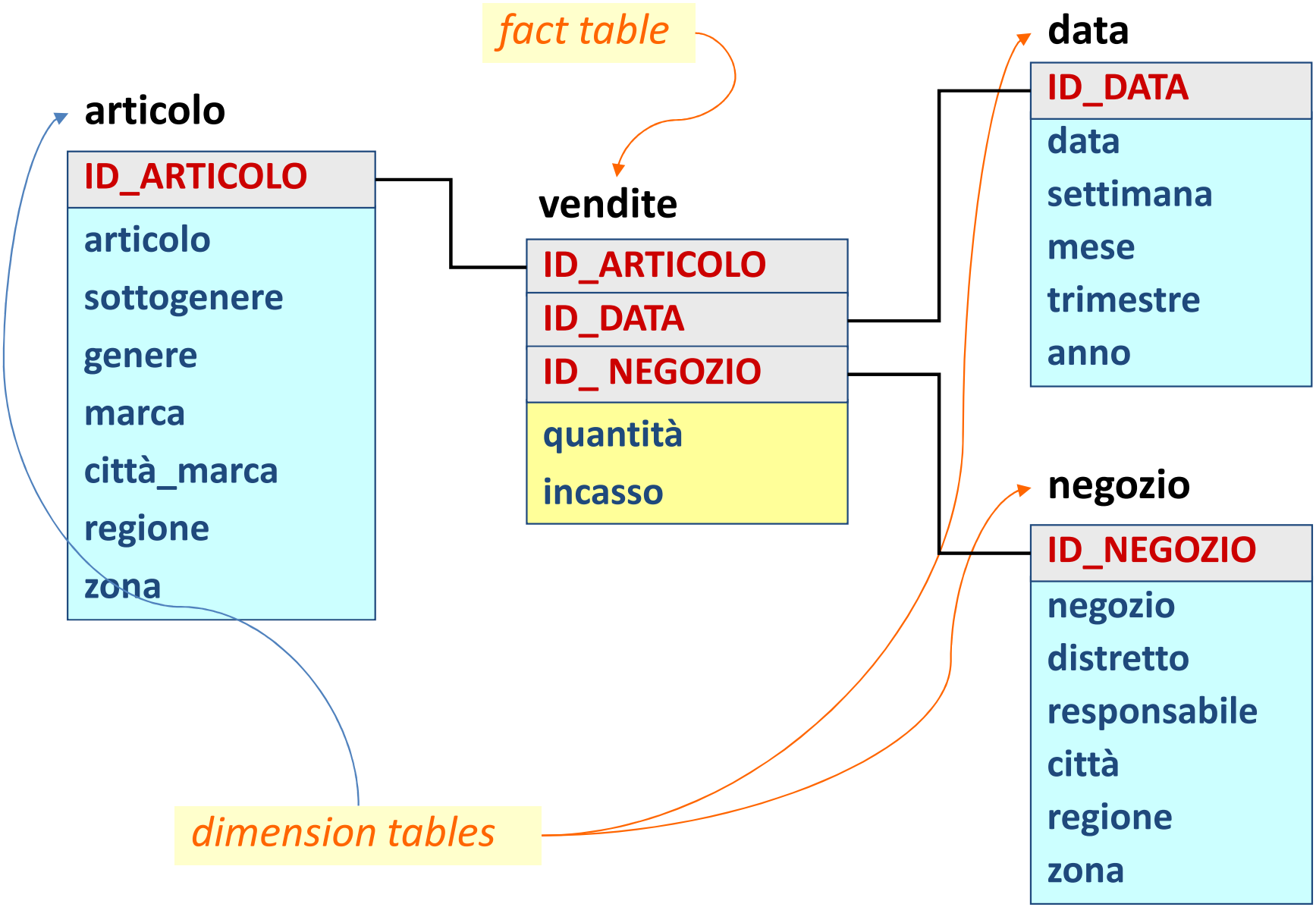
- mancanza di uno standard affermato sia per le strutture dati che per i linguaggi di accesso
- gestione della “sparsità” dei dati (frazione popolata del cubo multidimensionale)
  - elementi significativi individuati in base ad offset (collezione degli indici degli elementi non nulli)
  - partizionamento in cubi più piccoli a densità quasi uniforme (densi o molto sparsi)
  - strutture dati ad hoc (es.: kd-trees)

Modello logico (ROLAP): STAR-SCHEMA

- una *DIMENSION TABLE* per ciascuna dimensione:
  - chiave primaria (solitamente una *chiave surrogata*)
  - un insieme di attributi che descrivono i valori per tutti i livelli di aggregazione
- una singola *FACT TABLE*:
  - chiave primaria: una *foreign-key* per ciascuna delle *dimension tables*
  - un attributo per ciascuna *misura*

- Completa DENORMALIZZAZIONE (a parte la fact table)

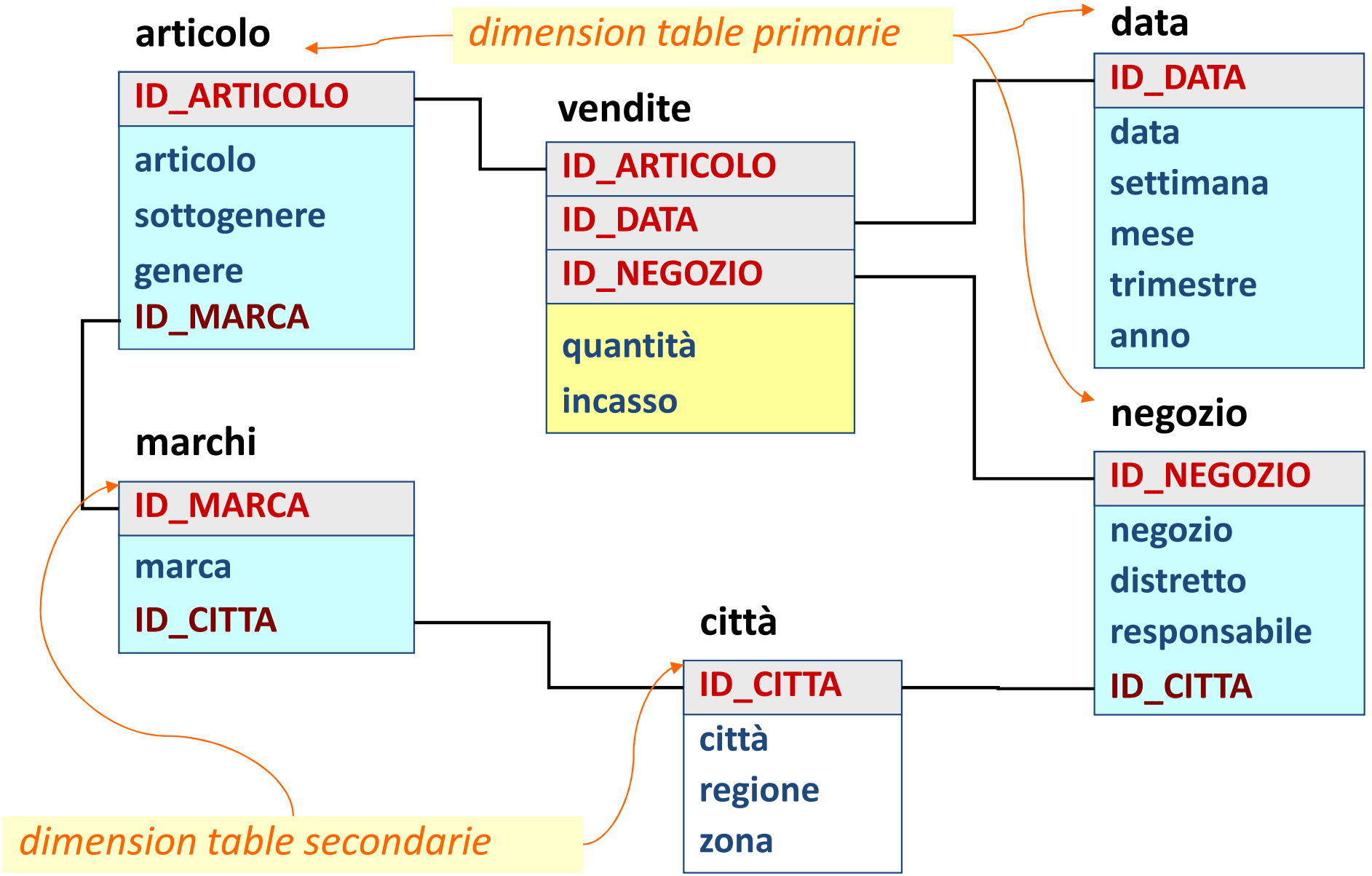
Esempio di STAR-SCHEMA



Modello logico (ROLAP): SNOWFLAKE

- A partire dallo STAR-SCHEMA, si opera una NORMALIZZAZIONE (parziale) delle *dimension tables*, ottenendo:
- per ciascuna dimensione, la singola *dimension table primaria* nello Star-Schema può essere decomposta dando luogo ad una collezione di *dimension table secondarie*
- una singola *fact table*:
  - chiave primaria: una *foreign-key* per ciascuna delle dimensioni (e per ciascuna *dimension table primaria*)
  - un attributo per ciascuna *misura*

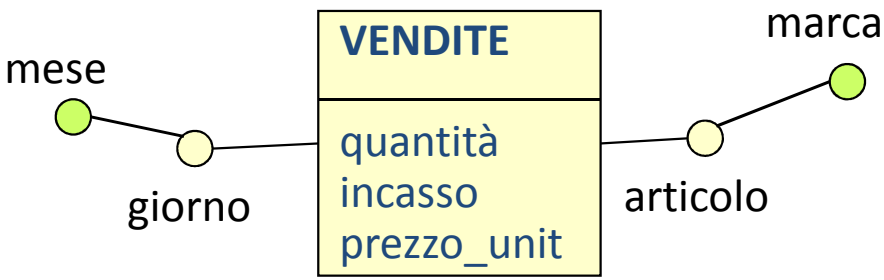
Esempio di schema SNOWFLAKE

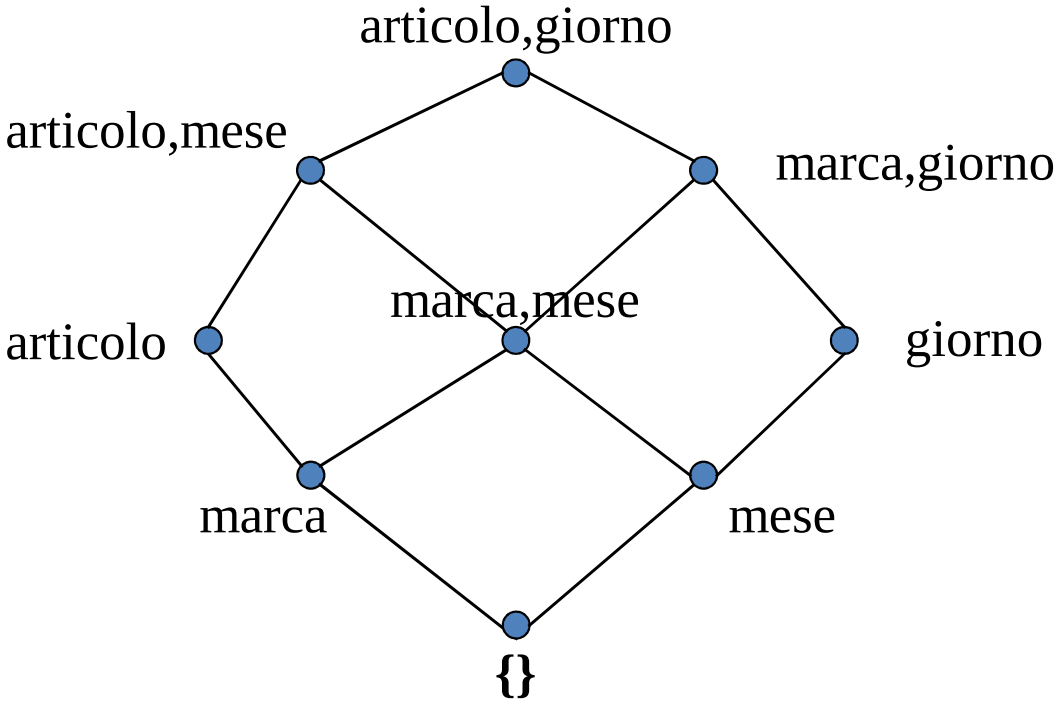


Progetto Fisico e VISTE

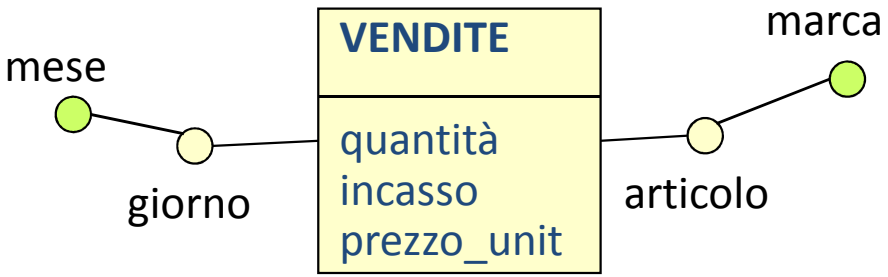
- Il principale problema operativo in un Data Warehouse è quello delle prestazioni
- Per contro, la ridondanza non costituisce un grave problema, a causa della essenziale staticità del DW
- Per conseguire migliori prestazioni, si opera una parziale materializzazione delle viste sulla Fact Table
- La contropartite legate alla materializzazione di viste sono:
  - spazio aggiuntivo (dati completamente ridondanti)
  - tempo di calcolo al momento del refresh del DW

Reticolo delle Viste

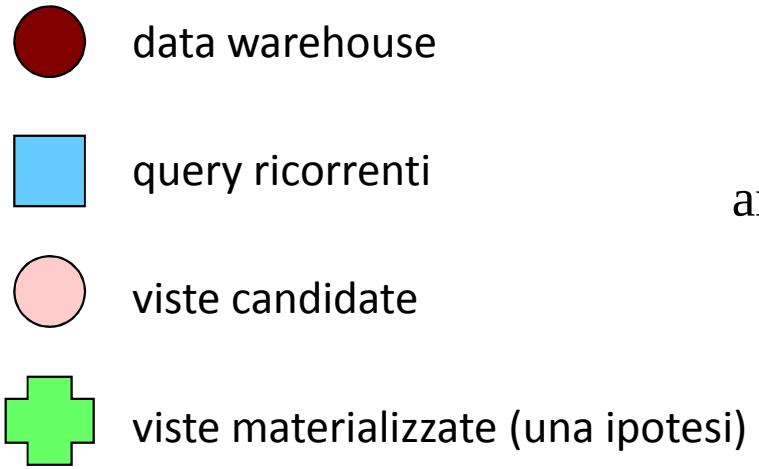


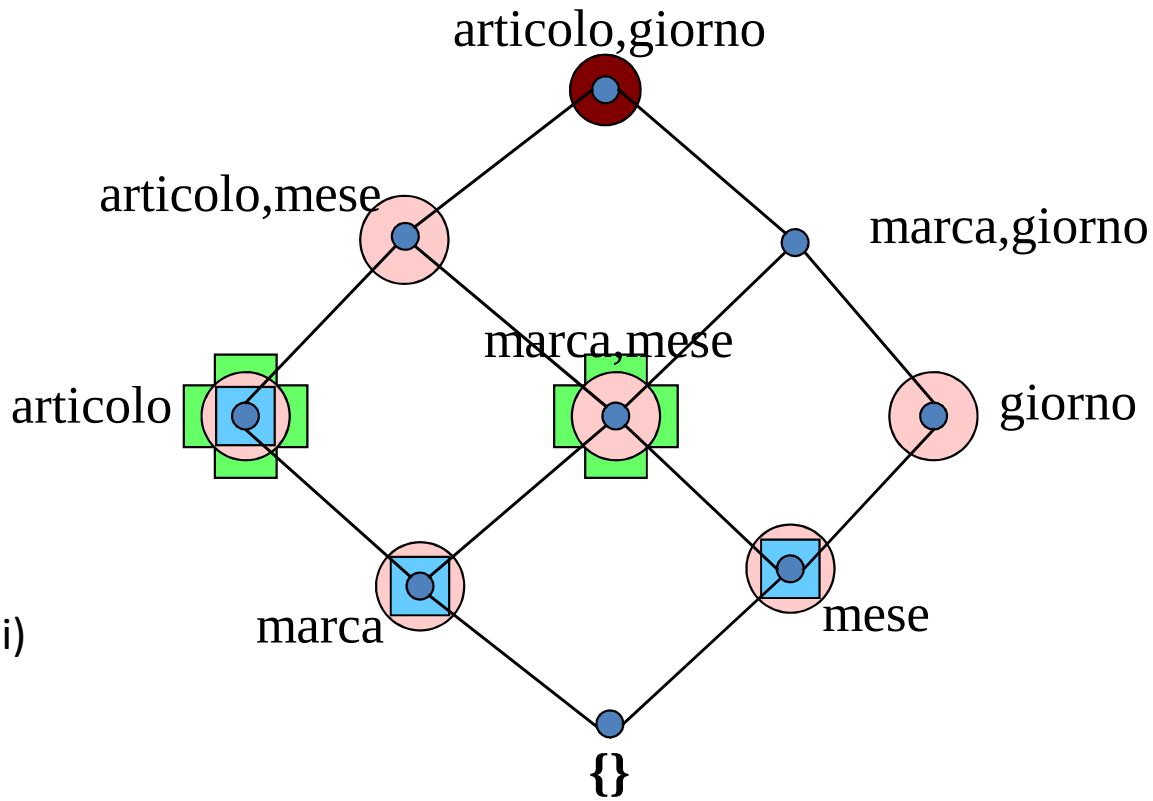


Ottimizzazione del calcolo basata sulle viste materializzate



FATTORI DI COSTO:
- **tempo di calcolo**
- **spazio**
- **tempo di refresh**





Bibliografia

M. Golfarelli, S. Rizzi.

Data Warehouse

Teoria e Pratica della Progettazione (2ª ed.)

McGraw-Hill, 2006.