Filippo Geraci

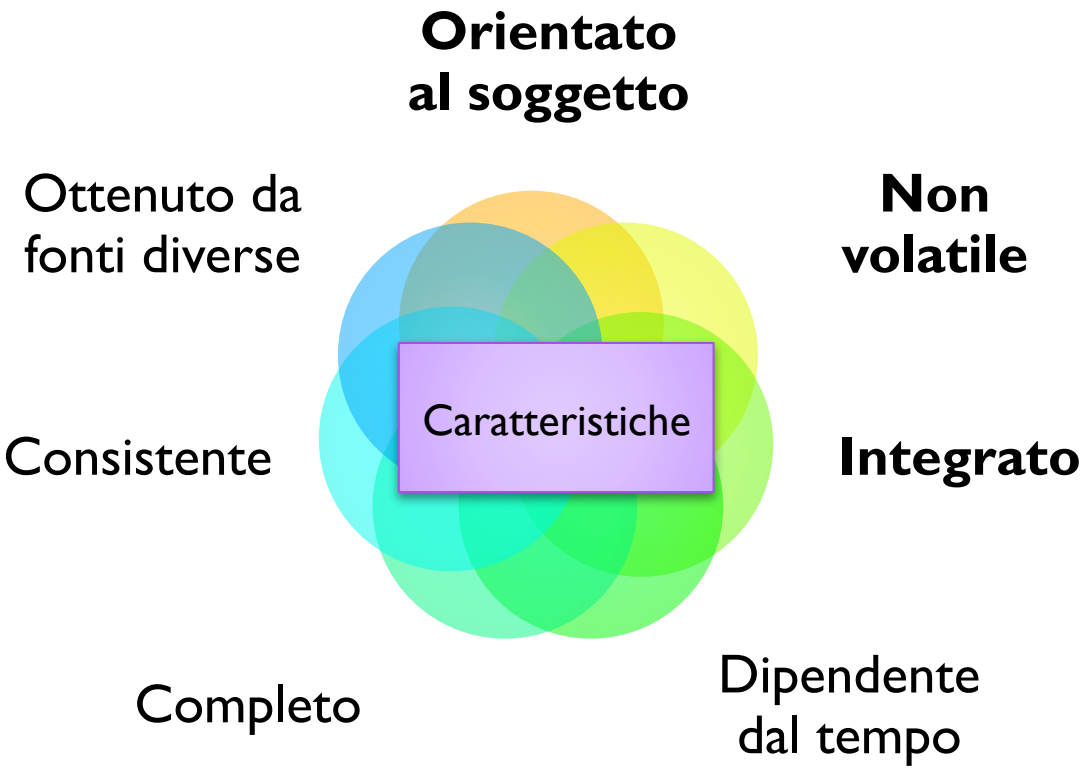
DATA WAREHOUSING

Data warehouse

- 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“

… un data warehouse

- **Obiettivo**:
  - supportare le decisioni
- **Obiettivo operativo**:
  - Estrarre, analizzare, presentare i dati



Orientato al soggetto

- I data warehouse sono progettati per aiutare l'utente ad analizzare i suoi dati in base al suo percorso logico senza schemi prestabiliti

- *Chi è stato il nostro miglior venditore di aspirapolveri lo scorso anno??*

  "miglior venditore di aspirapolveri": → analisi orientata al soggetto

Integrato

- Risoluzione dei conflitti tra nomi dei campi e dei problemi derivanti dal fatto che i dati si trovano espressi in unità di misure differenti.

- Nel database della succursale di Roma il Sig. Rossi ha venduto 1000 aspirapolveri a €900 mentre nel database della filiale di NY, Mr Smith ha venduto 900 aspirapolveri a $ 600
  - come confronto i dati?
  - Come risolvo i conflitti tra nomi?

Non volatile

- I dati non variano una volta entrati nel warehouse
- Il warehouse deve analizzare ciò che è accaduto

- *Il Sig. Rossi ha venduto 1000 aspirapolveri, ed ad oggi è il RecordMan di vendite*
  - *Se tra mezz'ora Mr. Smith ne vende altri 250, questa informazione non deve entrare nel data wharehouse*

Dipendente dal tempo

- La maggior parte delle analisi di business sono analisi di "trend". Per questo si ha bisogno di una grande mole di dati storici.

- *Voglio sapere negli ultimi tre anni l'andamento in borsa della Compagnia su Milano, Londra e Francoforte*

Correttetto e consistente

- Decisioni prese in base a dati non completi o non corretti possono portare a scelte errate

- Premio agente che fattura più di € 5000 annui. Il Sig. Verdi ha venduto 10 aspirapolveri da € 400 per la sede italiana, poi si è spostato a N.Y ed ha venduto 2 aspirapolveri. Non conosco il prezzo di vendita di N.Y.
  - Gli devo dare il premio?

Metodologie di accesso ai dati nei sistemi informativi

- OLTP: On Line Transaction Processing
  - Usato nei sistemi ERP per l'accesso ai dati
- OLAP: On Line Analytical Processing
  - Fornirsce supporto efficiente per l'analisi prendendo in considerazione più variabili contemporaneamente
- I dati usati dai sistemi OLAP sono gli stessi di quelli usati dai sistemi OLTP:
  - Cambia elaborazione
  - Cambia memorizzazione sul database

OLTP - On Line Transaction Processing

- Transazioni predefinite e di breve durata
- Dati dettagliati, recenti e aggiornati
- Dati residenti su un unico DB logico
- Read & write di pochi record
- Critiche le proprietà ACIDe
  - Atomicity (atomicità)
  - Consistency (consistenza)
  - Isolation (transazionalità)
  - Durability (robustezza)

OLAP - On Line Analytical Processing

- Interrogazioni complesse e casuali
- Interffaccia di interrogazione interattiva
- Dati storici e aggregati
- Dati provenienti da più DB eterogenei
- Moltissime operazioni di Read (nessuna di write)
- Visualizzazione dei dati su PC
- Scoperta di nuove relazioni tra le variabili

Caratteristiche sistemi di analisi OLAP

- FASMI - OLAP Report 1995
  - F ◦ Velocità di risposta (Fast)
  - A ◦ Analiticità (Analytical)
  - S ◦ Condivisione delle informazioni (Shared)
  - M ◦ Multidimensionalità (Multidimensional)
  - I ◦ Informatività (Informational)

Caratteristiche FASMI

- Velocità:
  - Sistema interattivo non deve interrompere il processo mentale.
  - Analisi OLAP in pochi secondi
  - Per il data mining non è sempre vero
- Analitico:
  - Report dati in forma grafica e tabellare
  - Deve seguire i percorsi mentali quindi:
    - Nuove analisi a partire dall'ultima elaborazione

Caratteristiche FASMI

- Condiviso:
  - Gestione di user management (ruoli diversi portano viste diverse)
- Multidimensionale:
  - Visione di un fatto da più prospettive
- Informativo:
  - Contiene tutti e soli i fatti di interesse per l'analisi
  - Dati completi e corretti

Confronto tro OLTP e OLAP

OLTP

- Utenti: Impiegati
- Operazioni giornaliere
- Operazioni. Correnti
- Operazioni. Ripetitivo
- Transazioni brevi
- Decine di record acceduti per volta
- Migliaia di utenti
- 100 MB – 1 GB

OLAP

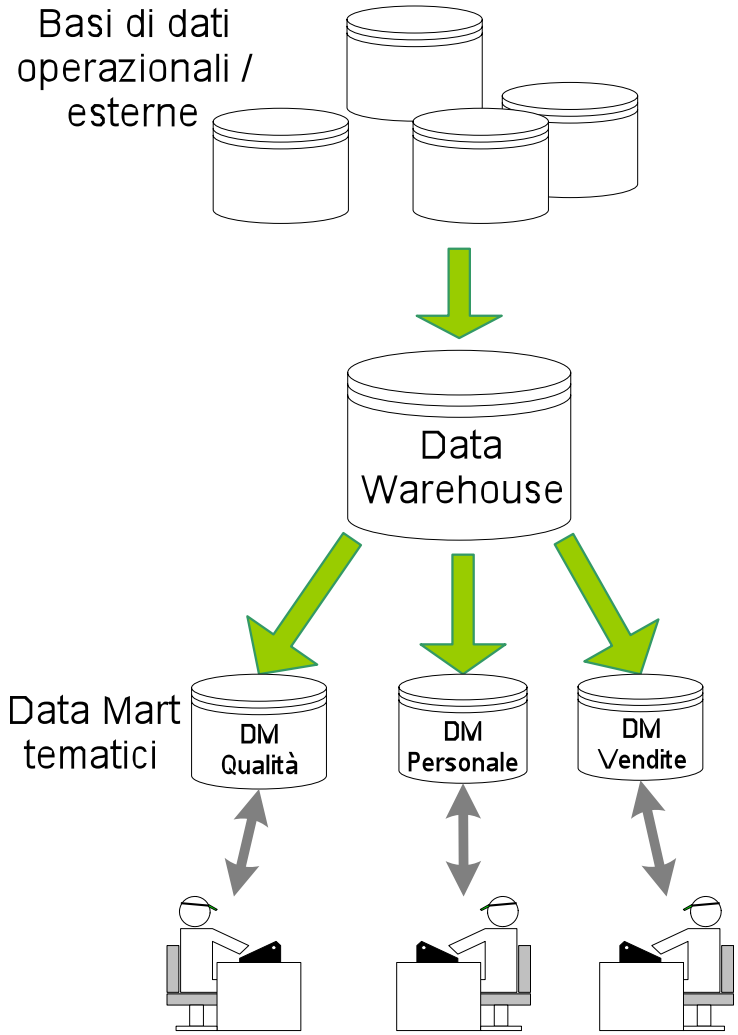
- Utenti : dirigenti
- Supporto Decisioni
- Dati Storici
- Oper. Casuali
- Int. Complesse
- Milioni di record acceduti
- Centinaia di utenti
- 100 GB – 1 TB

Data warehouse e Data mart

- Data warehouse
  - Magazzino unico, completo e consistente dell'informazione aziendale
  - Contiene dati articolati attorno a tutti i fatti di interesse aziendale (tutti i possibili ipercubi)
  - Può raggiungere dimensioni estremamente elevate
- Troppi dati nell'analisi rischiano di confondere e rallentano la computazione
- Troppo pochi dati possono fare perdere informazioni
- Data mart: porzione del data warehouse che contiene
  - Tutti i dati di suo interesse
  - Solo i dati di suo interesse

Data mart

- Data warehouse tematico, derivato dal data warehouse aziendale
  - Comprende i soli fatti che riguardano una certa area d'indagine
  - Estensione temporale ridotta
  - Granularità dei fatti minore



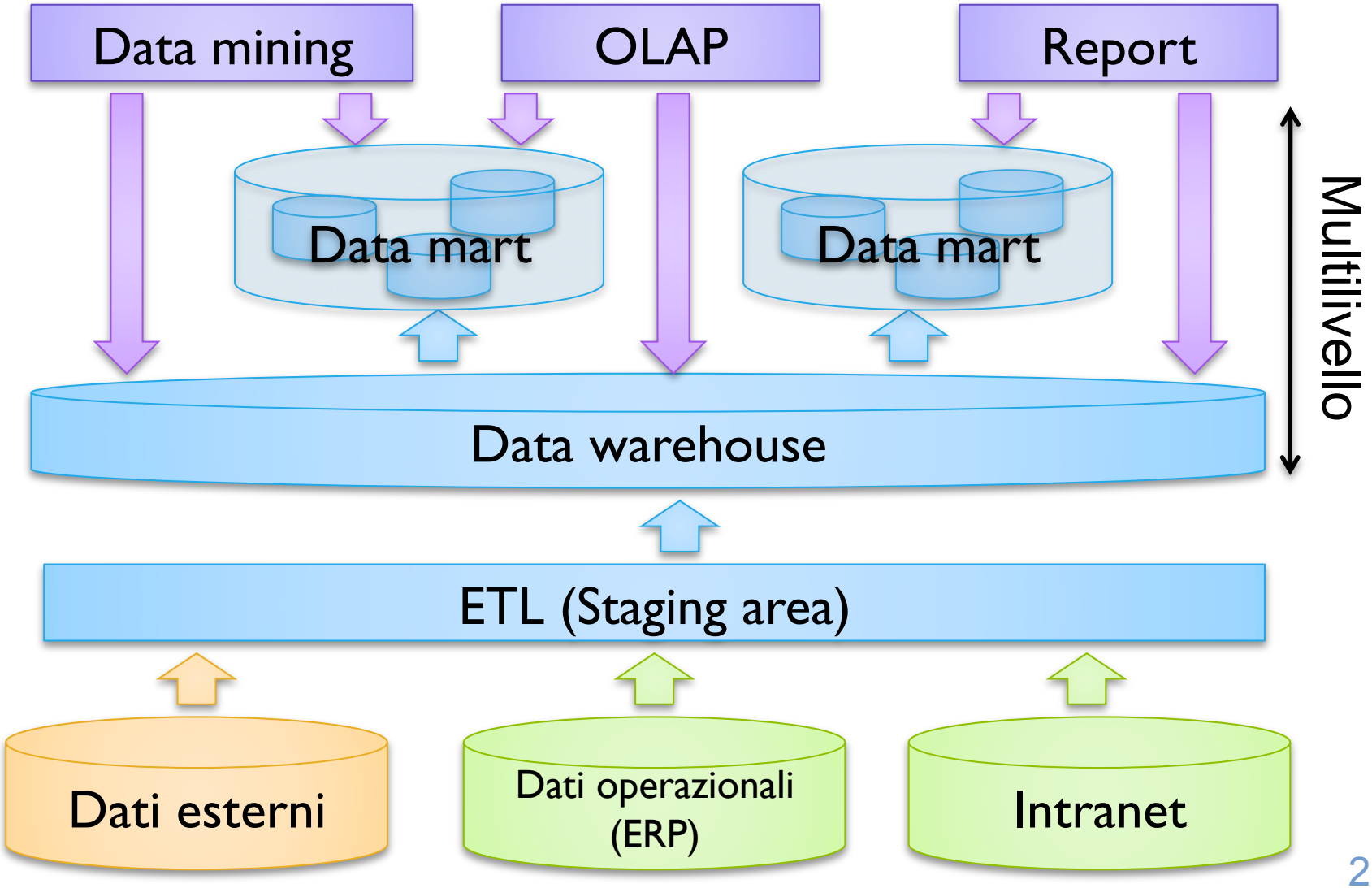
Architettura dei sistemi di data warehousing

- Sistema costituito da basi di dati a livelli distinti, diverse per: finalità, struttura e tipologia di dati
  - Sorgenti
    - basi di dati origine (operazionali o esterne)
  - Staging Area (opzionale)
    - area intermedia utilizzata come appoggio per le procedure di trasformazione dei dati
    - ETL (Extraction, Transformation Loading)
  - 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

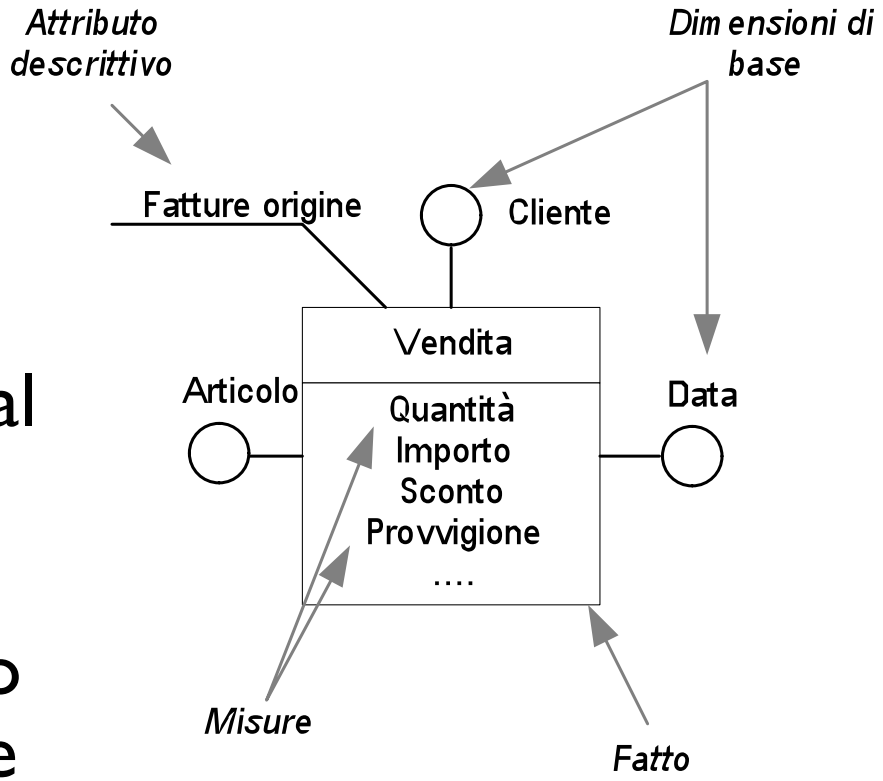


Modelli concettuali per il data warehouse: il DFM

- Il dimentional fact model DFM fornisce una visione ad alto livello e statica di ogni fatto
  - Descrive le misure associate
  - Descrive le dimensioni e le gerarchie
  - Descrive gli attributi descrittivi
- Ogni fatto è rappresentato tramite uno schema di fatto
  - Rappresentazione grafica

DFM – Schema di fatto

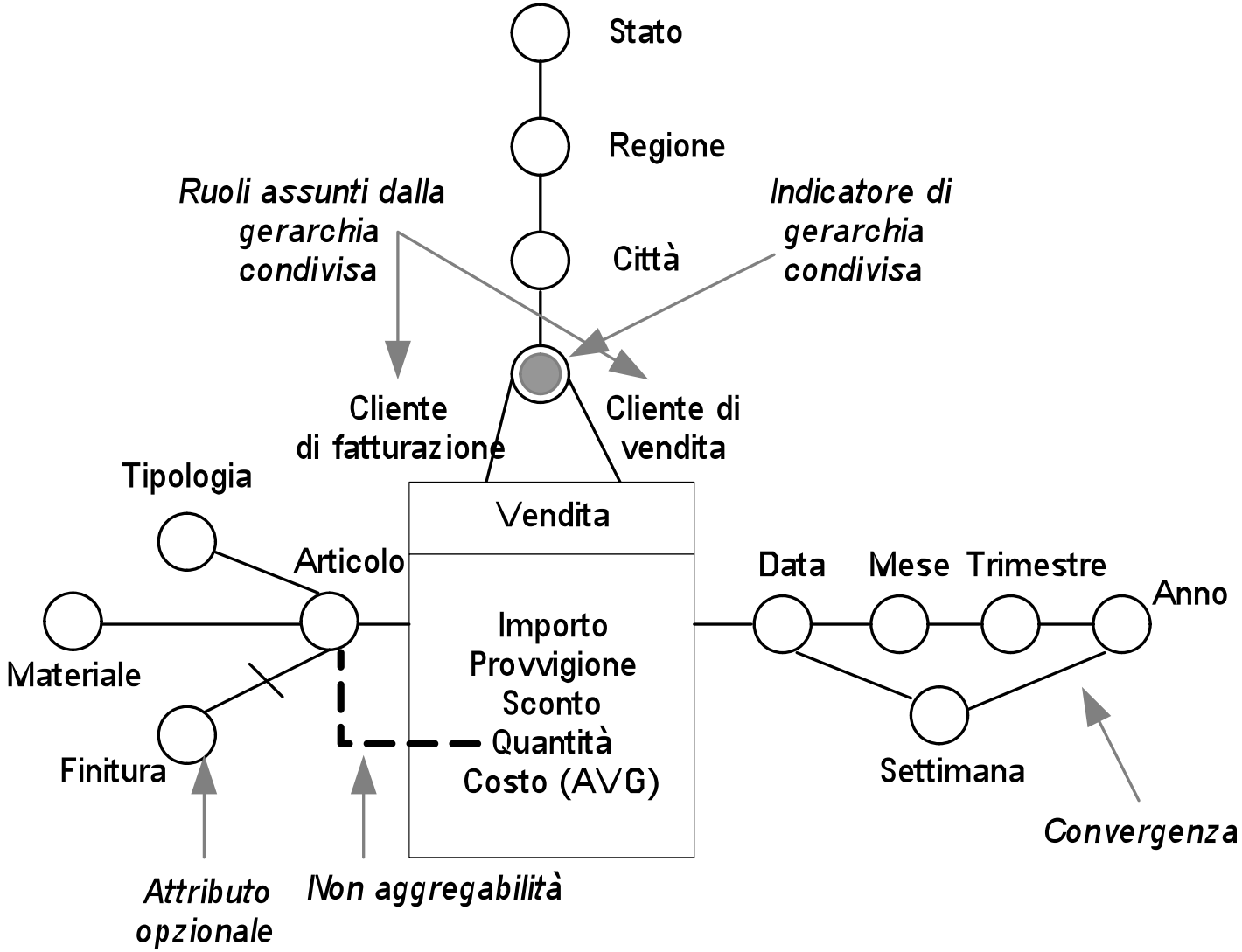
- **Fatto**: rettangolo contenente il nome del fatto e le sue misure
- **Dimensioni di base**: circoletti etichettati collegati al fatto
- **Attributi**: collegati con una linea al fatto o ad una dimensione



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

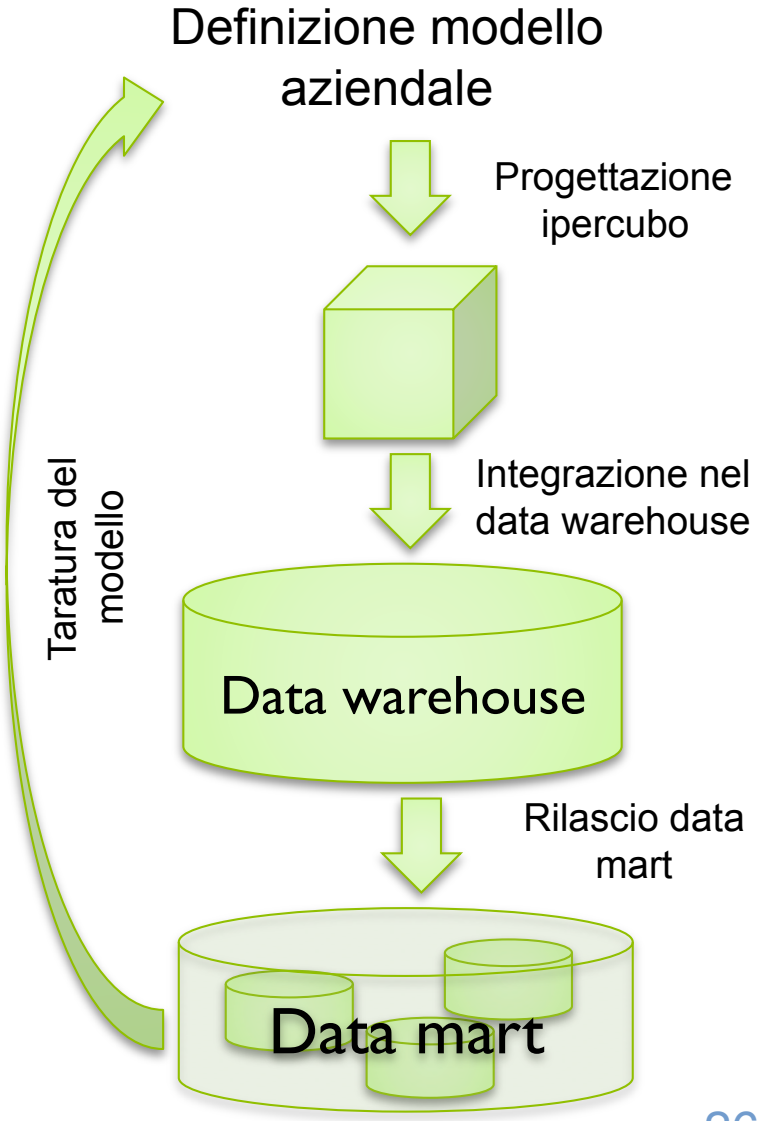
Modelli concettuali per il data warehouse: il DFM



PROGETTAZIONE DEL DATA WAREHOUSE

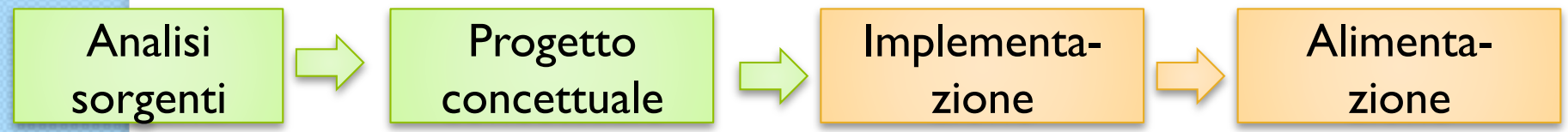
Il ciclo di vita dei sistemi di data warehousing

- Approccio costruzione iterattivo ed incrementale
  - 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
  - Sviluppo del modello in base ad uso effettivo
  - Espansione dimensioni in base ad uso effettivo



Costruzione del data warehouse

- 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
- Progettazione logica e ed implementazione fisica
  - 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



Modelli logici per il data warehouse – Architetture fisiche

- Bisogna scegliere il tipo di database ed il linguaggio di interrogazione

1. **Database**:
   - Relazionale: riporta il modello multidimensionale ad un modello relazionale
   - Multidimensionale
   - Ibrido (Data warehouse relazionele + data mart multidimensionale)
2. **Linguaggio di interrogazione**:
   - SQL
   - Proprietario del database multidimensionale
   - Proprietario di uno specifico prodotto

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
  - Facile trovare operatori con esperienza
  - Facilmente aggiornabili da ERP
- Svantaggi:
  - esecuzione di query poco efficiente
  - Miglioramento velocità di risposta implica aumento complessità e occupazione di spazio
    - Materializzazione delle viste
    - Denormalizzazione

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 proprietari
- Vantaggi
  - elevata efficienza nell'esecuzione di query complesse
  - stretta aderenza al modello concettuale
- Svantaggi
  - elevata occupazione di spazio
    - Allocato spazio per ogni possibile ennupla dimensionale
    - Solo poche celle contengono informazione (20%)
  - Nessuno standard, di rappresentazione e di interrogazione
  - Difficile trovare operatori con esperienza

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

Implementazione ROLAP

- Schemi multidimensionali su basi di dati relazionali
  - Schema a stella (star schema)
  - Schema a fiocco di neve (snowflake)

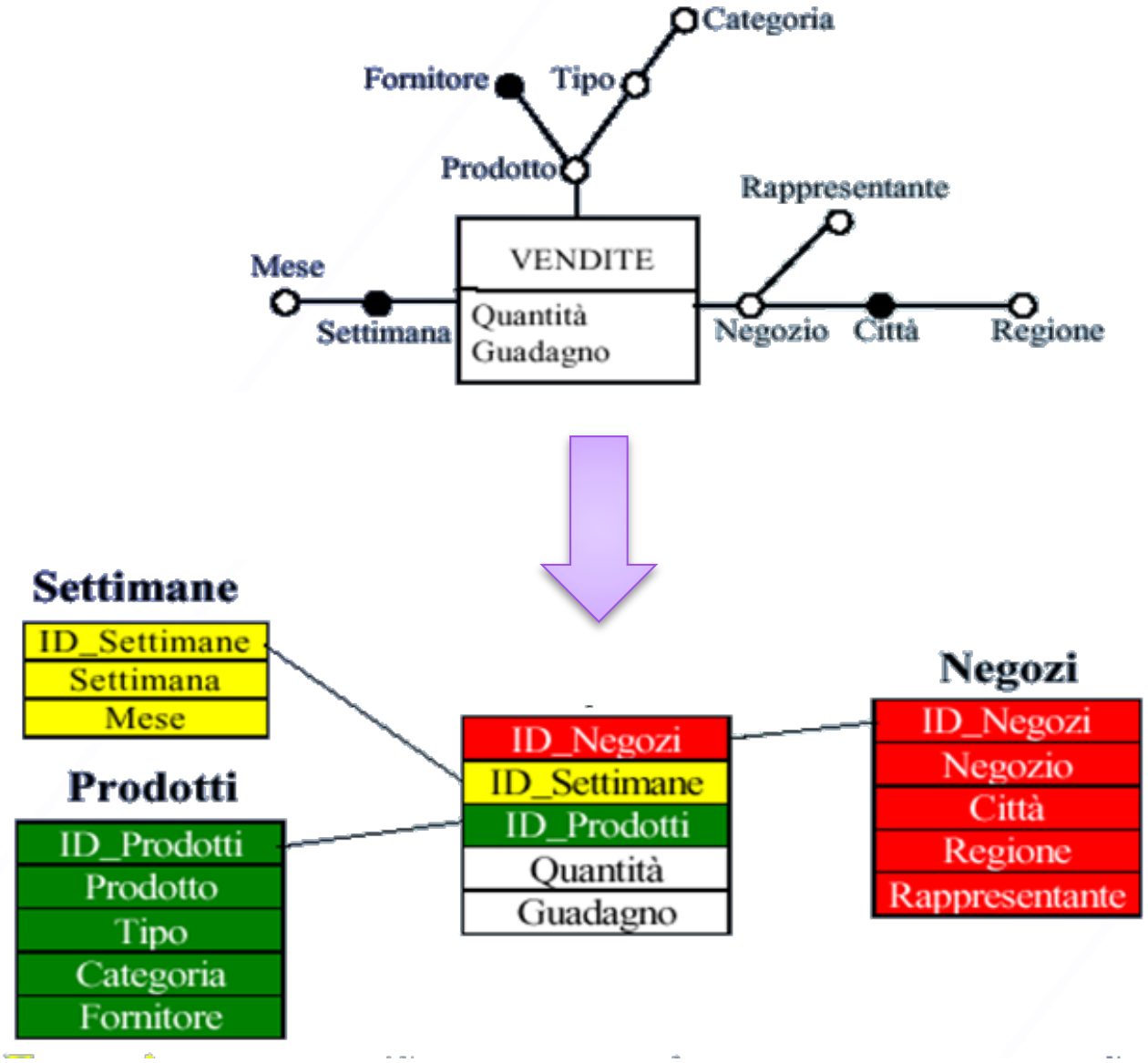
Schemi multidimensionali su basi di dati relazionali - Schema a stella

- 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
  - Ignora le ridondanze nelle gerarchie e le gerarchie condivise

Schemi multidimensionali su basi di dati relazionali - Schema a stella

- Vantaggi
  - massima velocità nel reperimento delle informazioni
  - Basta un unico join per recuperare tutti i dati
- Svantaggi
  - Ridondanza
  - spazio occupato
  - scarsa intuitività della struttura
  - elevata complessità di aggiornamento

Schema a stella - esempio



Schema a stella - esempio di Query

SELECT Settimane.ID_Settimane, Prodotti.Fornitore, Negozzi.Città,
SUM (vendite.Quantità)

FROM Vendite, Negozi, Settimane, Prodotti

WHERE Vendite.ID_Negozi = Negozi.ID_Negozi

AMD Vendite.ID_Settimane = Settimane.ID_Settimane

AMD Vendite.ID_Prodotto = ID_Prodotto

AMD Prodotti.Tipo = "Sport" AND Negozi.Regione = "Toscana"

GROUP BY Settimane.ID_Settimane, Prodotti.Fornitore,
Negozi.Città

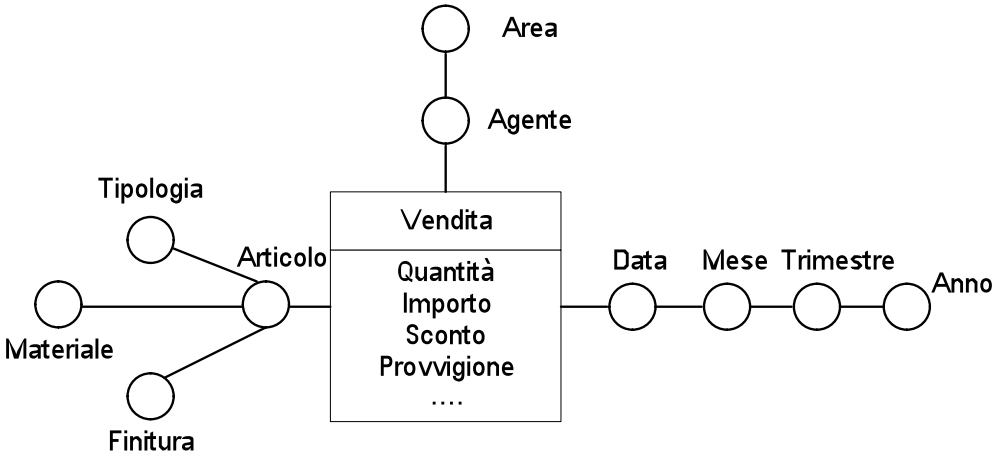
Settimana	Fornitore	Città	Fatturato
52	Rossi	Siena	350
52	Rossi	Pisa	200

- Restituisce i volumi di vendita degli articoli sportivi venduti in Toscana suddivisi per settimana, fornitore e città

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
  - Ottimizzazione query frequenti con materializzazione di viste
  - Minor sensibilità alle variazioni logiche delle gerarchie nel tempo
- Svantaggi
  - Più lento perchè deve fare molte join
- Costellazione
  - Tabelle dimensionali condivise da più tabelle dei fatti
  - Approccio da seguire quando più fatti coinvolgono gli stessi soggetti

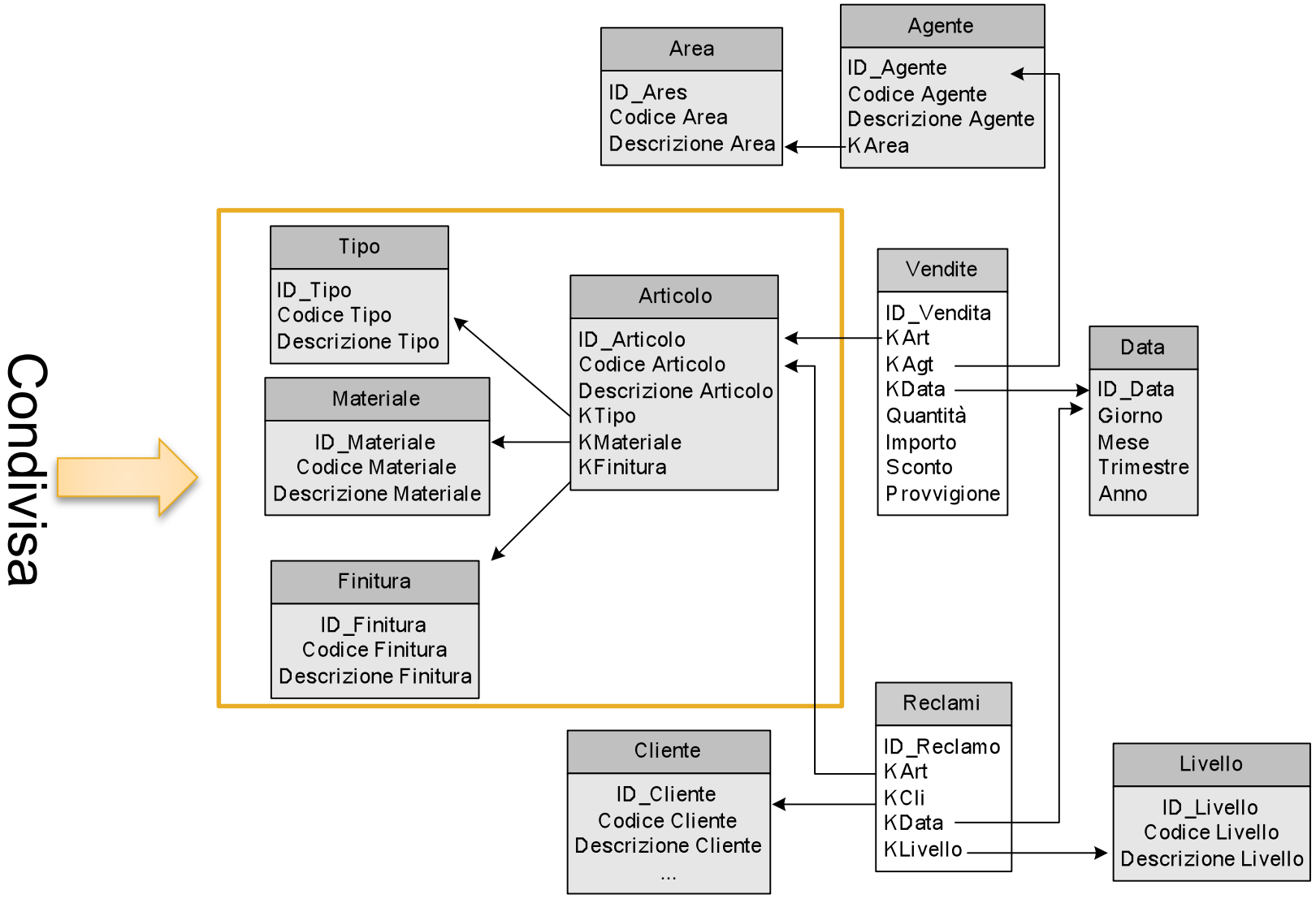
Schema a fiocco di neve



Modello concettuale

Modello logico su schema a fiocco di neve

Costellazione di fatti



Costellazione tra Vendite e Reclami

Popolamento del data warehouse

- Procedure complesse devono garantire:
  - Correttezza e completeza
  - Consistenza
- Problema del "query transformation"
- ETL (Extraction, transformatio, loading):
  - Estrazione, pulizia e caricamento
  - Operano nella staging area

Schema Matching

- ***Query Transformation***:
  - trasformare una interrogazione su un db con una certa struttura in una interrogazione equivalente su un db con una struttura diversa
- ***Schema Mapping***:
  - definisce le regole per passare da una struttura all'altra
- ***Schema Matching***:
  - Il problema di "imparare" le regole di mapping

Query transformation

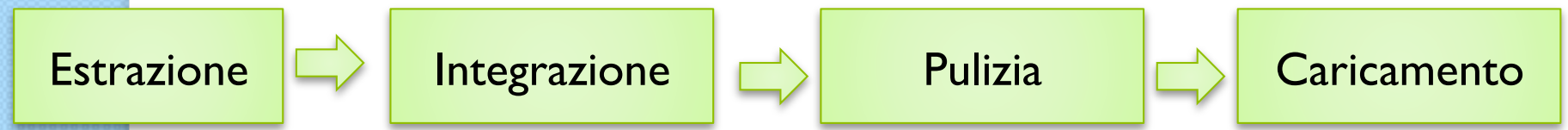
Docente	Facoltà	Insegnamento
Filippo Geraci	Ingegneria	SIA

Insegnante	Facoltà	Corso
Filippo Geraci	Ingegneria	SIA

- Mapping
  - Docente → Insegnante
  - Insegnamento → Corso

Fasi di popolamento del data warehouse

1. Estrazione
   - estrae dalle sorgenti i dati da portare sul data warehouse
2. Integrazione e trasformazione
   - riconduce i dati estratti al modello unificato definito per il data warehouse
3. Pulizia
   - aumenta la qualità dei dati, riconoscendo e risolvendo errori, incongruenze ed omissioni
4. 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
    - Aggregazione alla fonte
    - Estrazione al dettaglio massimo
- Tipi di estrazione
  - **statica**: tratta tutti i dati presenti nelle sorgenti
  - **incrementale**: tratta i soli dati inseriti o alteratidopo l'ultimo popolamento del data warehouse,
  - Identificazione nuovi dati:
    - Delegata alle applicazioni o al db
      - necessita di staging area
    - Pilotata da timestamp nei dati
    - Statica con successivo 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
  - standardizzazione
    - del formato
    - delle convenzioni
    - delle codifiche

Integrazione e trasformazione Esempio

Docente	Facoltà	Corso	Anno accademico
Filippo Geraci	Ingegneria	SIA	2012/13

Schema Mapping

Riconciliazione

Insegnante	Facoltà	Corso
F. Geraci	Ingegneria	SIA

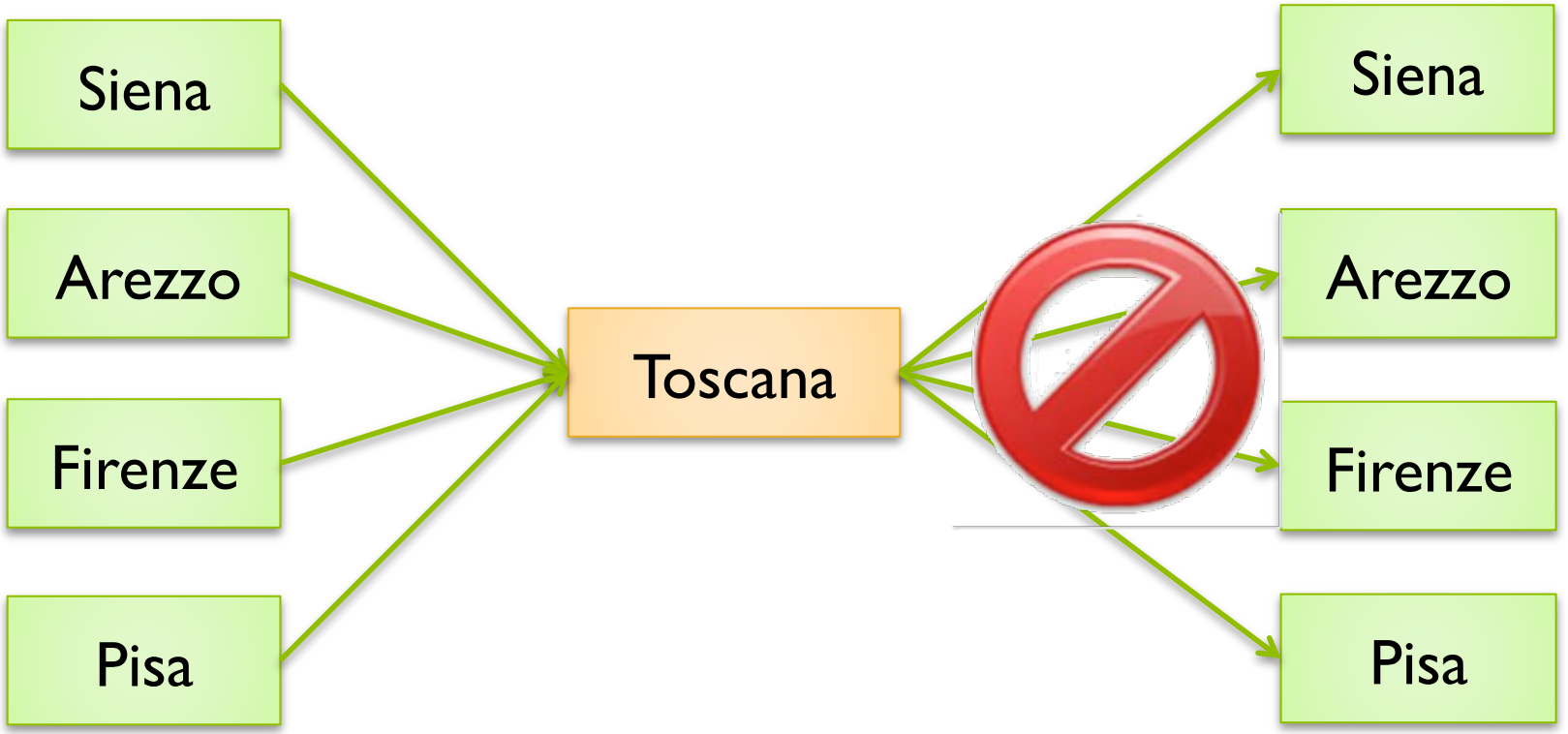
Standardizzazione

Rimozione duplicati

Docente	Facoltà	Corso	Anno accademico
Filippo Geraci	Ingegneria	SIA	2012/13

Standardizzazione della codifica e perdita di informazione

- E' possibile passare da una codifica più informativa ad una meno, ma non viceversa
- La conversione e' possibile per relazioni molti ad uno o uno ad uno



Popolamento del data warehouse - Pulizia

- Innalzamento del livello di qualità dei dati
- Non è necessariamente successiva alla integrazione
- Tipologie di errori trattati
  - dati incompleti
    - Strumenti: interpolazione
  - dati errati o incomprensibili
    - Esempio: codice fiscale errato
    - Strumenti: regole e dizionari
  - dati inconsistenti
    - Esempio: errore abbinamento CAP con comune
    - Strumenti: 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
  - I fatti non sono mai eliminati
  - Eventuale sovrascrittura degli elementi modificati

Popolamento del data warehouse -strategia aggiornamento dimensioni

- Non modificare dimensioni
  - Ogni fatto usa gli attributi dimensionali validi all'inserimento della dimensione
  - Dimensioni non corrispondenti al presente aziendale
- Sovrascrivere
  - Ogni fatto usa gli attributi dimensionali validi adesso
  - Sii perde l'informazione sul passato (analisi passate con dimensioni presenti
- Creare una nuova istanza
  - Associata ai fatti da oggi in poi
    - Massima corrispondenza con la realtà
  - Con marcatore temporale varibile
    - Rende possibili analisi su scenari
    - Esempio: cosa succederebbe se la dimensione cambiasse in altro data

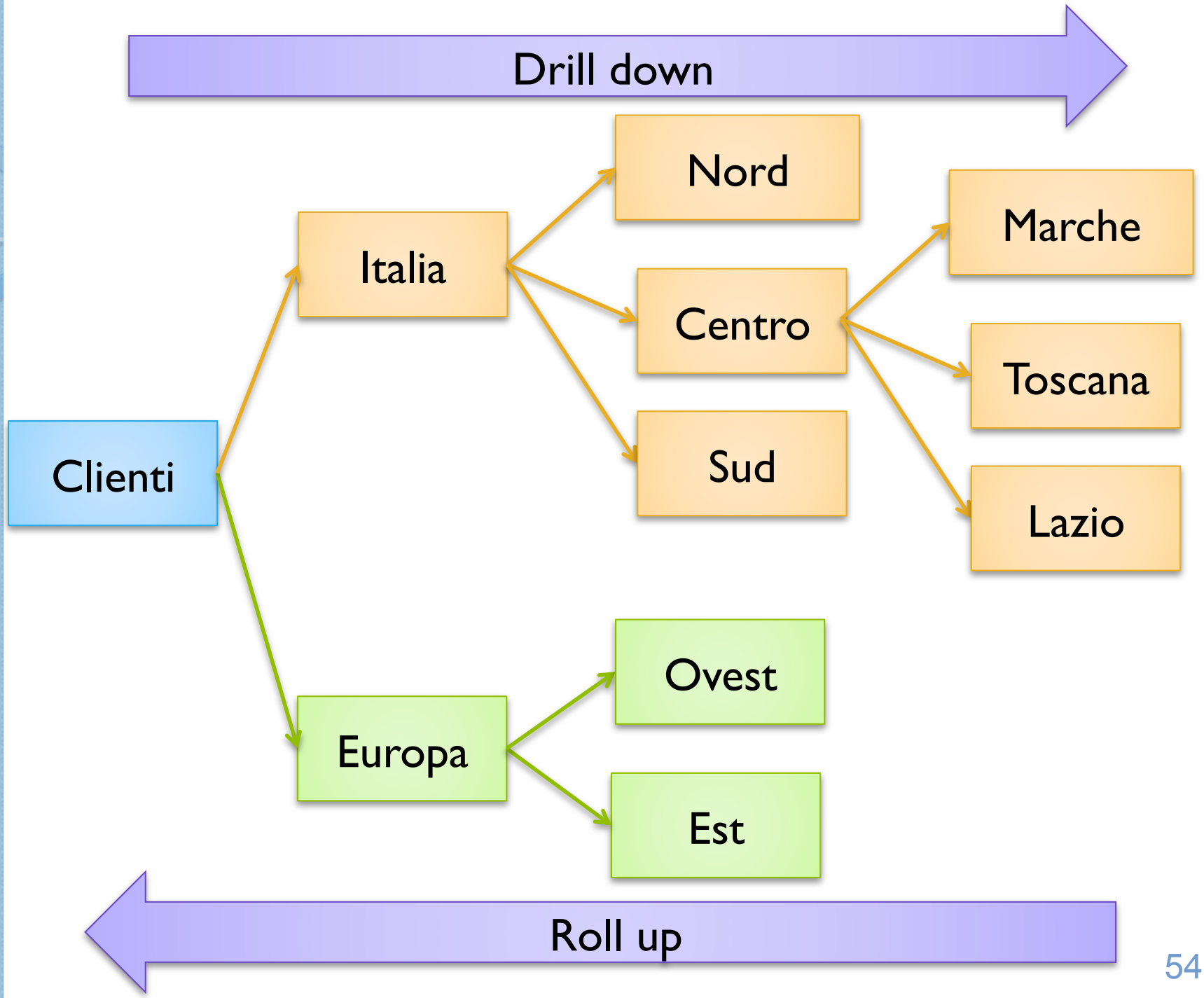
ANALISI OLAP

L'analisi OLAP

- Navigazione interattiva sui dati multidimensionali
- Esplorazione guidata da ipotesi
  - Esempio: presumo che fatturato 2010 sia superiore a quello 2009. Estraggo i due dati ed effettuo confronto
- 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

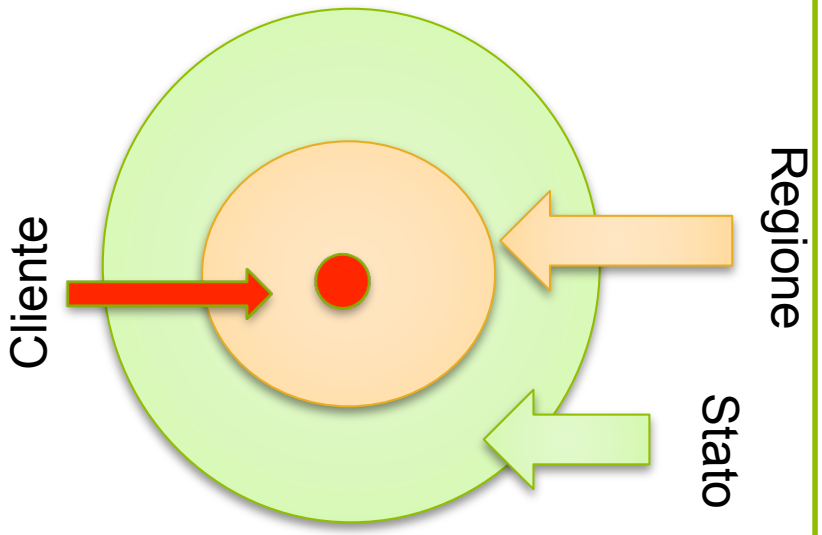
Operatori OLAP

- Drill down
  - Disaggregazione dei dati
    - Esempio: mostra le vendite giornaliere e dettagliate di ciascun negozio per una certa categoria di prodotti
- Roll Up
  - Aggregazione dei dati
    - Esempio volume di vendita totale dello scorso anno per categoria e regione
- Slice
  - limita l'analisi ad valore specifico per una dimensione
- Dice
  - limita l'analisi a valori specifici su più dimensioni
- Pivot
  - Riorientamento del cubo

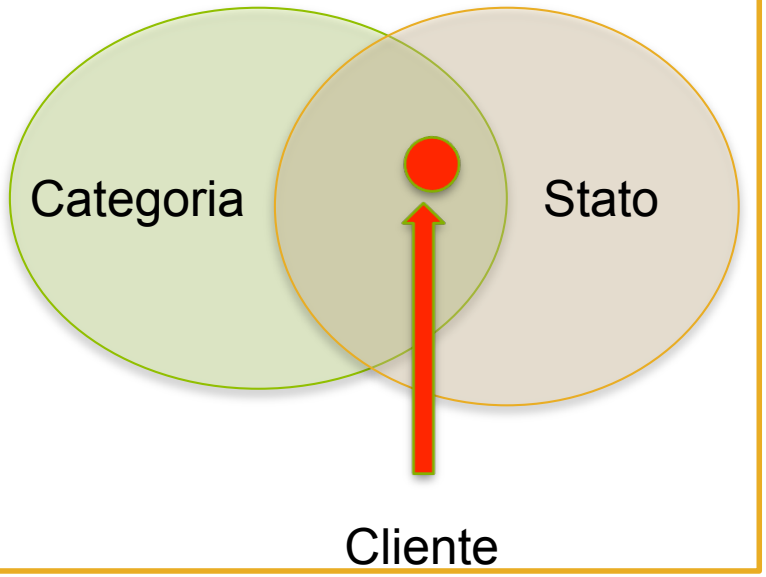

Drill down
Clienti
Italia
Nord
Centro
Marche
Toscana
Sud
Lazio
Europa
Ovest
Est
Roll up

Sottoinsiemi (MEMO)

- Sottoinsiemi su linea gerarchica

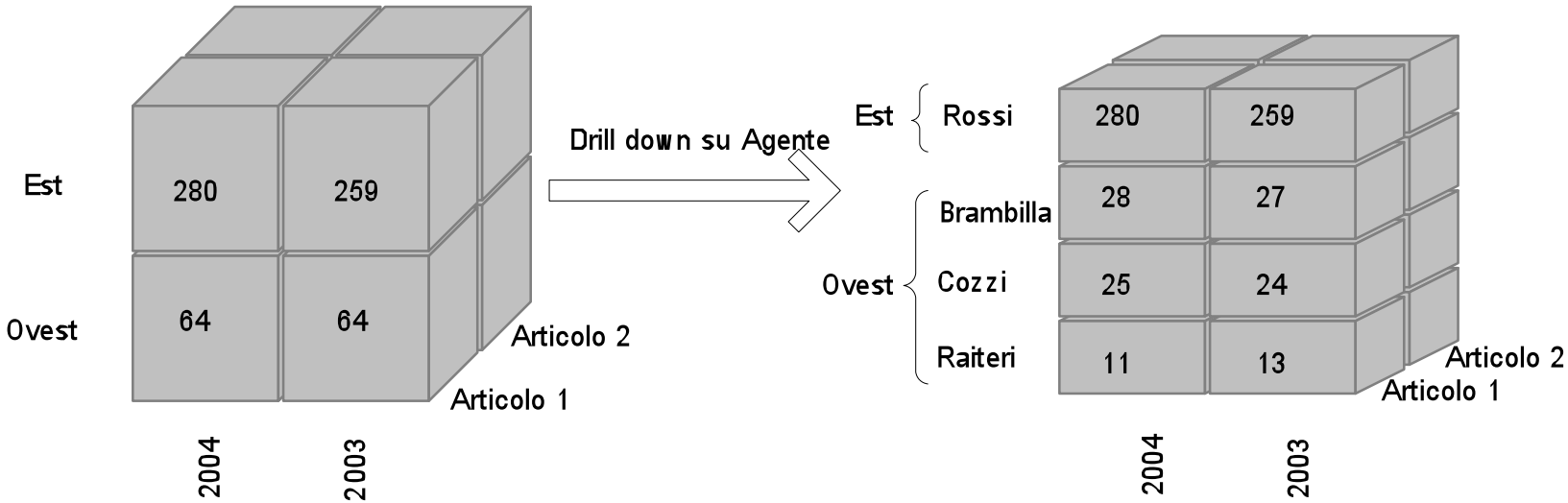


- Sottoinsiemi su gerarchie diverse



Operatori OLAP: Drill down

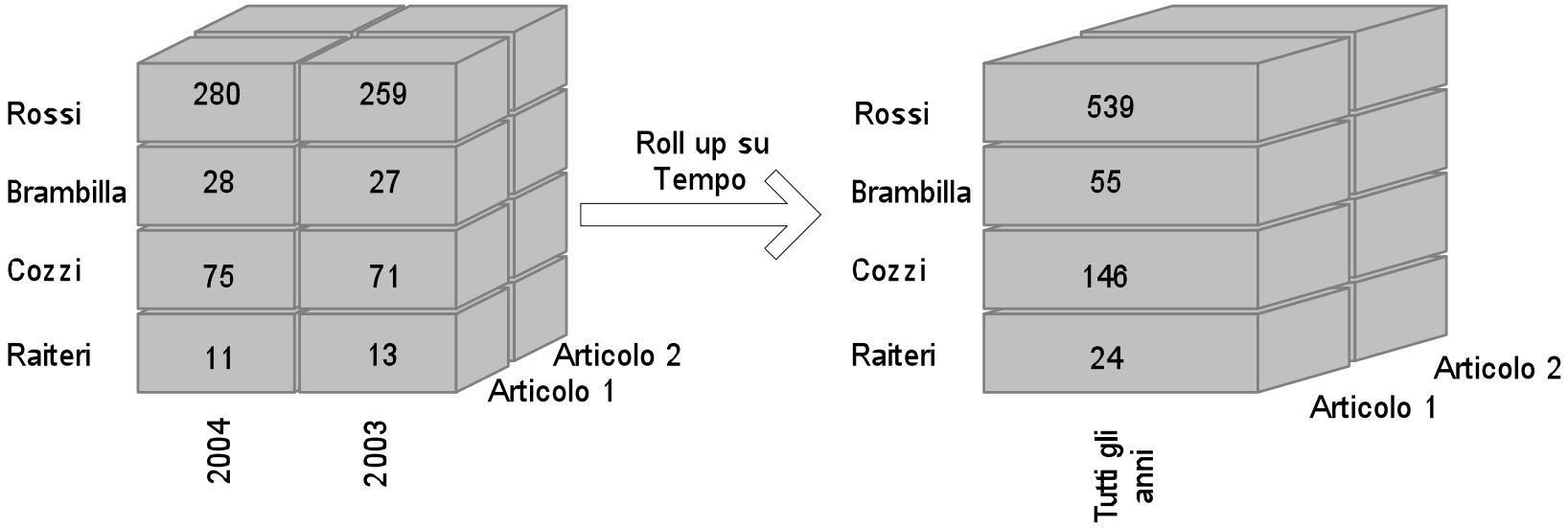
- Dettaglia i dati
  - Scendendo lungo una gerarchia
  - Aggiungendo una dimensione di analisi



- Esempio: mi domando perchè la zona ovest non ha incrementato il fatturato. Drill down su agente mostra che Ranieri ha perso e gli altri hanno guadagnato

Operatori OLAP: Roll up

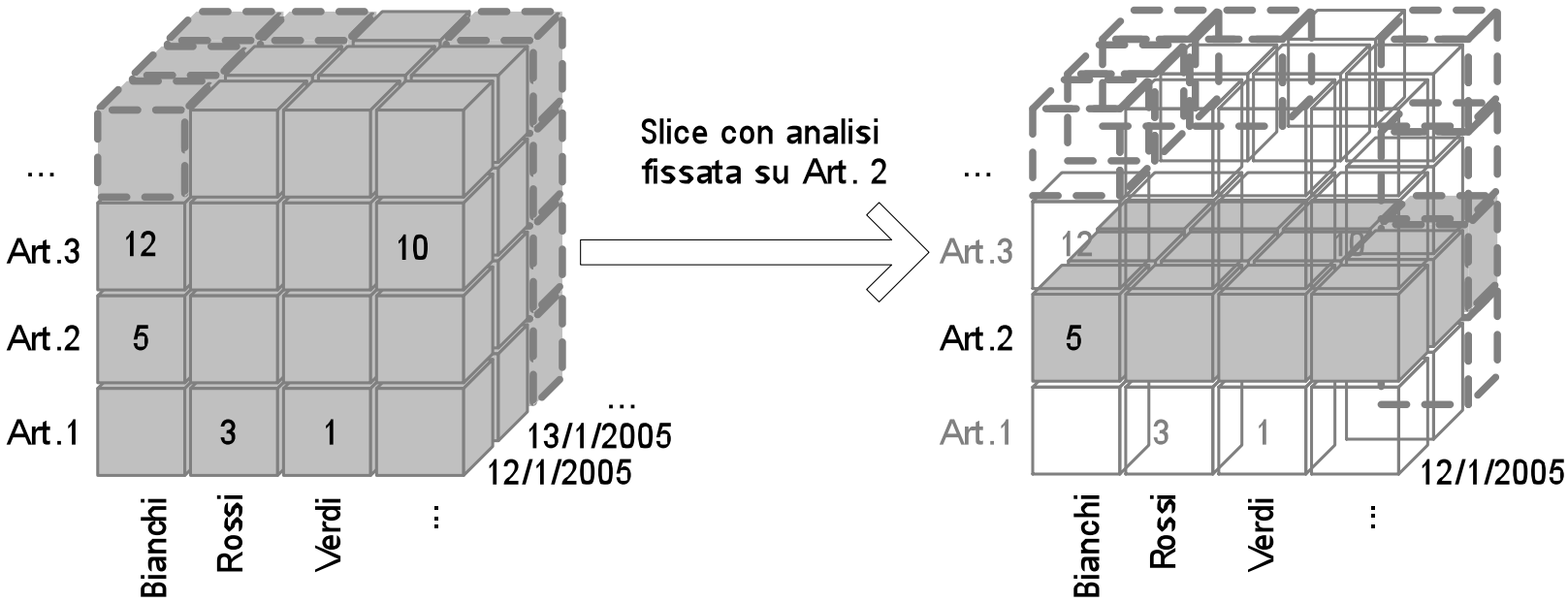
- Sintetizza i dati
  - Percorrendo le gerarchie nella direzione di maggior aggregazione
  - Eliminando una delle dimensioni di analisi



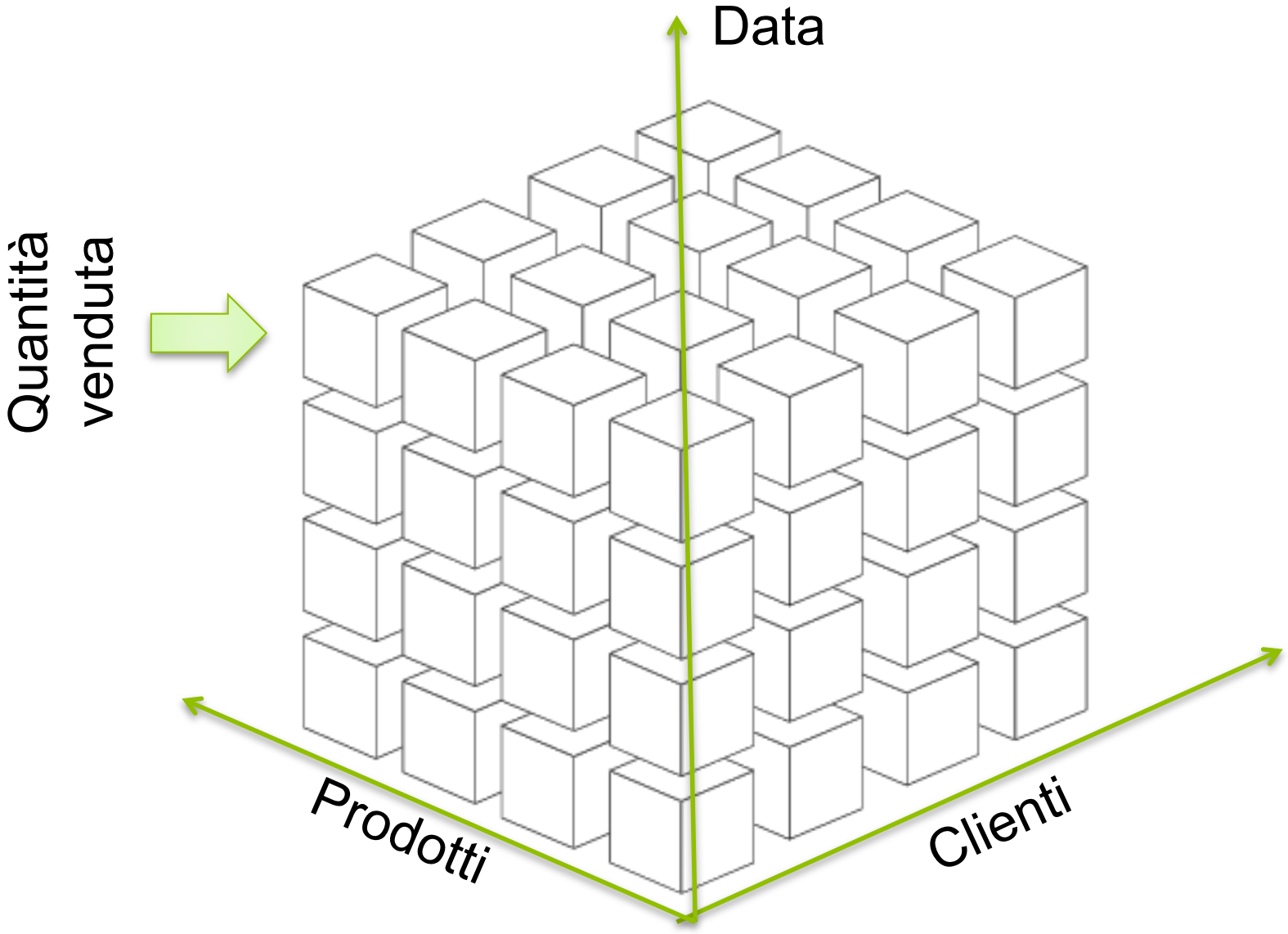
- Esempio: aggregando per anno scopro l'apporto complessivo dell'agente alla società

Operatori OLAP: Slice

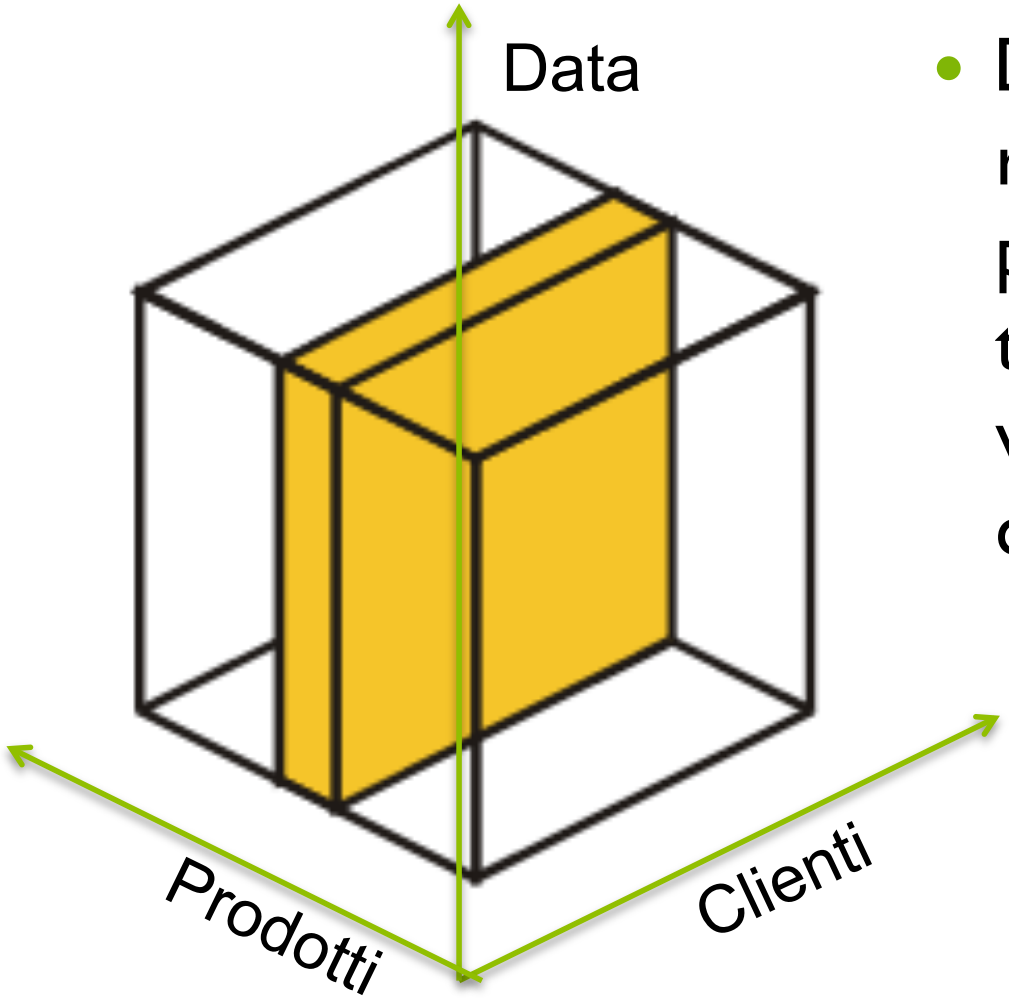
- Fissa il valore di una delle dimensioni base per analizzare la porzione di dati filtrati così ottenuta



Esempio - slice

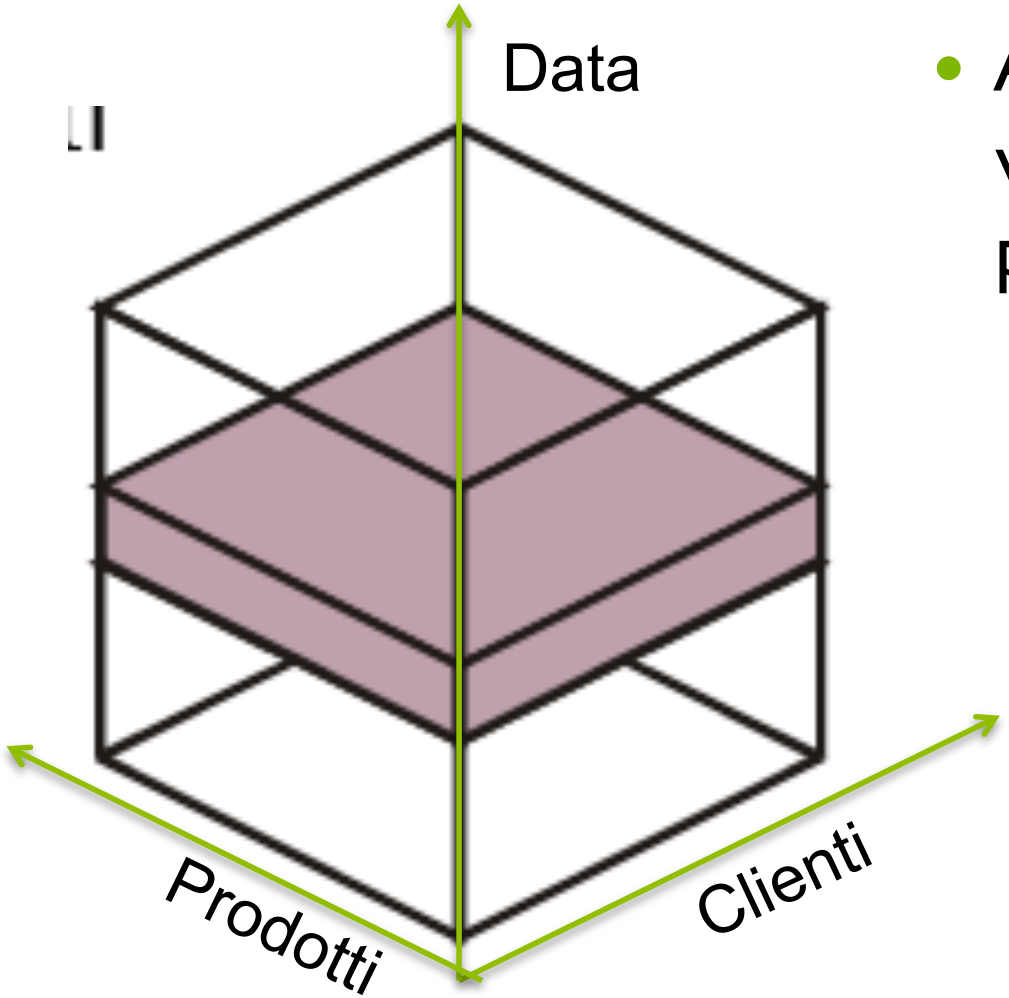


Esempio - slice



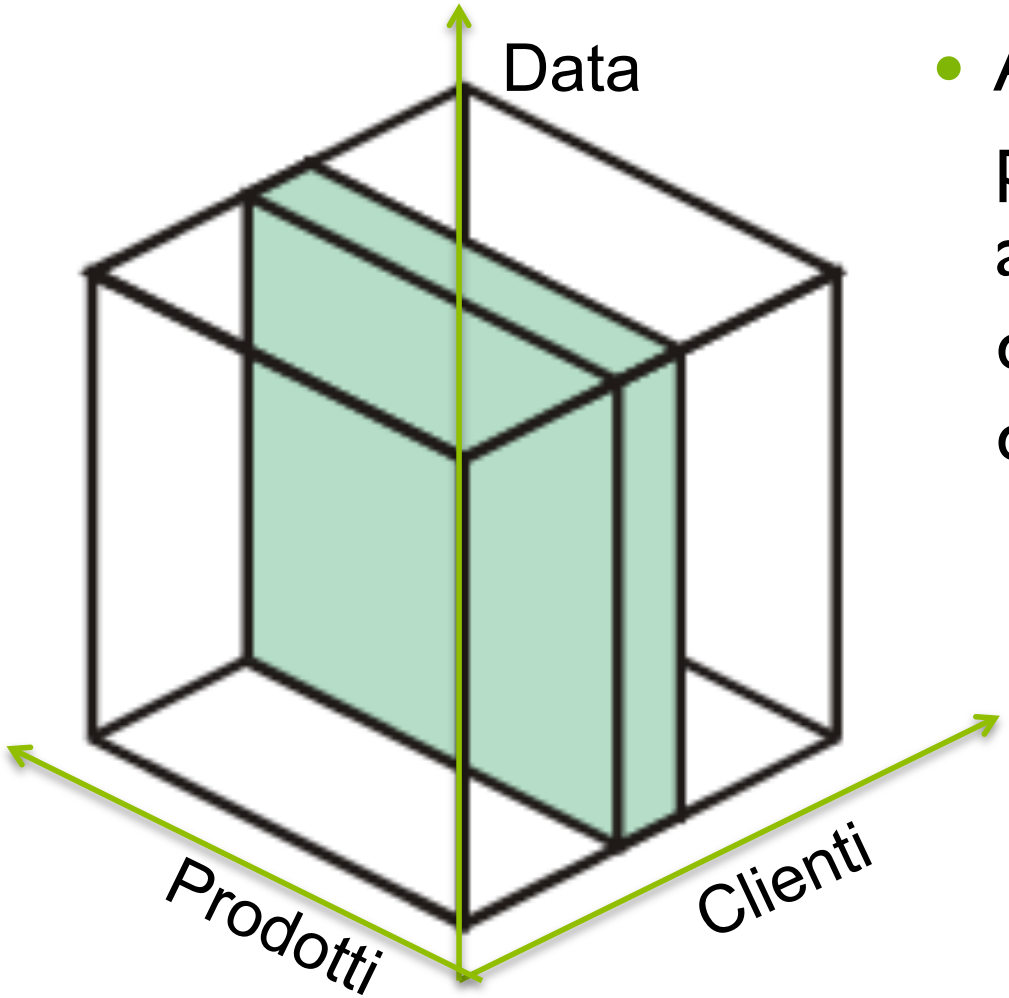
- Dato un prodotto rappresenta le sue performance in termini di quantità vendute per ogni cliente e periodo

Esempio - slice



- Analisi delle quantità vendute in un certo periodo temporale

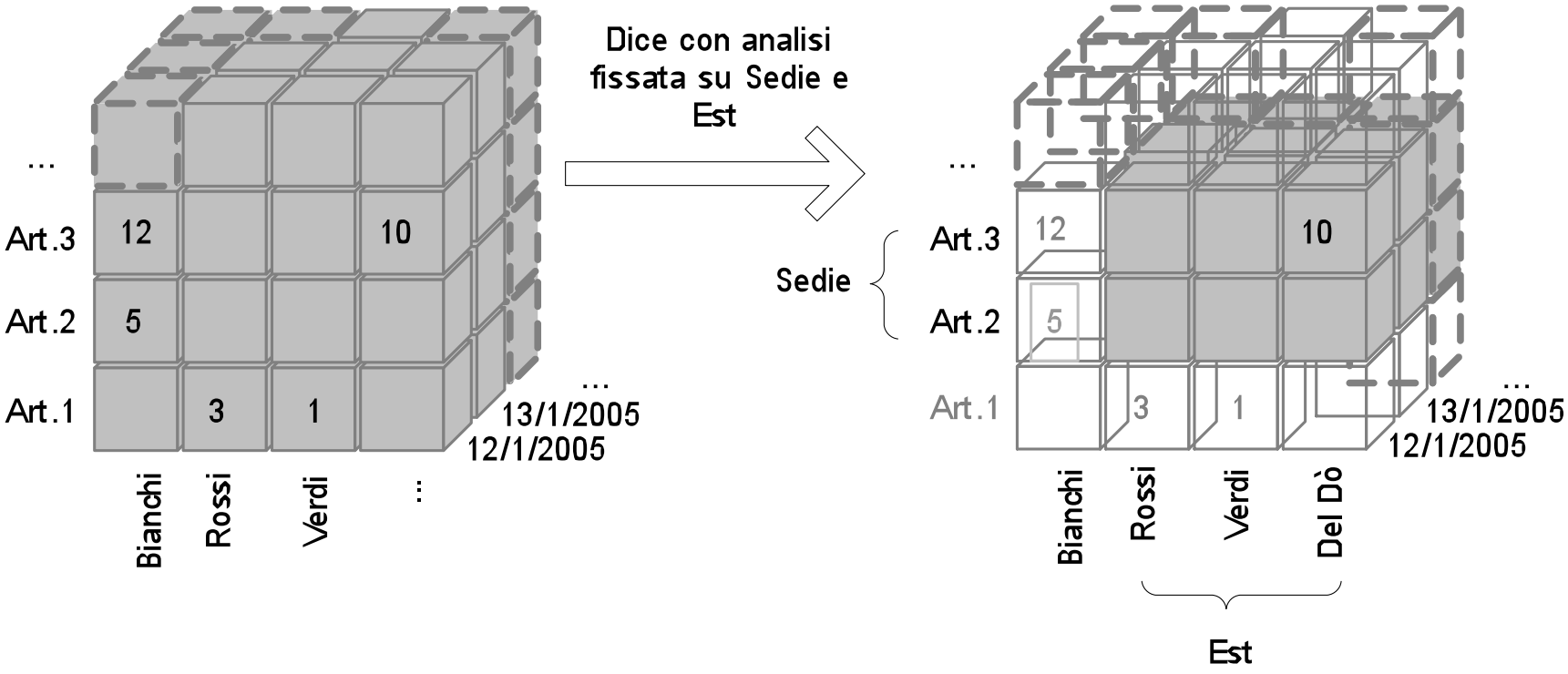
Esempio - slice



- Analisi per ogni prodotto delle abitudini di acquisto da parte di un determinato cliente

Operatori OLAP: Dice

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



Operatori OLAP: Pivot

- Inverte la relazione tra le dimensioni, realizzando una rotazione del cubo nell'analisi
- Particolarmente utile nell'analisi di dati presentati in forma tabellare
- Utile per rendere piu naturali certe misure
  - Esempio: chilometri/litri

Prodotto	**Area**	2003	2004
Articolo 1	Centro	60	56
	Est	203	220
	Ovest	64	64

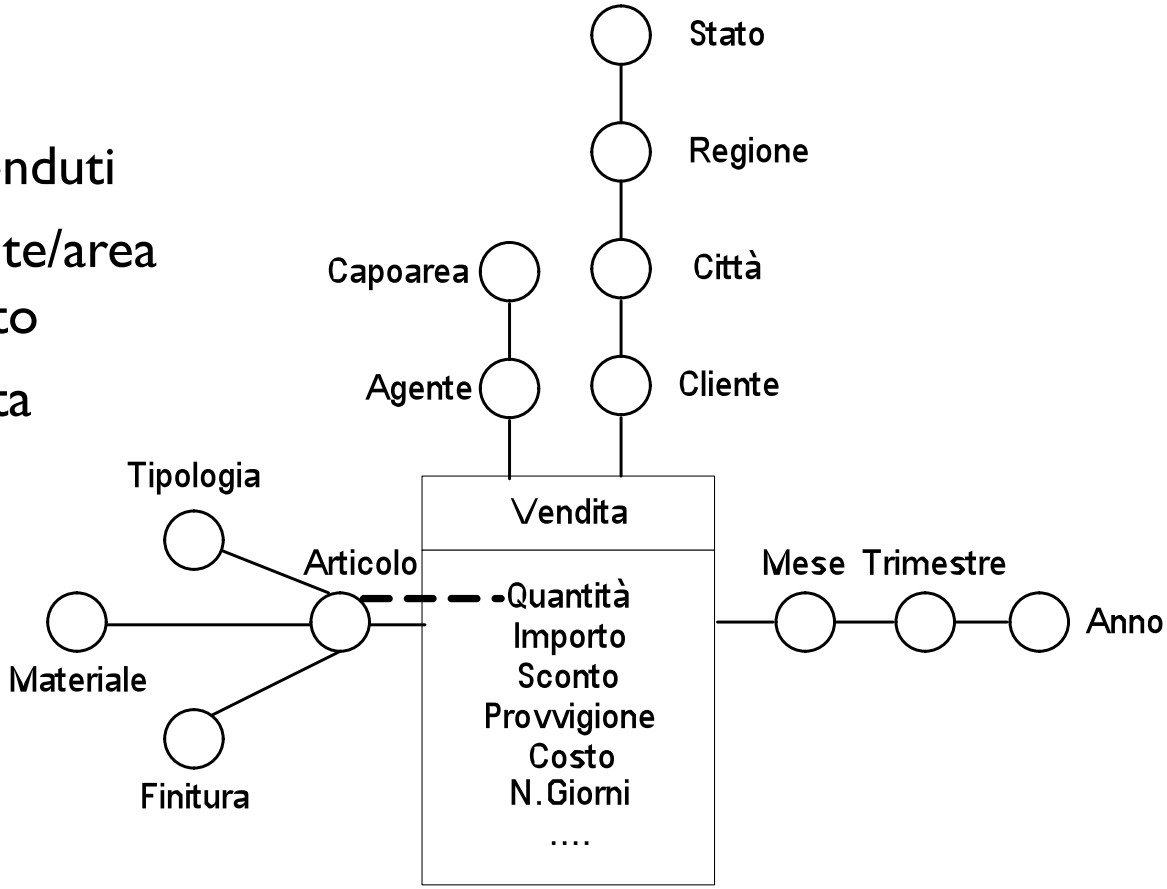
Prodotto	**Anno**	Centro	Est	Ovest
Articolo 1	2003	60	203	64
	2004	56	220	64

Pivoting tra le dimensioni Anno e Area

ESEMPI DI AREE DI APPLICAZIONE

Aree di applicazione: Flusso attivo

- Analisi tipiche
  - Mix di prodotti venduti
  - fatturato per cliente/area geografica/prodotto
  - Servizi post vendita
  - Efficienza rete di distribuzione
  - Abbandoni
- Eventi
  - Documenti del flusso attivo



Esempio di schema di fatto per analisi delle vendite

Contabilità analitica (MEMO)

Piano dei conti

Tipologia Immobile

Piano dei conti

Cantiere

- Posso sapere l'andamento di un centro di costo all'interno dell'altro.
  - Esempio: andamento grattacieli nel cantiere 2

Controllo di gestione: conto economico per cliente/prodotto

- analizza la redditività del singolo cliente, distinta per prodotto
  - Per esempio valuta bontà di una linea di prodotti
  - Struttura tridimensionale (cubo)
  - Ogni dimensione può essere una gerarchia
- Principali dimensioni possibili
  - Clienti
  - Prodotti (e le sue gerarchie)
    - Esempio: prodotti finiti, linea X, Y, Z, ecc
  - Ricavi
  - Costi
    - Esempio: Materiali, lavorazioni, fissi, diretti, indiretti