

# **Sistemi informativi aziendali ERP e sistemi di data analysis**

## **Data Warehousing**

Maurizio Pighin, Anna Marzona

Copyright © 2018 Pearson Italia



- Bill Inmon (seconda metà anni '80)
  - “[...] collezione di dati, a supporto del processo decisionale manageriale orientata al soggetto, integrata, non volatile e dipendente dal tempo”
- IBM System Journal (primi anni '90)
  - “Un singolo, completo e consistente deposito di dati, ottenuti da diverse fonti e resi disponibili agli utenti finali, in maniera tale da poter essere immediatamente fruibili”



# Data warehouse e metodologia OLAP

- OLAP: On Line Analytical Processing
  - *Identifica strumenti interattivi orientati a semplificare il processo decisionale aziendale*
- Caratteristiche richieste ai sistemi per l'analisi dei dati (FASMI - OLAP Report 1995)
  - *Velocità di risposta (Fast)*
  - *Analiticità (Analytical)*
  - *Condivisione delle informazioni (Shared)*
  - *Multidimensionalità (Multidimensional)*
  - *Informatività (Informational)*



# Architettura dei sistemi di data warehousing

- Il sistema è costituito da basi di dati poste a livelli distinti, diverse per finalità, struttura e tipologia di dati contenuti
  - *Sorgenti*
    - basi di dati origine (operazionali o esterne)
  - *Staging Area (opzionale)*
    - area intermedia utilizzata come appoggio per le procedure di trasformazione dei dati
  - *Data warehouse*
    - base di dati centrale; contiene tutti i dati necessari all'analisi articolati su un modello unificato concettualmente multidimensionale
  - *Data mart*
    - basi di dati multidimensionali su cui si appoggia l'analisi



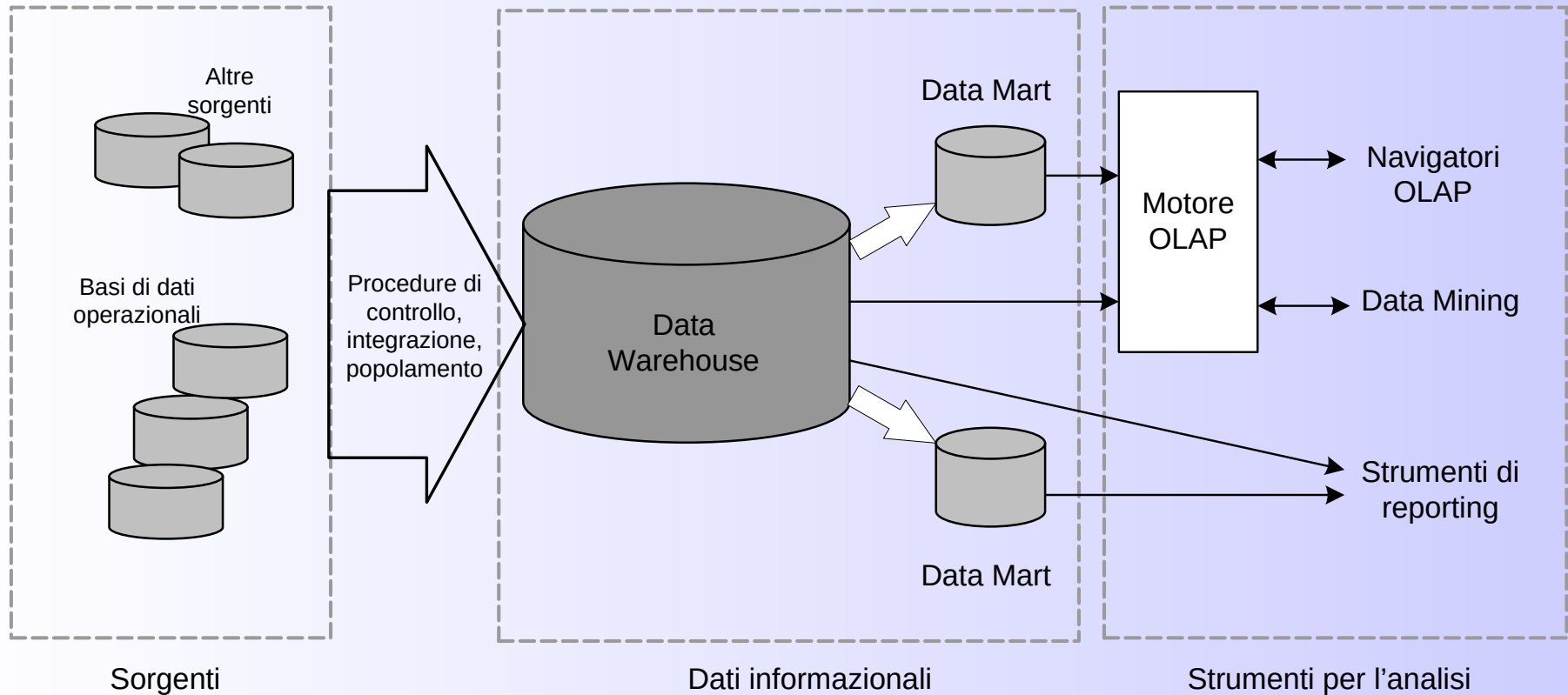
# Architettura dei sistemi di data warehousing

- Architetture a due livelli
  - *Sorgenti, Data warehouse, Data mart*
- Architetture a tre livelli
  - *Comprendono anche l'area di trasformazione dei dati (staging area)*
- Appartengono al sistema
  - *Procedure per il trasferimento dei dati tra le diverse basi di dati*
  - *Strumenti per l'analisi dei dati*



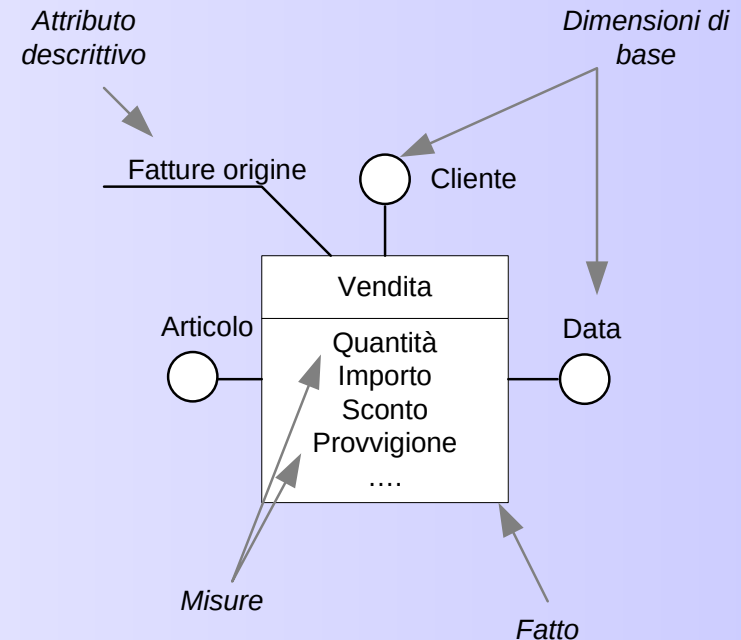
# Architettura dei sistemi di data warehousing

Sistemi informativi aziendali  
ERP e sistemi di data analysis  
*Cap.12 - Data Warehousing*  
Maurizio Pighin, Anna Marzona



# Modelli concettuali per il data warehouse: il DFM

- Il DFM (Dimensional Fact Model) descrive graficamente i fatti attorno a cui si struttura un data warehouse
  - Ogni fatto è rappresentato tramite uno schema di fatto
- Schema di fatto
  - *Fatto*
    - rettangolo contenente il nome del fatto e le sue misure
  - *Dimensioni di base*
    - circoletti etichettati collegati al fatto



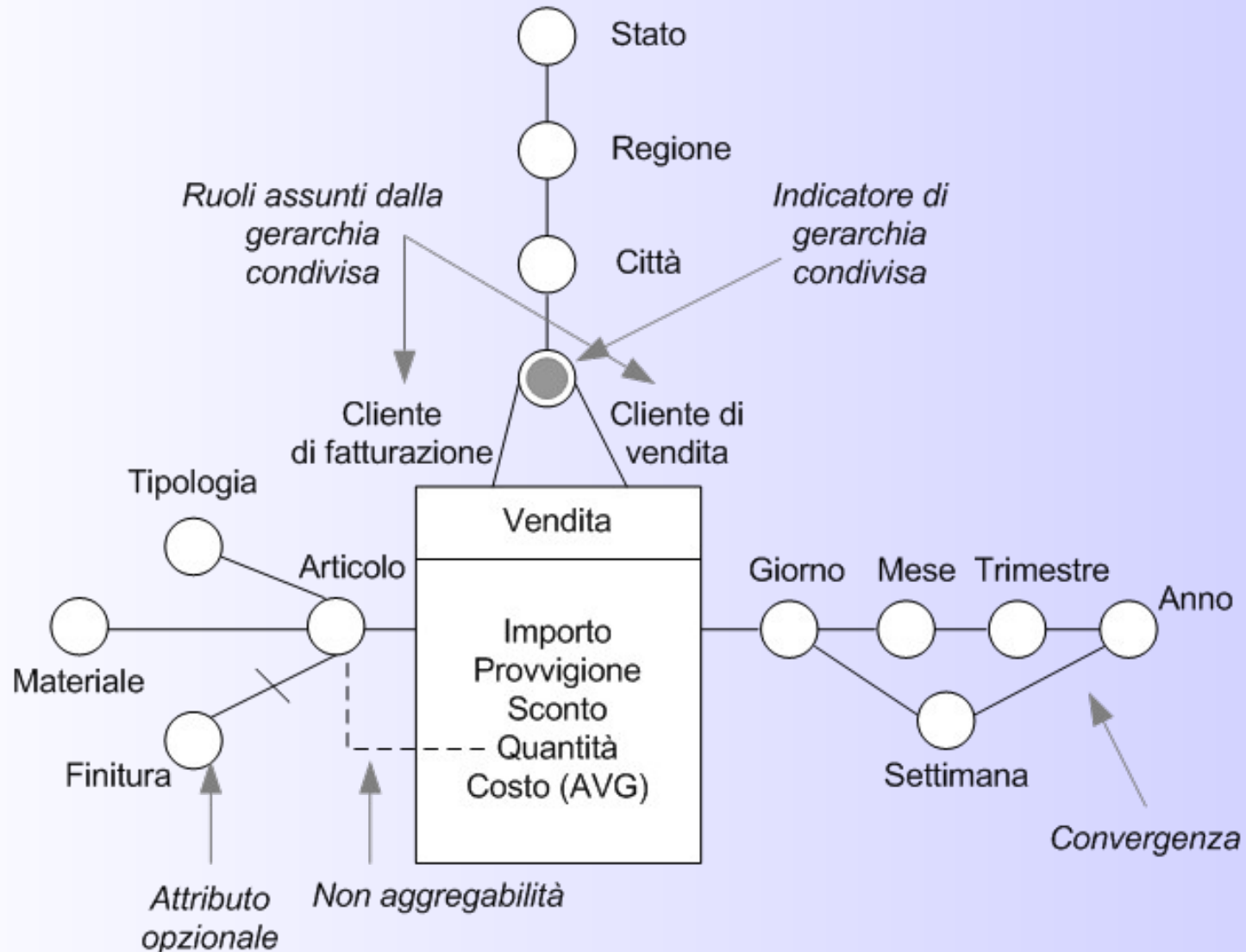
# Modelli concettuali per il data warehouse: il DFM

- Le gerarchie dimensionali sono alberi con radice nelle dimensioni di base
  - *Gli attributi dimensionali sono i nodi dell'albero*
- DFM permette di rappresentare caratteristiche proprie dei sistemi multidimensionali
  - *Opzionalità*
  - *Gerarchie condivise*
  - *Convergenze*
  - *Non aggregabilità*



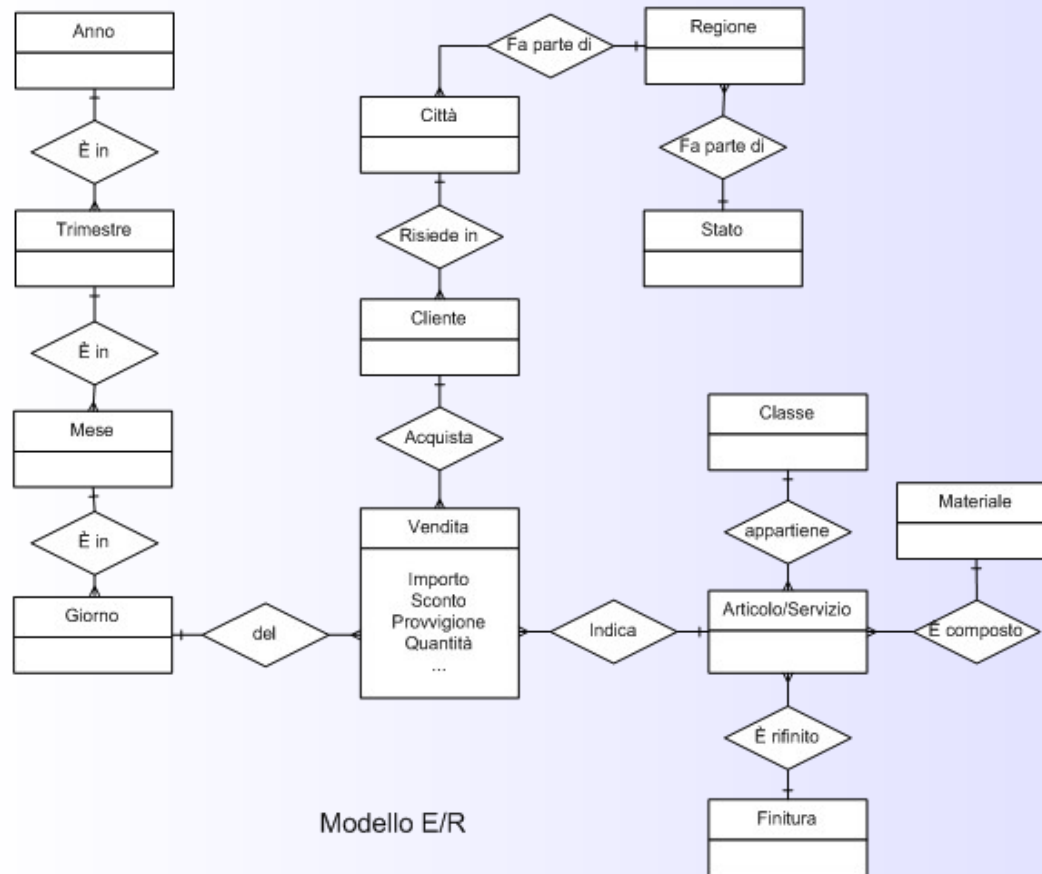


# Modelli concettuali per il data warehouse: il DFM

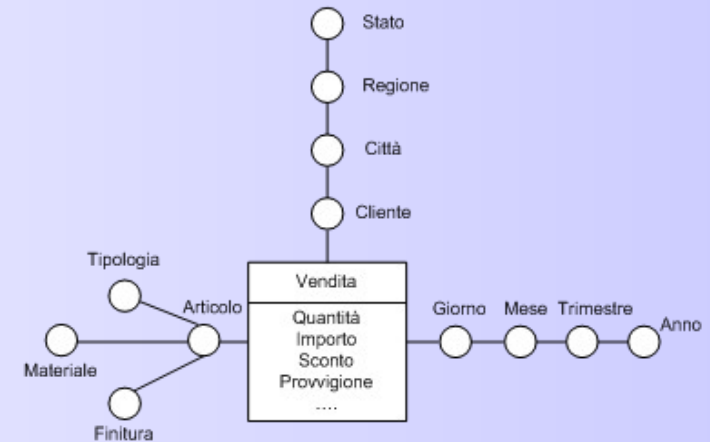


# Corrispondenze con il modello Entità-Relazione

Sistemi informativi aziendali  
ERP e sistemi di data analysis  
*Cap.12 - Data Warehousing*  
Maurizio Pighin, Anna Marzona



Modello E/R



Modello DFM



# Modelli logici per il data warehouse

- ROLAP
  - *La struttura multidimensionale dei fatti viene realizzata su database relazionale*
  - *Interrogazioni tramite query SQL standard*
  - *Vantaggi*
    - minima occupazione di spazio
    - elevata conoscenza degli strumenti relazionali da parte degli operatori
  - *Svantaggi*
    - esecuzione di query poco efficiente
    - le soluzioni per il miglioramento della velocità di risposta (denormalizzazione, materializzazione delle viste) implicano un aumento della complessità e dell'occupazione di spazio



# Modelli logici per il data warehouse

- MOLAP
  - *La struttura dei fatti viene realizzata su database multidimensionale, con accesso di tipo posizionale*
  - *Interrogazioni ottimizzate tramite strumenti di query proprietari*
  - *Vantaggi*
    - elevata efficienza nell'esecuzione delle query complesse
    - stretta aderenza al modello concettuale
  - *Svantaggi*
    - elevata occupazione di spazio (viene allocato lo spazio per ogni possibile ennupla dimensionale)
    - mancanza di standard, sia di rappresentazione dei dati che di interrogazione
    - scarsa familiarità con il modello da parte degli operatori



# Modelli logici per il data warehouse

- HOLAP
  - *Soluzione intermedia che combina i vantaggi di MOLAP e ROLAP*
  - *Data warehouse: realizzato su base relazionale*
    - semplicità di sviluppo e di manutenzione delle procedure di popolamento dei fatti
    - scalabilità del sistema
  - *Data mart: realizzati su base multidimensionale*
    - efficienza nelle interrogazioni
    - dimensioni contenute

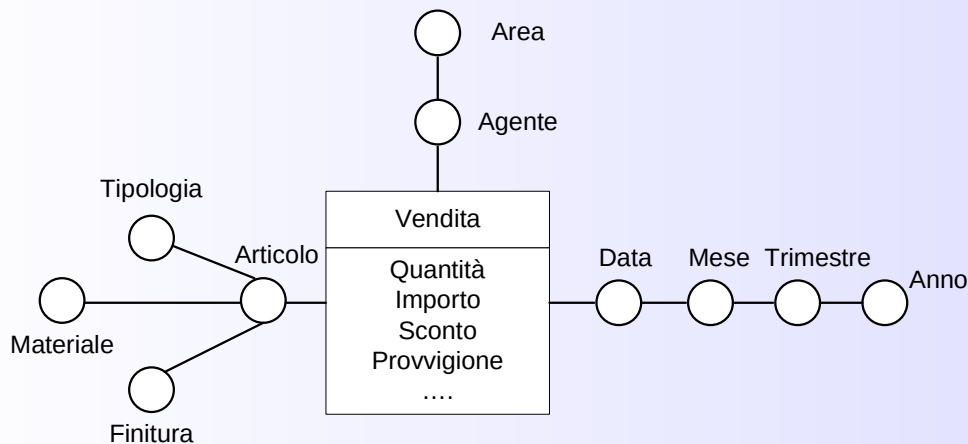


# Schemi multidimensionali su basi di dati relazionali

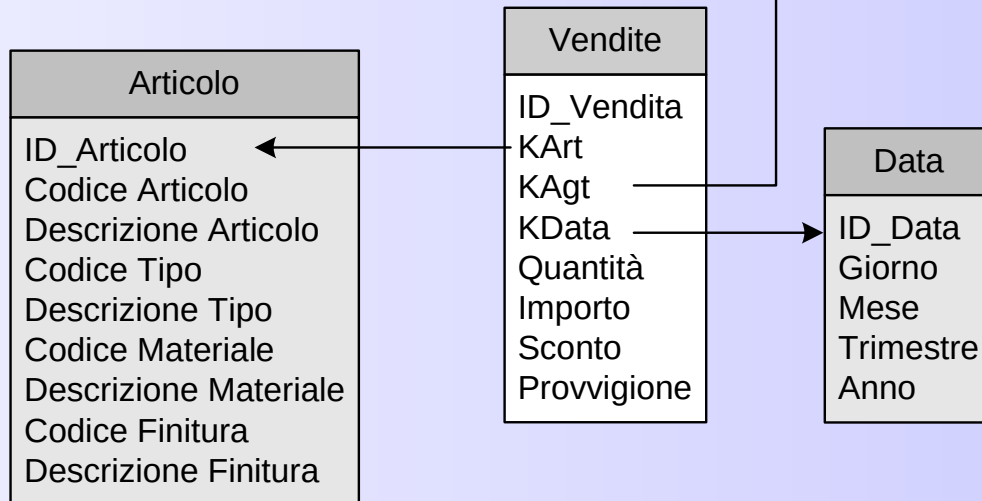
- Schema a stella
  - *Tabella dei fatti*
    - una tabella per ogni fatto
    - un campo per ogni misura ed una chiave esterna per ogni dimensione di base
  - *Tabelle delle dimensioni*
    - una per ogni dimensione di base
    - un campo per ogni attributo dimensionale della gerarchie che ha radice nella dimensione rappresentata (denormalizzazione completa)
  - *Vantaggi*
    - massima velocità nel reperimento delle informazioni
  - *Svantaggi*
    - ridondanza, spazio occupato, scarsa intuitività della struttura, elevata complessità di aggiornamento



# Schema a stella



## Modello concettuale



## Modello logico su schema a stella



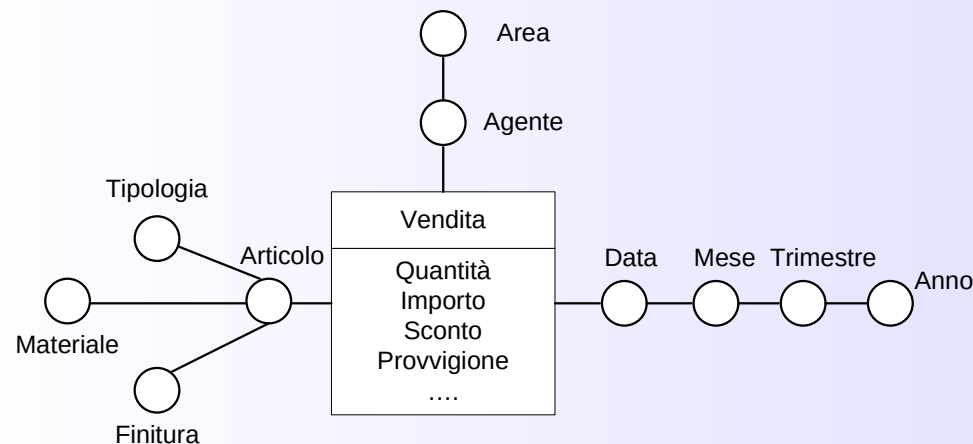
# Schemi multidimensionali su basi di dati relazionali

- Schema a fiocco di neve
  - *Riduce la denormalizzazione delle tabelle delle dimensioni esplicitando alcune gerarchie*
  - *Vantaggi*
    - chiara separazione logica sui soggetti, migliori prestazioni nel caso di materializzazione di viste, minor sensibilità alle variazioni logiche delle gerarchie nel tempo
  - *Svantaggi*
    - velocità di risposta alle interrogazioni minore rispetto allo schema a stella
- Costellazione
  - *Tabelle dimensionali condivise da più tabelle dei fatti*
  - *Approccio da seguire quando più fatti coinvolgono gli stessi soggetti*

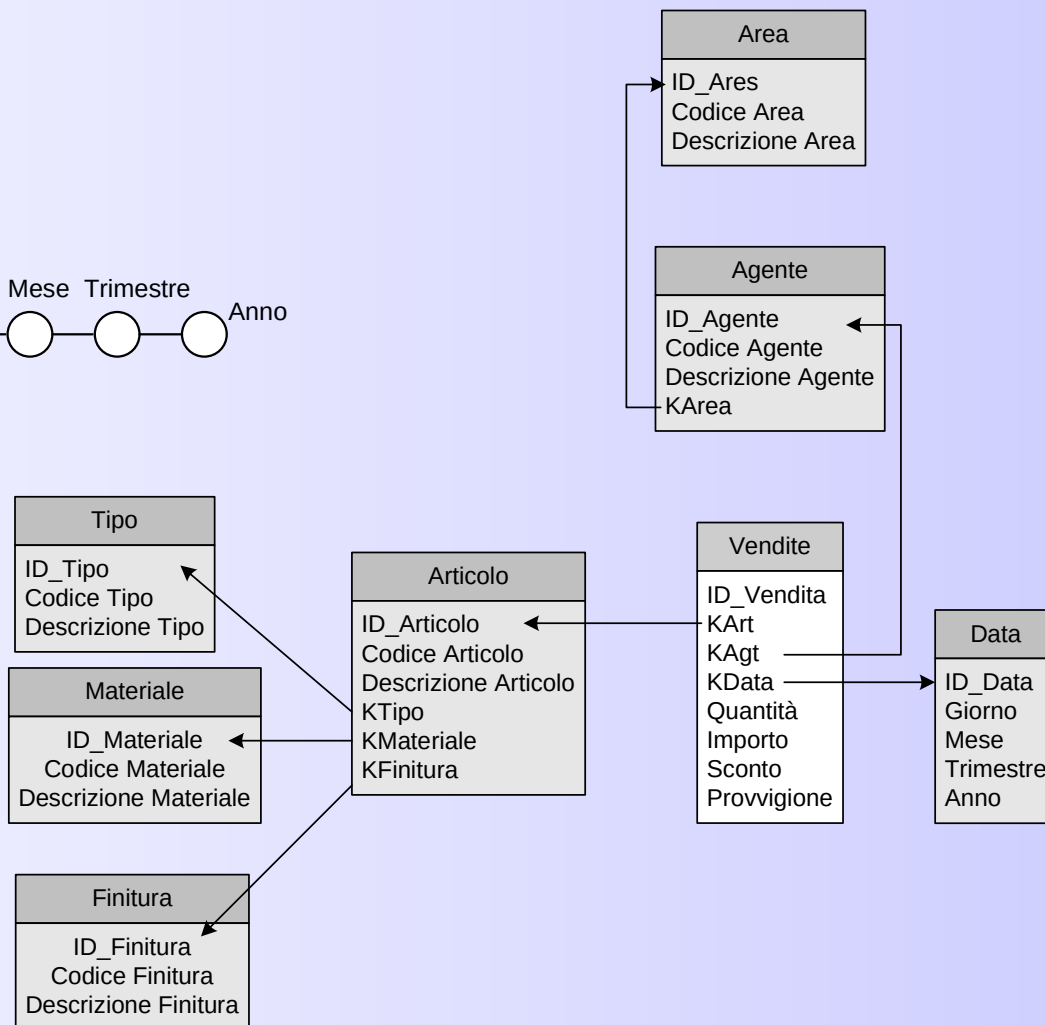




# Schema a fiocco di neve



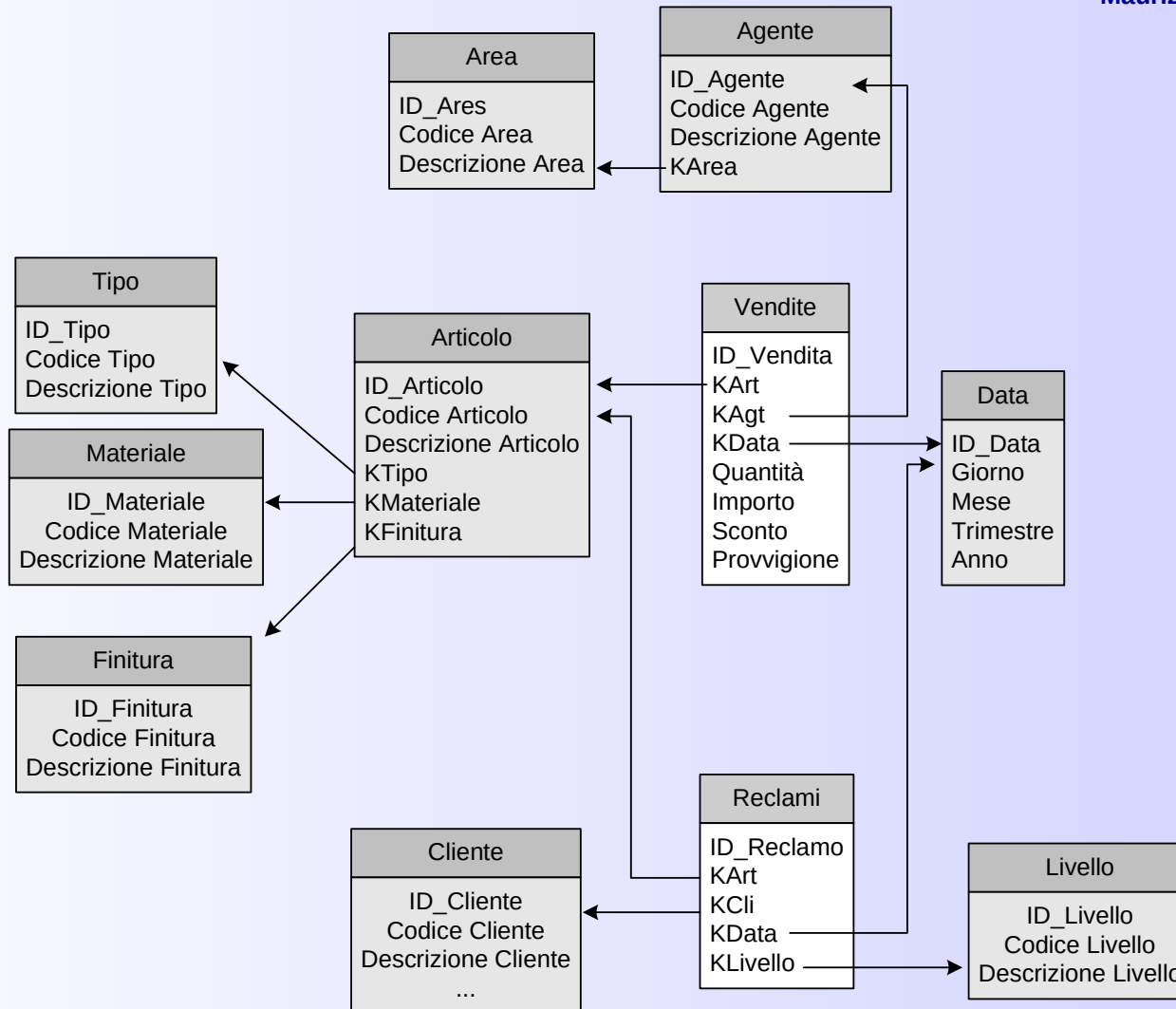
**Modello concettuale**



**Modello logico su schema a fiocco di neve**



# Costellazione di fatti

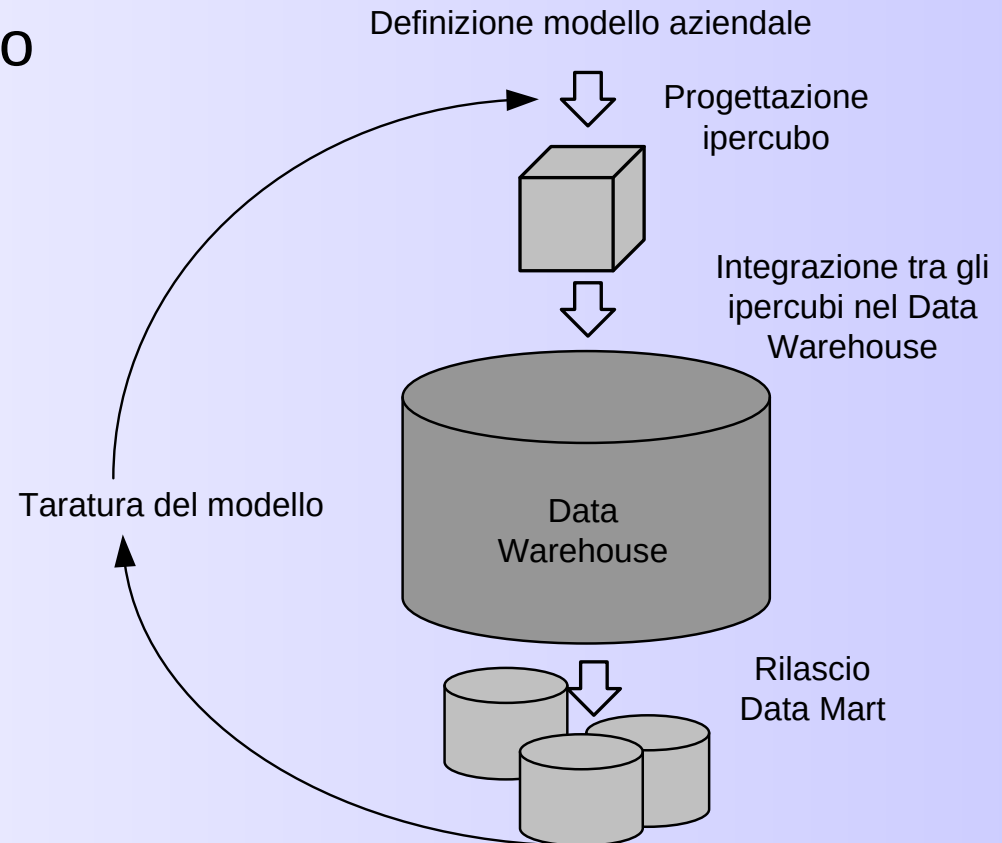


Costellazione tra Vendite e Reclami



# Il ciclo di vita dei sistemi di data warehousing

- La costruzione avviene con un approccio iterativo
  - *Costruzione del primo ipercubo relativamente al fatto più significativo*
  - *Integrazione progressiva degli altri fatti*
  - *Rilascio di data mart*
- Vantaggi
  - *Primi risultati disponibili in breve tempo*
  - *Investimenti diluiti*
  - *Possibilità di tarare e di sviluppare il modello sulla base delle indicazioni emerse dall'uso effettivo*



# Costruzione di un data mart

- Analisi delle sorgenti
  - *Descrizione dei dati disponibili*
  - *Verifica della compatibilità con i requisiti dell'utente*
  - *Creazione schema concettuale unico ed uniforme*
- Progettazione concettuale degli schemi di fatto
  - *Identificazione di misure, dimensioni, gerarchie dimensionali, limiti di aggregabilità delle misure per ogni fatto*
- Progettazione logica e ed implementazione fisica dei fatti nel data warehouse
  - *Uso di schemi a stella o a fiocco di neve, costruzione di viste materializzate o di ipercubi ad alto livello di aggregazione*
- Progettazione dell'alimentazione
  - *Definizione delle procedure di popolamento del data warehouse a partire dalle sorgenti*



# Popolamento del data warehouse

- Fasi di popolamento
  - *Estrazione*
    - estrae dalle sorgenti i dati da portare sul data warehouse
  - *Integrazione e trasformazione*
    - riconduce i dati estratti al modello unificato definito per il data warehouse
  - *Pulizia*
    - aumenta la qualità dei dati, riconoscendo e risolvendo errori, incongruenze ed omissioni
  - *Caricamento*
    - popola il data warehouse con i dati estratti, trasformati e ripuliti



# Popolamento del data warehouse

- Estrazione
  - *Informazioni di base*
    - quali informazioni devono essere acquisite (tabelle, campi)
    - come devono essere trattati gli eventi origine (es. aggregazione alla fonte o estrazione al dettaglio massimo)
  - *Tipi di estrazione*
    - **statica**: tratta tutti i dati presenti nelle sorgenti
    - **incrementale**: tratta i soli dati inseriti o alterati dalla data dell'ultimo popolamento del data warehouse, identificandoli tramite una delle seguenti metodologie
      - *estrazione delegata alle applicazioni (necessita di staging area)*
      - *estrazione delegata a trigger (necessita di staging area)*
      - *estrazione pilotata da timestamp*
      - *estrazione statica con successiva selezione per confronto diretto*



# Popolamento del data warehouse

- Integrazione e trasformazione
  - *Riporta i dati estratti al modello aziendale*
  - *Fasi di integrazione e trasformazione*
    - **riconciliazione** dei dati provenienti da fonti diverse riferite allo stesso soggetto
    - **riconoscimento di duplicati**
    - **trasformazione** di dati continui utilizzati come dimensioni in parametrizzazioni discrete (es. creando fasce di valori per misure come il peso o le quantità per vendita)
    - **standardizzazione**
      - *del formato*
      - *delle convenzioni*
      - *delle codifiche*



# Popolamento del data warehouse

- Pulizia
  - *Innalzamento del livello di qualità dei dati*
  - *Non è necessariamente successiva alla integrazione*
  - *Tipologie di errori trattati*
    - dati incompleti
    - dati errati o incomprensibili
    - dati inconsistenti
  - *Strumenti utilizzati per il riconoscimento e la correzione*
    - dizionari
    - regole
    - classificatori, predittori





# Popolamento del data warehouse

- Caricamento
  - *Caricamento vero e proprio dei dati sul data warehouse*
  - *Aggiornamento dall'esterno (dimensioni più esterne) all'interno (fatti), con applicazione delle politiche di aggiornamento agli elementi già esistenti*
  - *Aggiornamento dei fatti*
    - inserimento dei fatti nuovi
    - eventuale sovrascrittura degli elementi modificati



# Popolamento del data warehouse

- *Aggiornamento delle dimensioni:*
  - inserimento dei nuovi valori per le dimensioni
  - eventuale modifica dei valori presenti, secondo diverse strategie
    - *non fare nulla (ogni fatto usa gli attributi dimensionali validi all'inserimento della dimensione)*
    - *sovrascrivere (ogni fatto usa gli attributi dimensionali validi adesso)*
    - *creare una nuova istanza da associare ai fatti che si verificano da oggi in avanti (ogni fatto usa gli attributi dimensionali validi all'epoca)*
    - *creare una nuova istanza con marcatori temporali (massima flessibilità)*

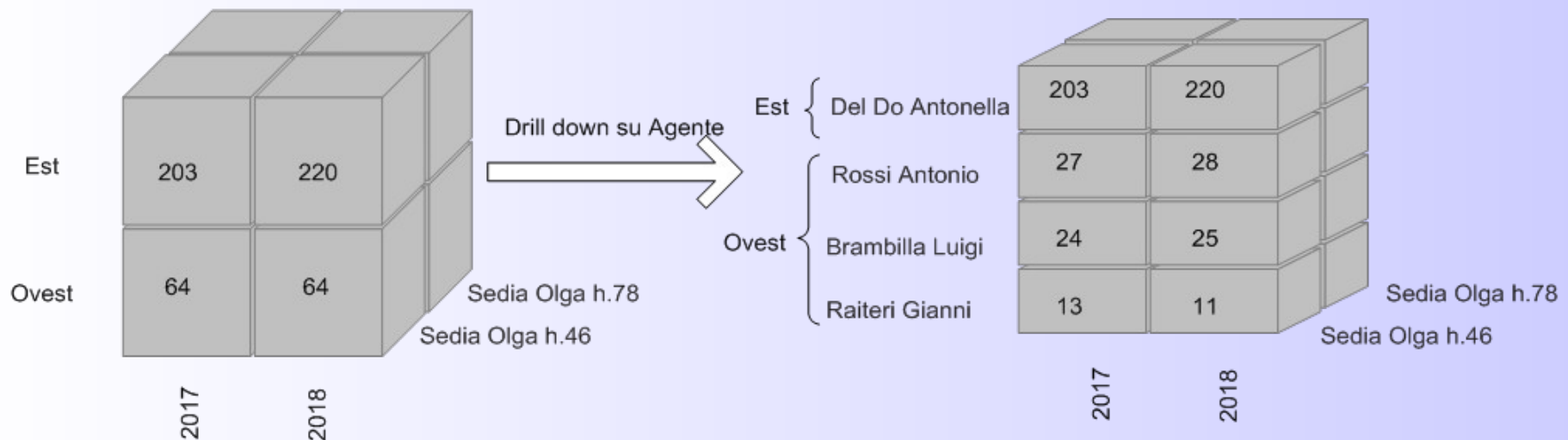


- Navigazione interattiva sui dati multidimensionali
- Esplorazione guidata da ipotesi
- Sessione di analisi complessa
  - *Ciascun passo è conseguenza dei risultati ottenuti al passo precedente*
  - *Le interrogazioni operano per differenza rispetto all'interrogazione precedente*
- Passo di navigazione
  - *Applicazione di un operatore OLAP all'insieme di dati estratto al passo precedente*
- Risultati presentati in forma tabellare o grafica



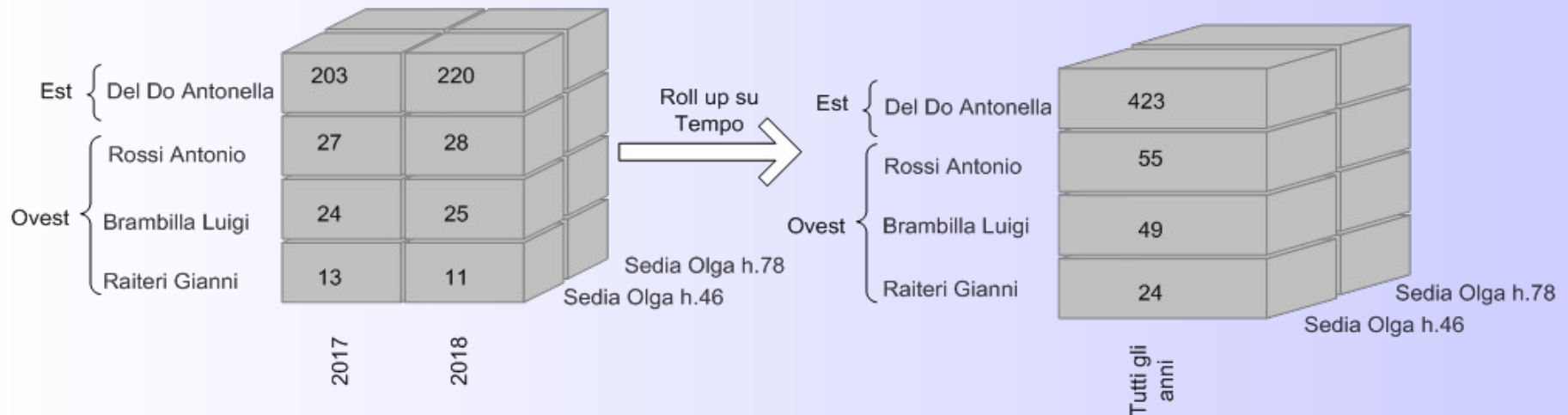
# Operatori OLAP: Drill down

- Dettaglia i dati
  - Scendendo lungo una gerarchia
  - Aggiungendo una dimensione di analisi



# Operatori OLAP: Roll up

- Sintetizza i dati
  - Percorrendo le gerarchie nella direzione di maggior aggregazione
  - Eliminando una delle dimensioni di analisi



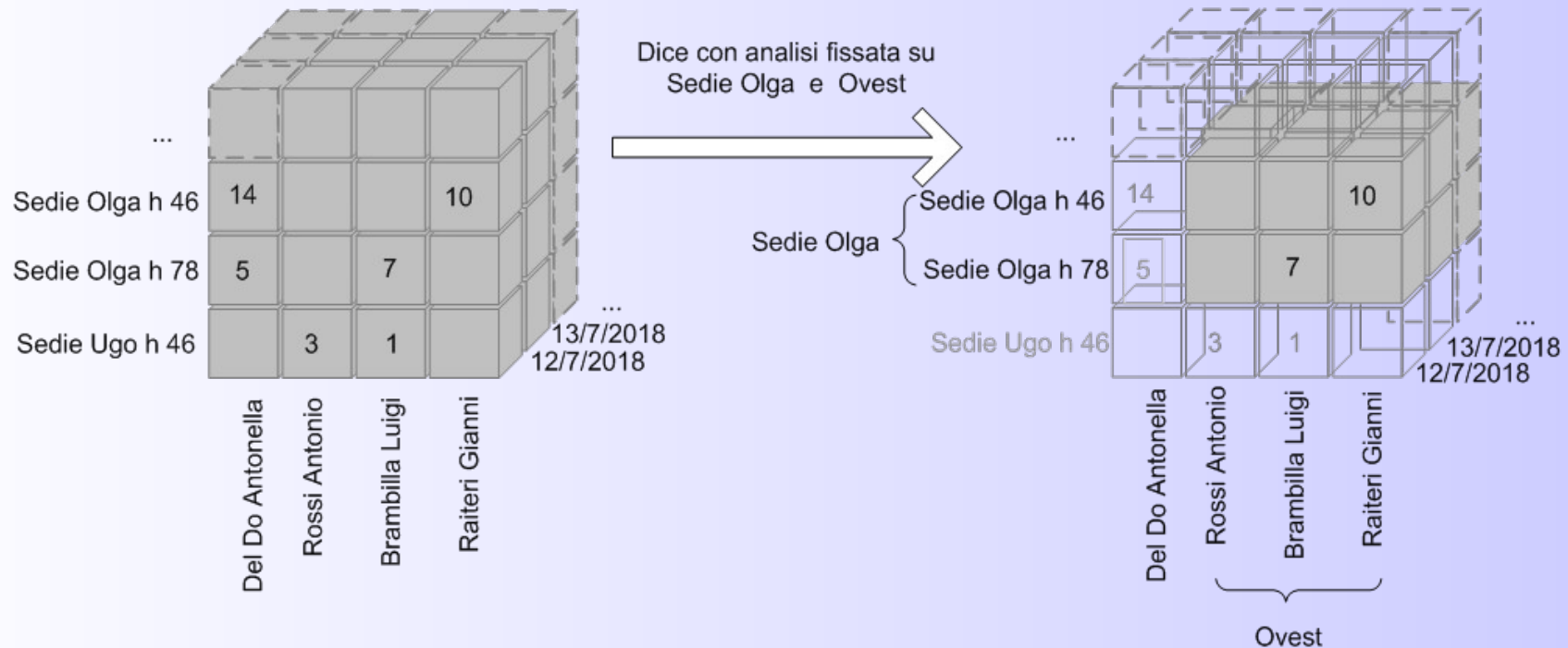
# Operatori OLAP: Slice

- Fissa il valore di una delle dimensioni base per analizzare la porzione di dati filtrati così ottenuta



# Operatori OLAP: Dice

- Filtra i fatti elementari considerati nell'analisi fissando valori per coordinate dimensionali di qualsiasi livello



# Operatori OLAP: Pivot

Prodotto	Area/Anno	2017	2018
Sedia Olga h.46	Est	203	220
	Ovest	64	64

↓

Prodotto	Anno/Area	Est	Ovest
Sedia Olga h.46	2017	203	64
	2018	220	64

Asse rotazione

**Pivoting tra le dimensioni Anno e Area**

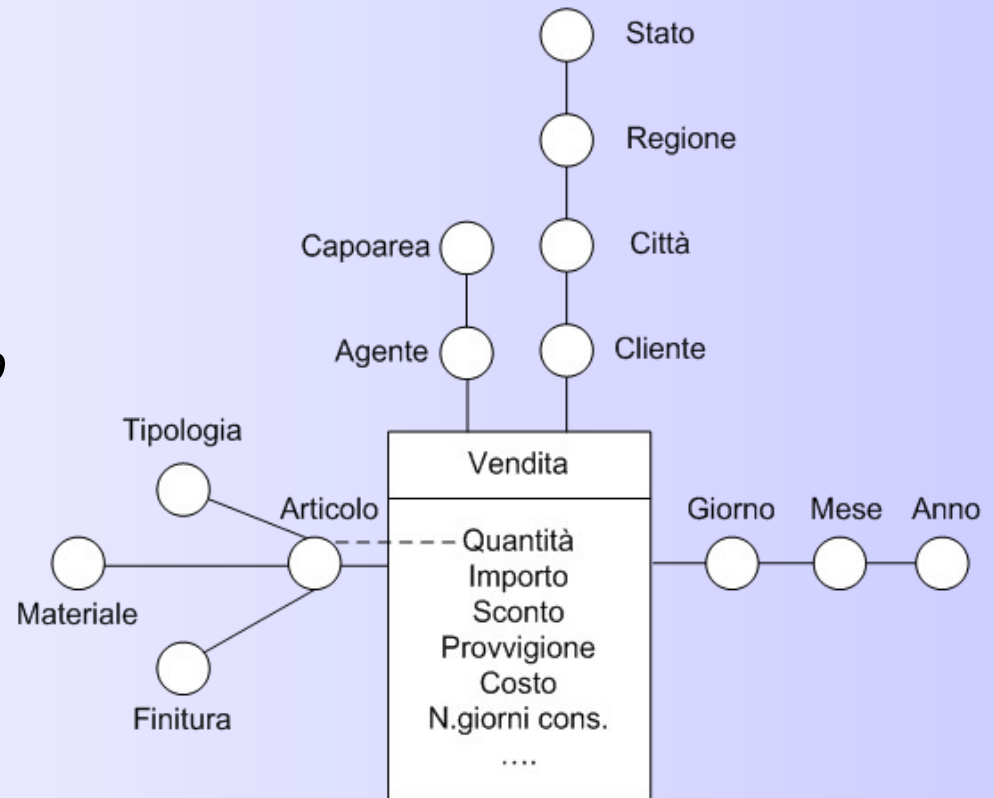
- Inverte la relazione tra le dimensioni, realizzando una rotazione del cubo nell'analisi
- Particolarmente utile nell'analisi di dati presentati in forma tabellare





# Aree di applicazione: Flusso attivo

- Analisi tipiche
  - *Mix di prodotti venduti, fatturato per cliente/area geografica/prodotto, efficienza della rete di distribuzione, rilevamento abbandoni silenziosi, puntualità del servizio al cliente*
- Eventi
  - *Documenti del flusso attivo*

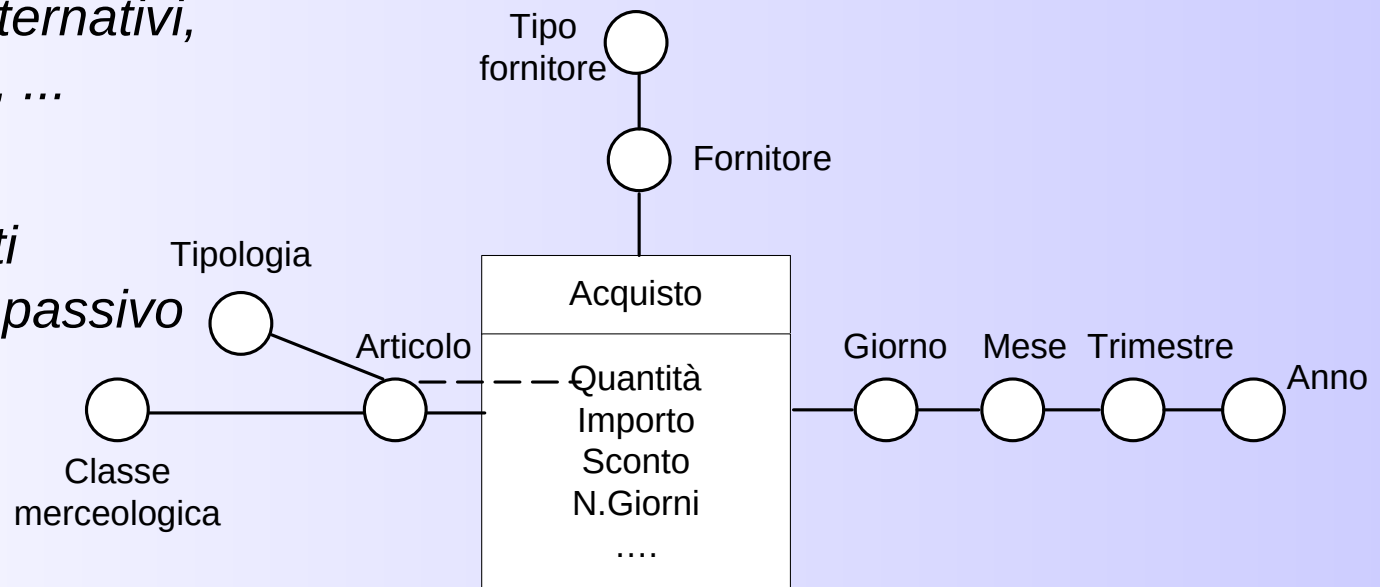


**Esempio di schema di fatto  
per analisi delle vendite**



# Aree di applicazione: Flusso passivo

- Analisi tipiche
  - *Incidenza del costo degli articoli di acquisto, descrizione e confronto di fornitori alternativi, puntualità, ...*
- Eventi
  - *Documenti del flusso passivo*

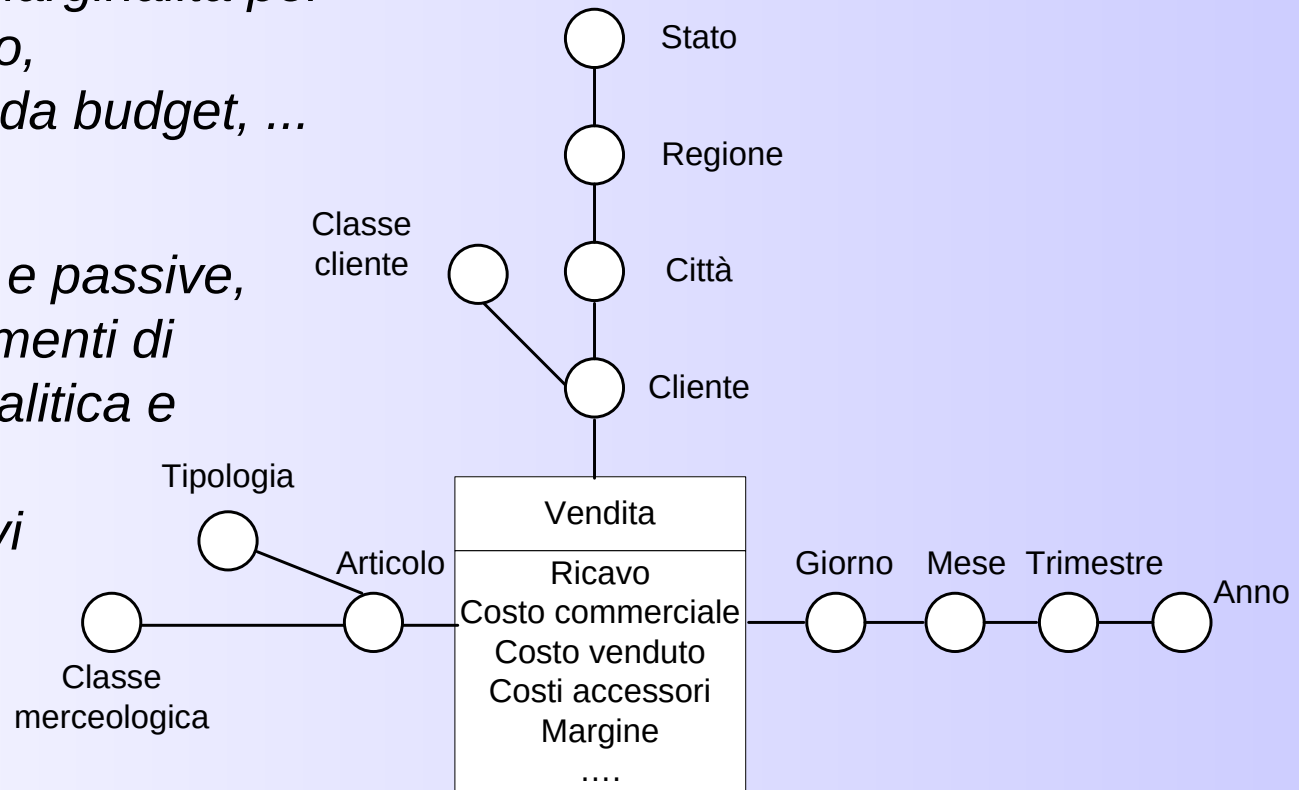


**Esempio di schema di fatto  
per analisi degli acquisti**



# Aree di applicazione: Controllo gestione

- Analisi tipiche
  - *Costi/ricavi, marginalità per cliente/articolo, scostamento da budget, ...*
- Eventi
  - *Fatture attive e passive, budget, movimenti di contabilità analitica e ordinaria, costi produttivi*

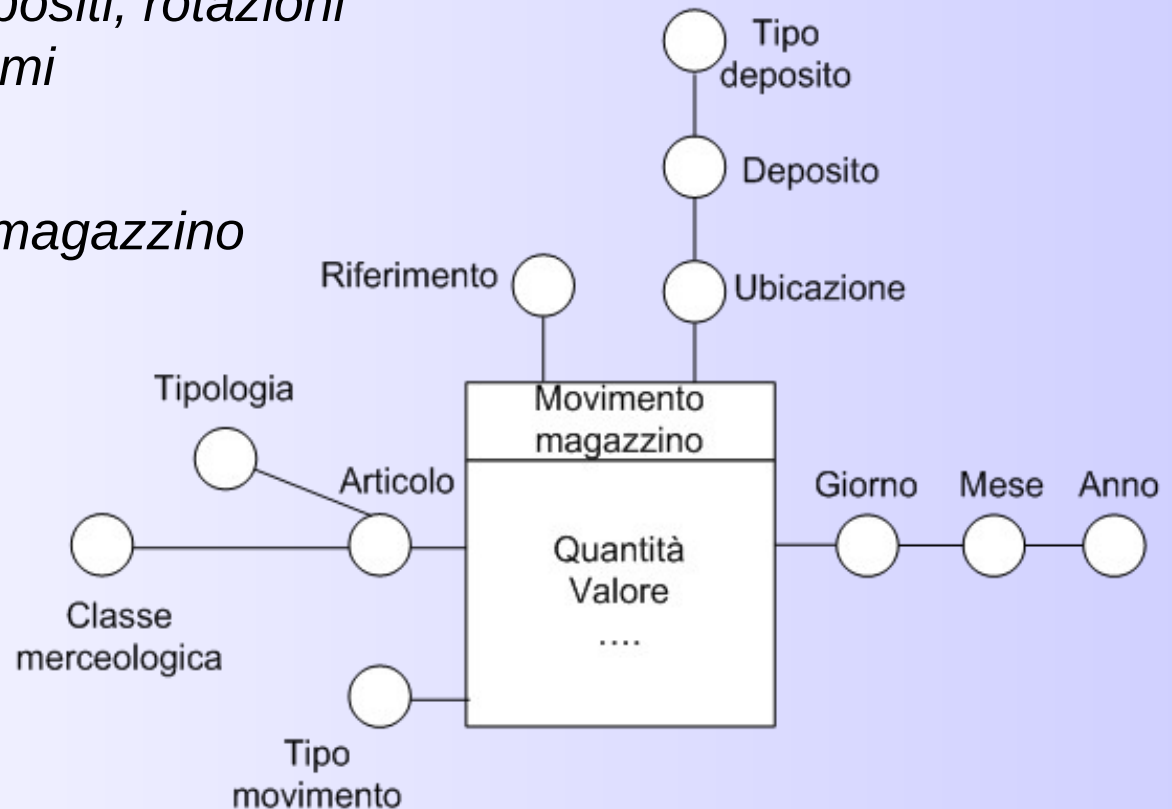


**Esempio di schema di fatto  
per analisi di marginalità**



# Aree di applicazione: Logistica

- Analisi tipiche
  - Attività sui depositi, rotazioni articoli, consumi
- Eventi
  - Movimenti di magazzino

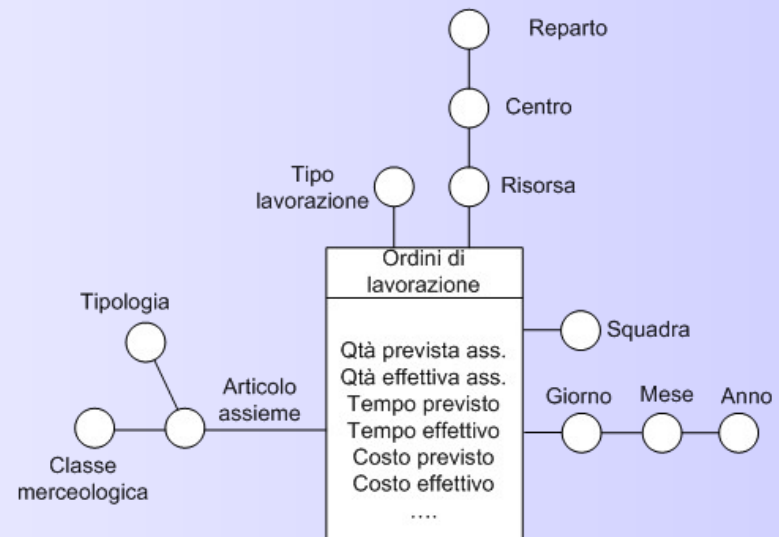
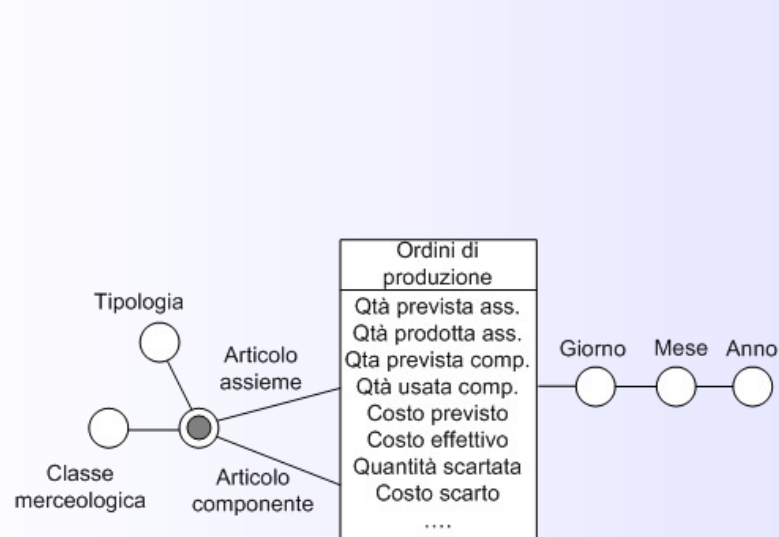


**Esempio di schema di fatto  
per analisi sui movimenti logistici**



# Aree di applicazione: Produzione

- Analisi tipiche
  - *Costi e efficienza del processo produttivo*
- Eventi
  - *Ordini di produzione e di lavorazione, consuntivazione lavorazioni*

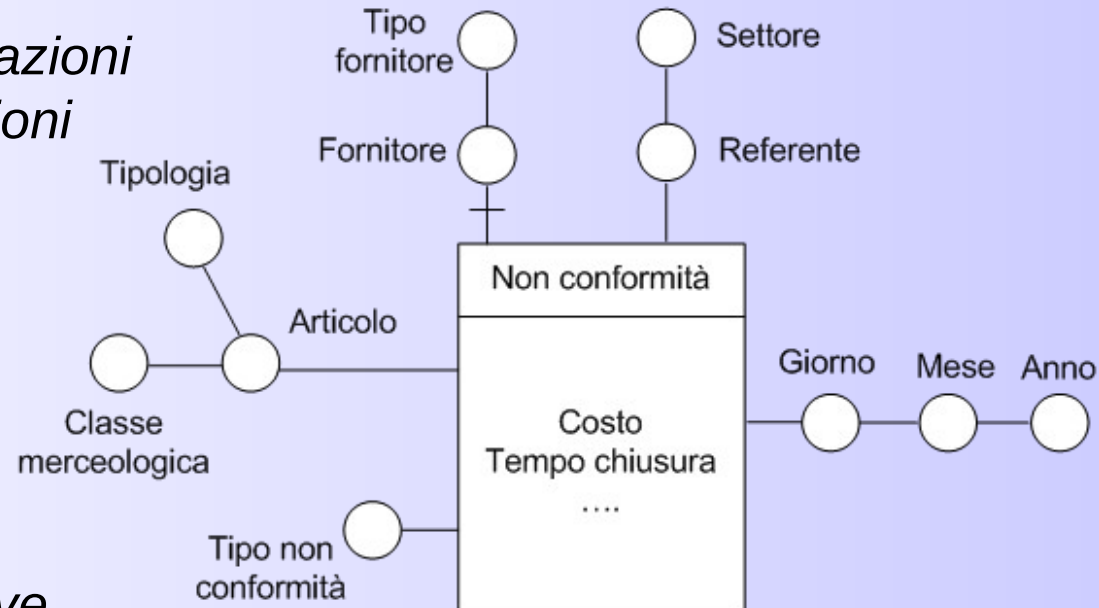


Esempi di schemi di fatto per analisi di produzione con riferimento ai materiali ed alle lavorazioni



# Aree di applicazione: Qualità

- Analisi tipiche
  - *Difettosità degli articoli, puntualità e difettosità dei fornitori, puntualità aziendale, attuazione azioni correttive, efficacia azioni preventive, ...*
- Eventi
  - *Rilevamento non conformità, spedizioni, ingressi a magazzino, reclami, azioni correttive, ...*

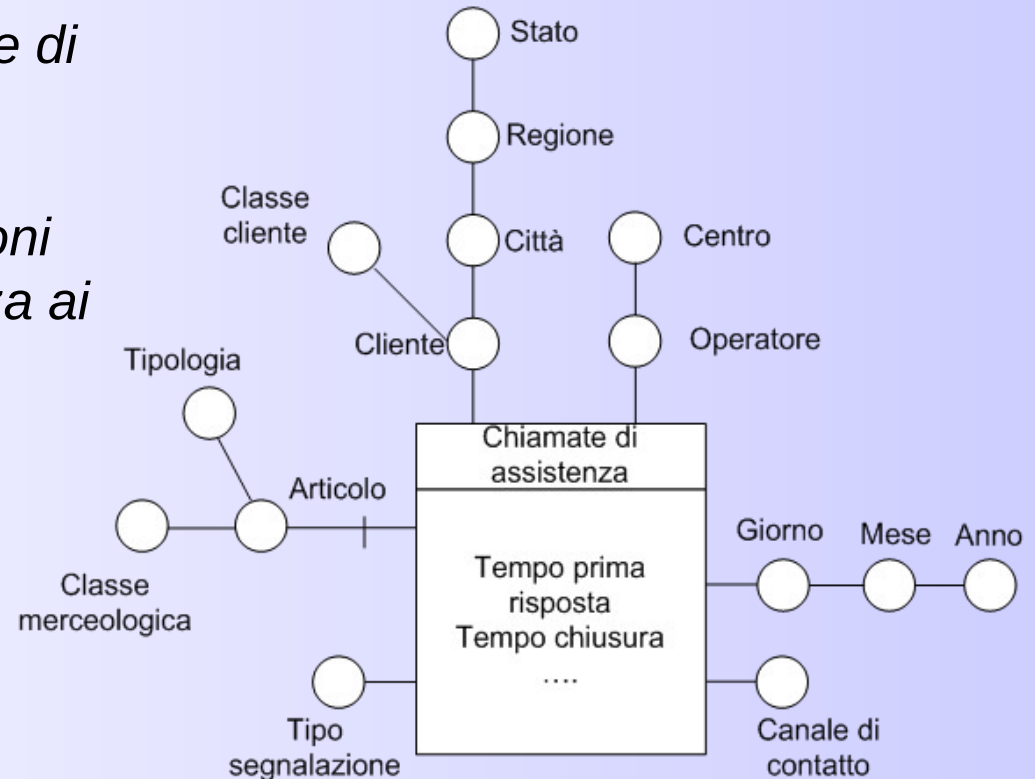


**Esempio di schema di fatto per  
analisi delle non conformità**



# Aree di applicazione: CRM

- Analisi tipiche
  - *Efficacia di promozioni e di azioni di fidelizzazione, esito di campagne di telemarketing, prestazioni del servizio di assistenza ai clienti*
- Eventi
  - *Azioni commerciali, vendite, chiamate di assistenza, ...*

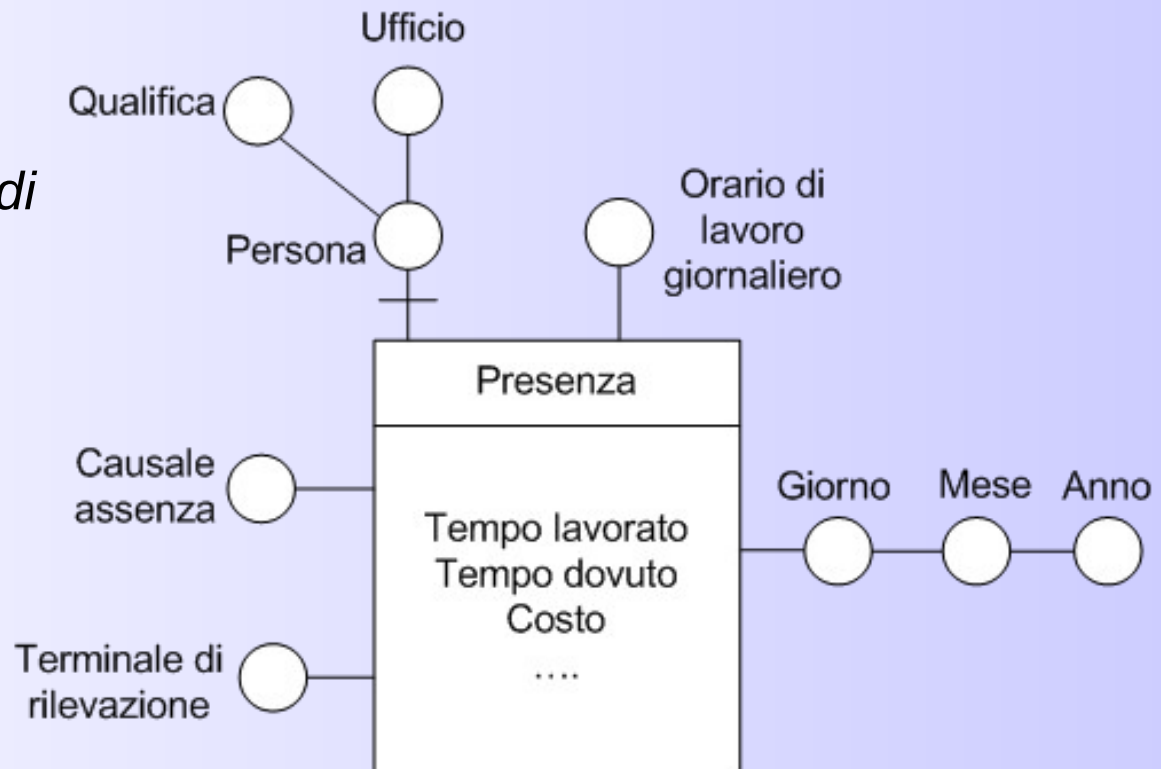


**Esempio di schema di fatto  
per analisi sul servizio di assistenza  
clienti**



# Aree di applicazione: Risorse umane

- Analisi tipiche
  - Presenze, retribuzioni, ...
- Eventi
  - Ingressi/Uscite, emissione cedolini di pagamento, ...



Esempio di schema di fatto per  
analisi delle presenze

