Data warehousing

Introduzione

- A partire dagli anni novanta è risultato chiaro che i database per i DSS e le analisi di business intelligence vanno separati da quelli operazionali.

- In questa lezione vedremo le caratteristiche di tali database (data warehouse e data mart), analizzando le differenze rispetto ai sistemi operazionali.

- Analizzeremo gli aspetti funzionali dei data warehouse, dando alcuni dettagli relativamente alla loro progettazione.

Un po' di storia

- Nel 1992, William Inmon pubblica il libro *Building the data warehouse*.
- Inmon propone di creare un grande magazzino di dati in cui riversare tutti i dati dell'azienda.
- Dalla aggregazione dei dati è possibile effettuare statistiche su dati aggregati utili ai fini delle analisi.
- Nel 1996, Ralph Kimball propone di eliminare il magazzino e di effettuare le analisi direttamente sulle basi di dati dipartimentali.

Definizione di data warehouse

Magazzino per i dati

+

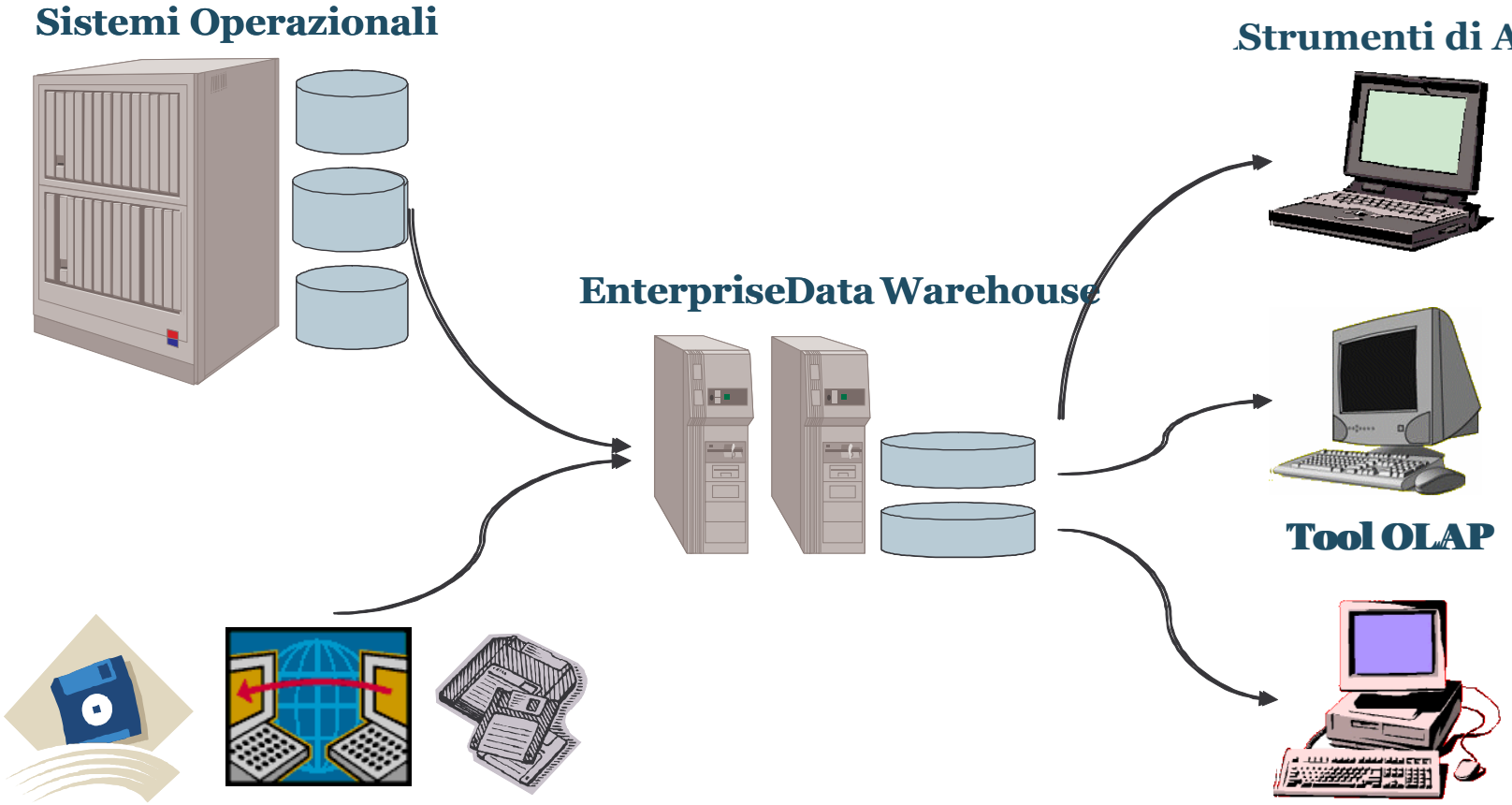
Procedure di acquisizione,
organizzazione e elaborazione

=

Data Warehouse

- Data warehousing = l'insieme delle attività per la progettazione, realizzazione ed uso dei data warehouse.

Architettura di data warehouse



Motivazione

- Molteplici ragioni inducono a realizzare data warehouse (OLAP) separati dai database transazionali (OLTP):
  - **Integrazione**: i data warehouse integrano dati da fonti diverse
  - **Qualità**: i dati influenzano i risultati
  - **Efficienza**: le analisi devono essere rapide
  - **Estensione temporale**: i dati devono avere sufficiente profondità storica.

Caratteristiche

- Collezione di dati che soddisfa le seguenti proprietà:
  - **Orientata ai soggetti**: considera i dati di interesse ai soggetti dell'organizzazione e non quelli rilevanti ai processi organizzativi
  - **Integrata:** a livello aziendale e non dipartimentale
  - **Storicizzata:** con ampio orizzonte temporale
  - **Consolidata (aggregata)**: non interessa "chi" ma "quanti"
  - **Denormalizzata**: le ridondanze permettono tempi di risposta più rapidi.
  - **Fuori linea**: dati aggiornati periodicamente

Differenze OLTP-OLAP

	OLTP	OLAP
Funzione	gestione giornaliera	supporto alle decisioni
Progettazione	orientata alle applicazioni	orientata al soggetto
Frequenza	giornaliera	sporadica
Dati	recenti, dettagliati	storici, riassuntivi, multidim.
Sorgente	singola DB	DB multiple
Uso	ripetitivo	ad hoc
Accesso	read/write	Read
Flessibilità accesso	programmi precompilati	generatori di query
# record acceduti	decine	Migliaia
Utenti	operatori	Manager
# utenti	migliaia	Centinaia
Tipo DB	singola	multiple, eterogenee
Performance	alta	Bassa
Dimensione DB	100 MB - GB	100 GB – TB

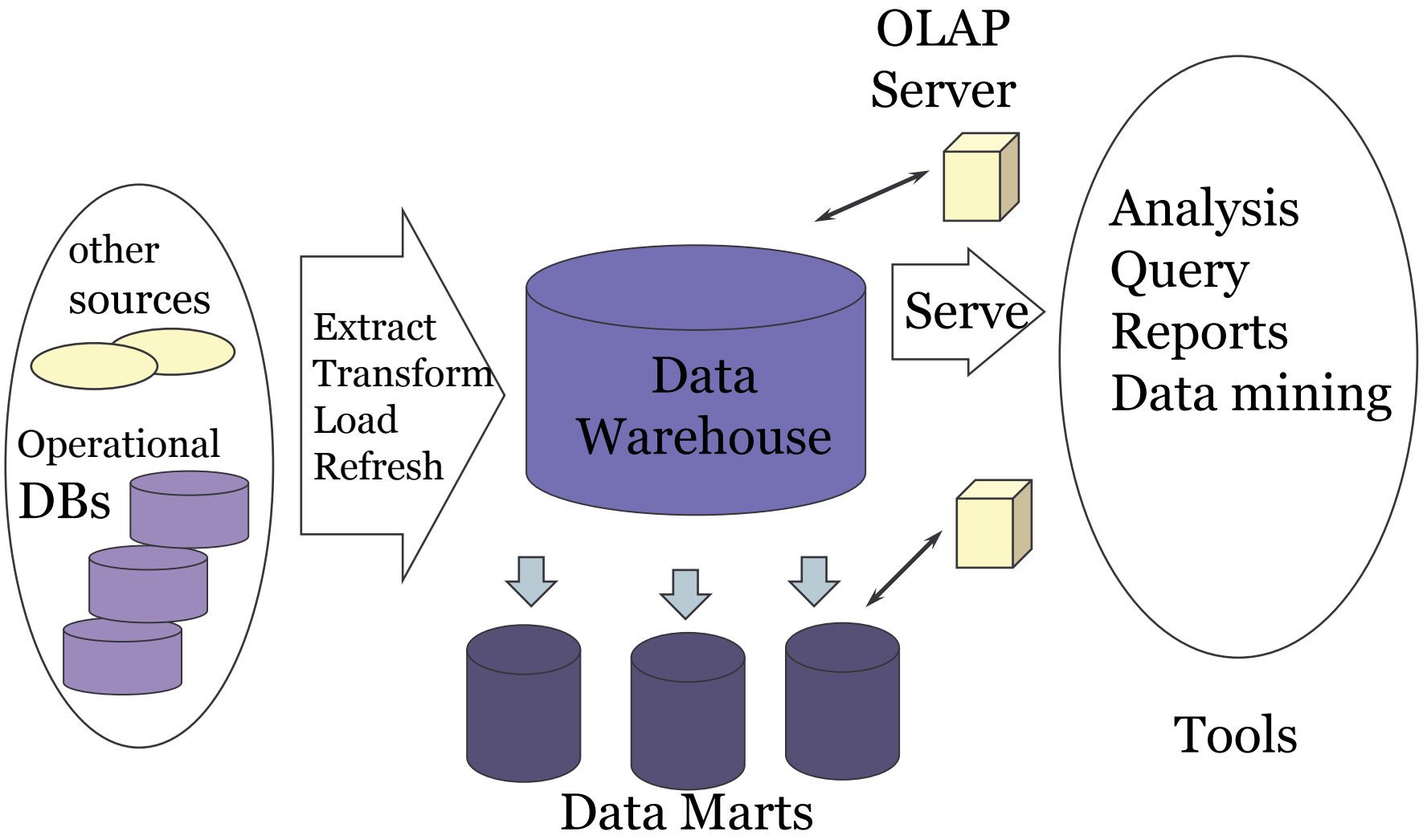
Data mart

- Data warehouse dipartimentale che raccoglie i dati di una specifica funzione aziendale.
- Sistema specializzato che mette insieme i dati necessari ad un dipartimento.
- Implementato creando viste specifiche alle applicazioni.
  - Un data mart di marketing contiene informazioni relative ai clienti e alle transazioni di vendita, risultati di campagne,...
- Sottoinsiemi materializzati di viste dipartimentali che focalizzano su soggetti determinati.

Qualità dei dati

- Verificare, preservare e incrementare la qualità dei dati rappresenta una preoccupazione costante per i responsabili della progettazione e manutenzione.
- Principali inconvenienti:
  - Dati non corretti
  - Dati non aggiornati
  - Dati inesistenti
- I principali fattori che influenzano le analisi di BI riguardano l'accuratezza, completezza, consistenza, attualità, non ridondanza, rilevanza, interpretabilità e accessibilità dei dati.

Architettura del data warehouse



Strumenti ETL

- Extract, Transform, Load: insieme degli strumenti che permettono di estrarre, trasformare e caricare i dati
- Nella prima fase i dati vengono estratti dai database operazionali.
  - Estrazione iniziale e incrementale
- Nella fase di trasformazione vengono eliminate le inconsistenze, le duplicazioni, i valori inammissibili.
- I dati corretti e trasformati vengono caricati nel data warehouse.

Realizzazione

- **Top-Down**: Si parte dalle analisi per determinare I dati necessari.
  - Pro: maggiore probabilità di successo,
  - Contro: tempi lunghi.
- **Bottom-down**: Si parte dai dati e si arriva alle analisi
  - Pro: tempi brevi
  - Contro: maggiore probabilità di errore.
- **Ibrida**: Si procede in entrambe le direzioni, realizzando prototipi successivi delle varie parti del sistema
  - Pro: risultati immediati
  - Contro: integrazione incerta.

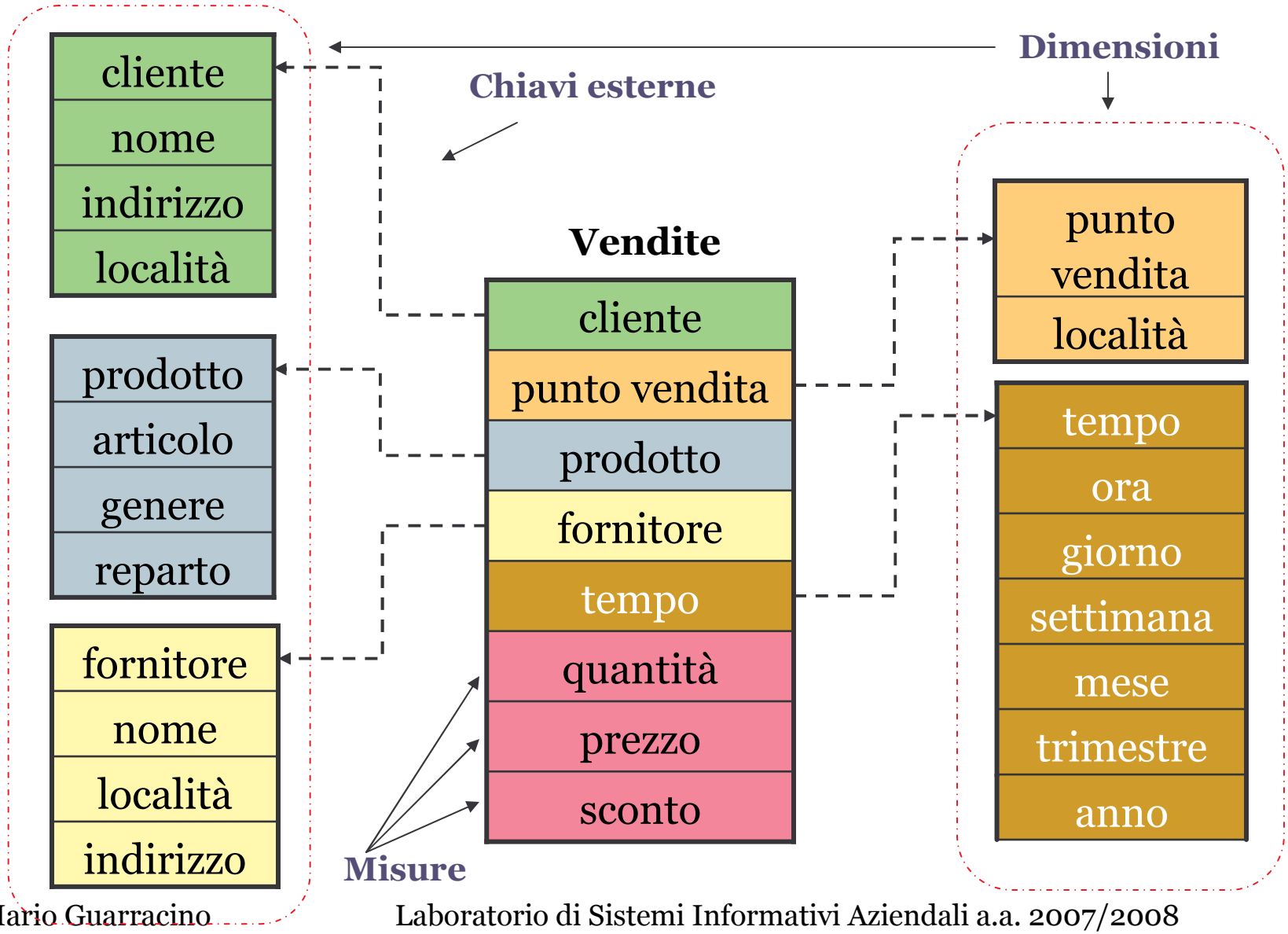
Cubi e analisi multidimensionali

- La modellazione concettuale di un data warehouse utilizza:
- **Schema a stella**: Un singolo oggetto (tabella dei fatti) in mezzo connessa ad un numero di oggetti (tabella delle dimensioni).
- **Schema a fiocco di neve**: Un raffinamento dello schema a stella in cui la gerarchia dimensionale è rappresentata esplicitamente (normalizzando le tabelle delle dimensioni).
- **Galassie**: tabelle dei fatti multiple condividono la tabelle delle dimensioni.

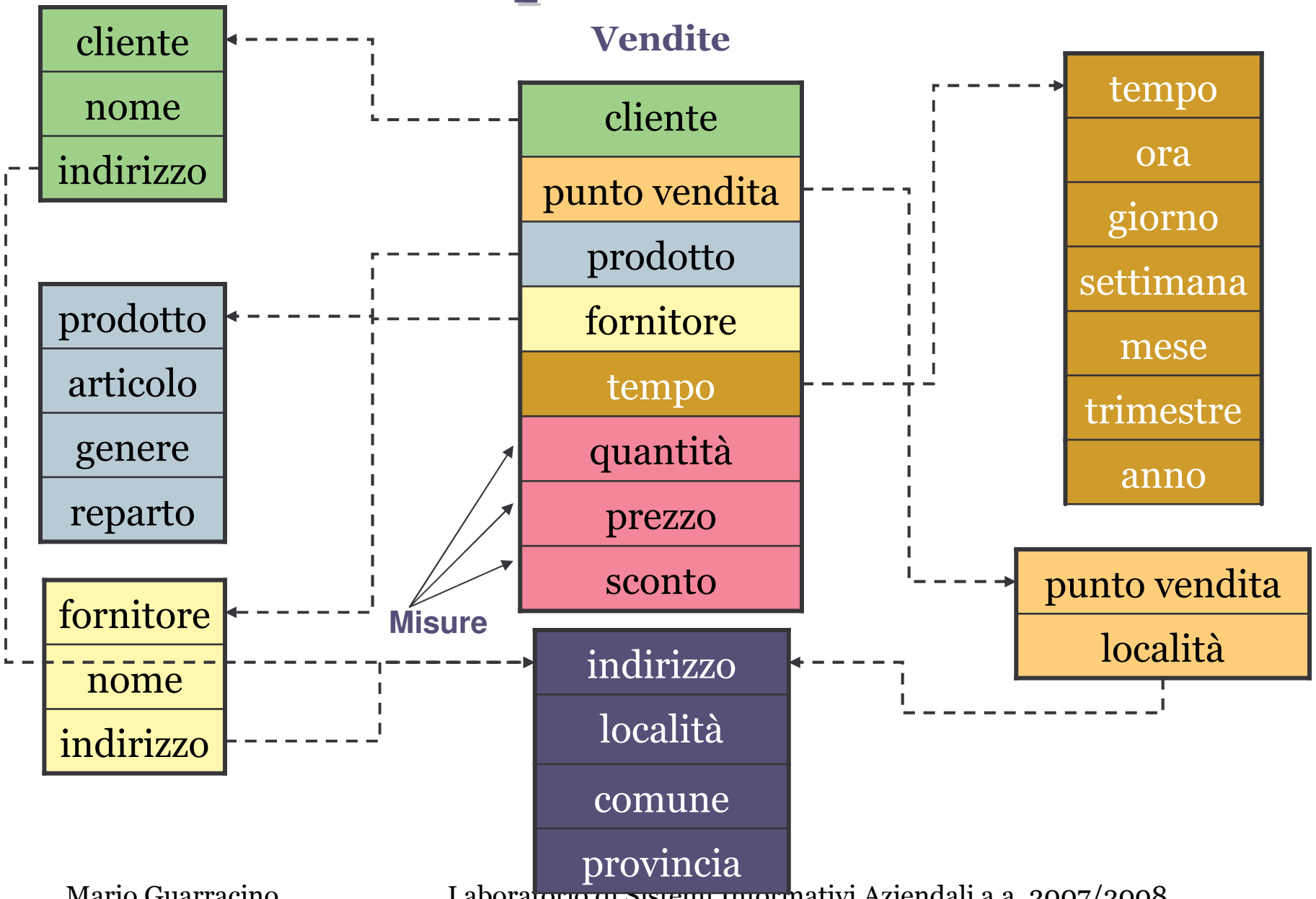
Schema a stella

- Un **fatto** è un evento di interesse per l'impresa
  - vendite, spedizioni, acquisti,...
- Le **misure** sono attributi che descrivono quantitativamente il fatto da diversi punti di vista
  - numero di unità vendute, prezzo unitario, sconto,...
- Una **dimensione** determina la granularità minima di rappresentazione dei fatti
  - il prodotto,il punto vendita, la data
- Una **gerarchia** determina come le istanze di un fatto possono essere aggregate e selezionate - descrive una dimensione.

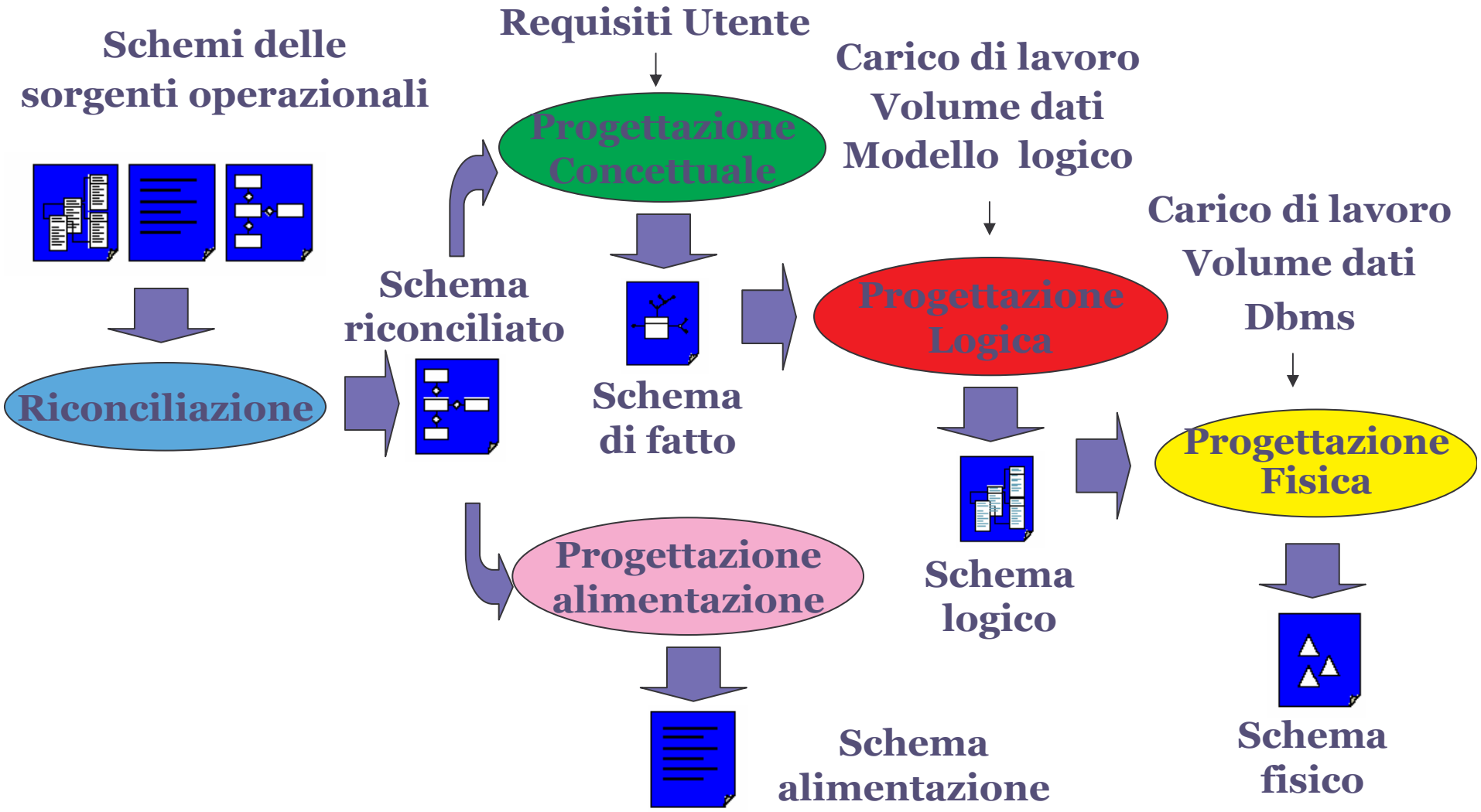
Esempio di schema a stella



Esempio di schema a fiocco



Progettazione dei data mart

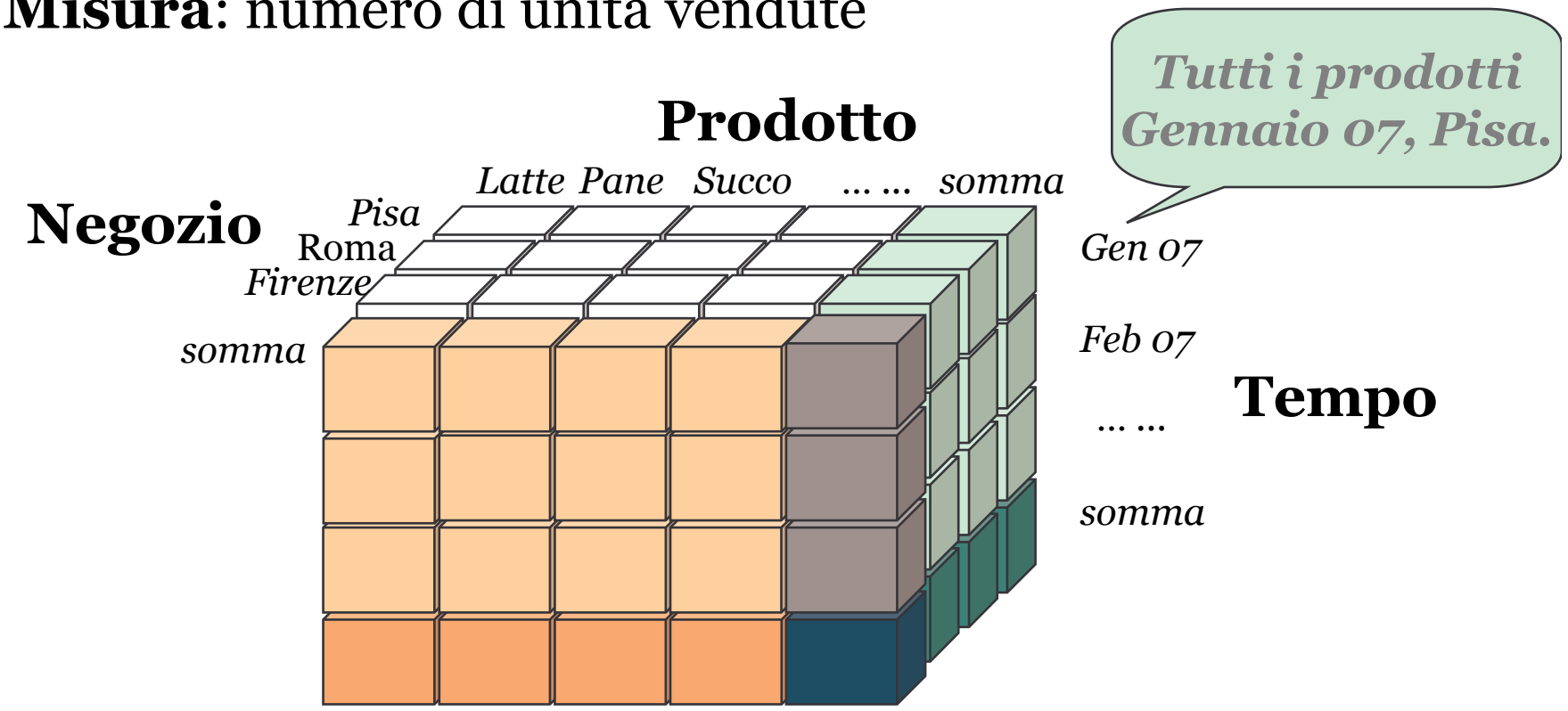


Cubi di dati

- Una tabella dei fatti collegata a *n* tabelle delle dimensioni può essere rappresentata mediante un cubo di dati a *n* dimensioni.
- Ogni dimensione contiene una gerarchia di valori e una cella del cubo contiene i valori aggregati
  - count, sum, max, …
- Essi rappresentano una naturale evoluzione dei fogli elettronici.

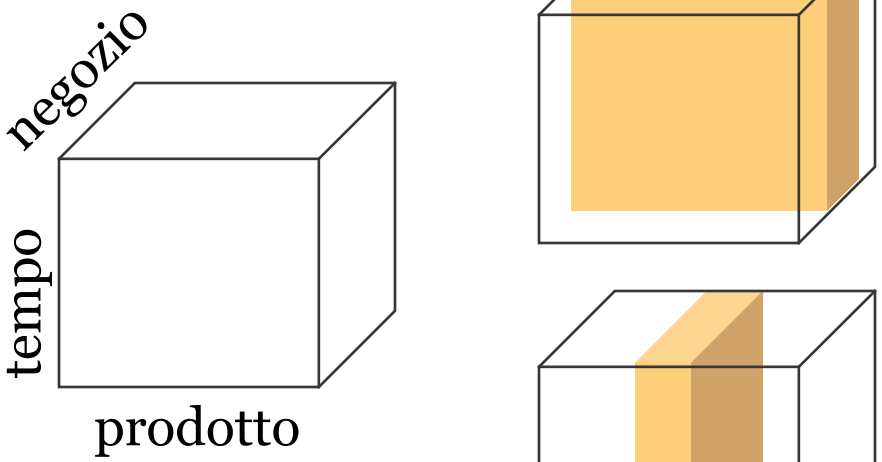
Esempio

- **Tabella dei fatti**: Vendita
- **Dimensioni**: {tempo, prodotto, negozio}
- **Misura**: numero di unità vendute

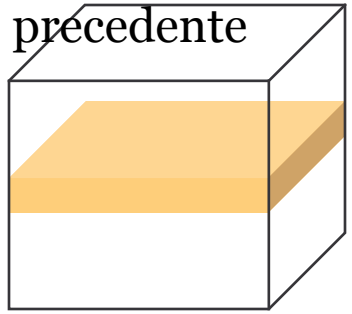


Esempio

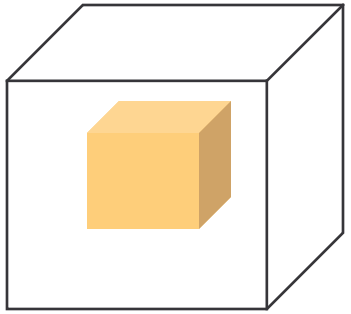
Il manager regionale esamina la vendita di tutti i prodotti in tutti i periodi relativamente ai propri mercati



Il manager finanziario esamina la vendita di tutti i prodotti, in tutti i mercati relativamente al periodo corrente e quello precedente



Il manager di prodotto esamina la vendita di un prodotto in tutti i periodo e in tutti i mercati



Il manager strategico si concentra su una categoria di prodotti, un'area regionale e un orizzonte temporale medio

Operazioni sui cubi

- **Roll up**: riassunto dei dati
  - *il volume totale di vendite per categoria di prodotto e per regione*
- **Roll down**, **drill down**, **drill through**: passa da un livello di dettaglio basso ad un livello di dettaglio alto
  - *per un particolare prodotto, trova le vendite dettagliate per ogni venditore e per ogni data*
- **Slice and dice**: select & project
  - *Vendite delle bevande nel sud negli ultimi 6 mesi*
- **Pivot**: riorganizza il cubo

Sommario

- Abbiamo visto:
  - Cosa sono i data warehouse;
  - Cosa siano i data mart;
  - Quali siano le architetture dei data warehouse;
  - Cosa siano le tabelle dei fatti, le dimensioni e gli indici;
  - Cosa siano gli schemi a stella e a fiocco;
  - Come si usino i cubi ed analisi multidimensionali;

Nella prossima lezione

- Modelli matematici per le decisioni:
  - Struttura dei modelli matematici
  - Fasi di sviluppo
  - Classi principali di modelli