

# Introduzione al data warehousing

Luca Cabibbo  
marzo 2010

Luca Cabibbo

## Che cosa è un data warehouse?

Un **data warehouse** è una base di dati

- collezione di dati di grandi dimensioni, persistente e condivisa gestita in maniera efficace, efficiente ed affidabile

con delle caratteristiche “peculiari”

- utilizzata principalmente per il supporto alle decisioni direzionali
- integrata — aziendale e non dipartimentale
- con dati storici — con un ampio orizzonte temporale
- con dati tipicamente aggregati — per effettuare stime
- fuori linea — aggiornata periodicamente
- mantenuta separatamente dalle basi di dati operazionali

## Motivazioni

I sistemi informatici permettono di aumentare la produttività delle organizzazioni automatizzandone la gestione quotidiana dei **processi operativi**

- vendite nelle catene di supermercati
- instradamento e la contabilizzazione delle telefonate

Questi dati — se opportunamente **accumulati** e **analizzati** — possono essere utilizzati per supportare i **processi gestionali** e **direzionali**, ovvero per la pianificazione e il supporto alle decisioni

- promozioni dei prodotti
- offerta di contratti diversificati

Perché?

- una corretta gestione dei dati storici può essere occasione di un grande vantaggio competitivo

## Analisi dei dati

Un data warehouse ha lo scopo di supportare le decisioni direzionali, ad esempio permettendo di calcolare (in modo efficiente) le seguenti interrogazioni

- quali sono stati i volumi di vendita dello scorso anno per regione e categoria di prodotto?
- quali prodotti hanno aumentato il livello delle vendite a fronte di una certa offerta promozionale?
- qual è stata la profittabilità delle campagne promozionali degli ultimi cinque anni?
- quali prodotti vanno pubblicizzati e venduti in offerta nella prossima campagna promozionale estiva?

## OLTP

I sistemi di gestione di basi di dati relazionali sono normalmente ottimizzati per supportare le operazioni transazionali (OLTP, **On Line Transaction Processing**)

- le transazioni sono predefinite e di breve durata
- i dati di interesse sono dettagliati, aggiornati e recenti
- i dati risiedono su una unica base di dati
- leggono e/o modificano pochi record
- le proprietà “transazionali” sono critiche
- architettura (principalmente) centralizzata

## OLAP

I sistemi di supporto alle decisioni dovrebbero invece supportare l’elaborazione analitica (OLAP, **On-Line Analytical Processing**), che ha le seguenti caratteristiche

- le interrogazioni sono complesse e casuali
- i dati di interesse sono tipicamente storici e aggregati
- i dati possono provenire da più basi di dati — possibilmente non omogenee
- leggono un numero enorme di record — non scrivono mai
- le risposte alle interrogazioni sono attese in linea
- la visualizzazione dei dati è fondamentale
- architettura client-server

## OLTP e OLAP

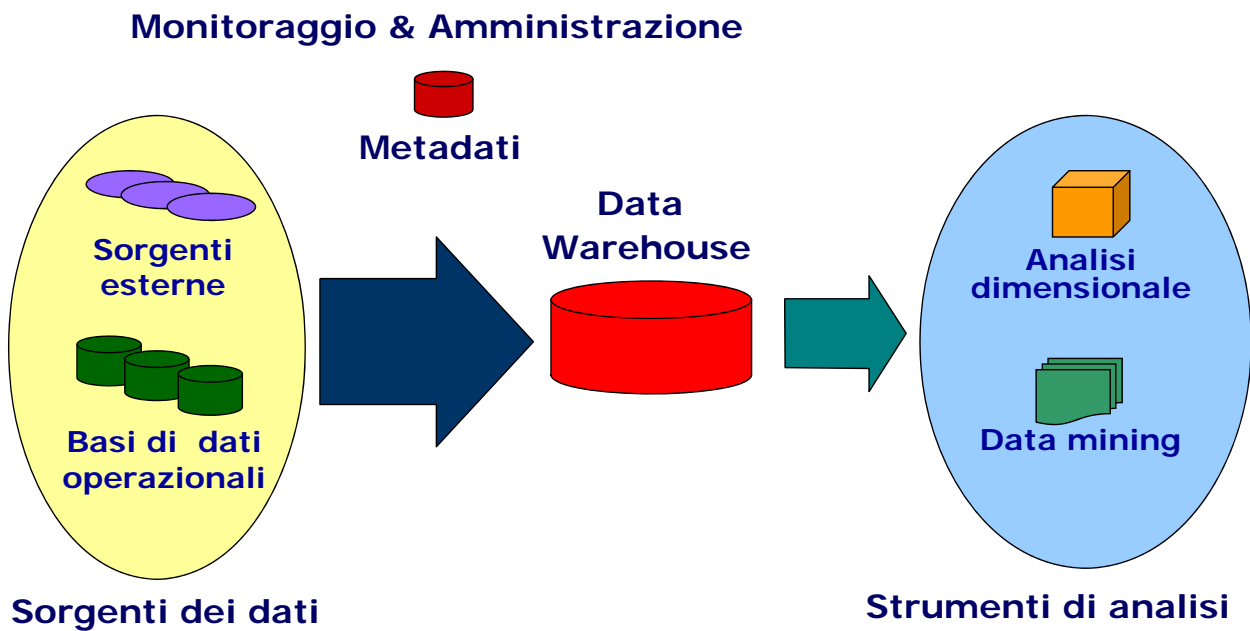
	OLTP	OLAP
<b>Utente</b>	impiegato	dirigente
<b>Funzione</b>	operazioni giornaliere	supporto alle decisioni
<b>Progettazione</b>	orientata all'applicazione	orientata ai dati
<b>Dati</b>	correnti, aggiornati, dettagliati, relazionali, omogenei	storici, aggregati, multidimensionali, eterogenei
<b>Uso</b>	ripetitivo	casuale
<b>Accesso</b>	read-write, indicizzato	read, sequenziale
<b>Unità di lavoro</b>	transazione breve	interrogazione complessa
<b>Record acc.</b>	decine	milioni
<b>N. utenti</b>	migliaia	centinaia
<b>Dimensione</b>	100MB - 1GB	100GB - 1TB
<b>Metrica</b>	throughput	tempo di risposta

## Definizione di data warehouse

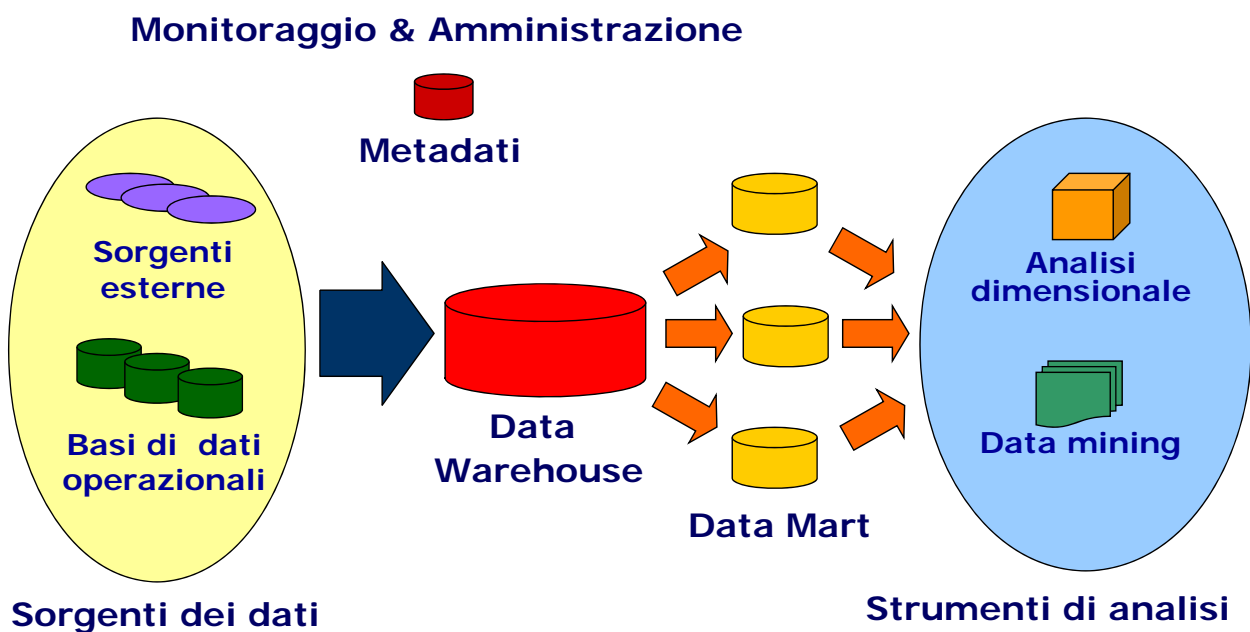
Un **data warehouse** è una base di dati

- utilizzata principalmente per il supporto alle decisioni direzionali
- integrata — aziendale e non dipartimentale
- con dati storici — con un ampio orizzonte temporale, e indicazione di almeno un elemento di tempo
- con dati tipicamente aggregati — per effettuare stime
- fuori linea — i dati sono aggiornati periodicamente
- mantenuta separata dalle basi di dati operazionali

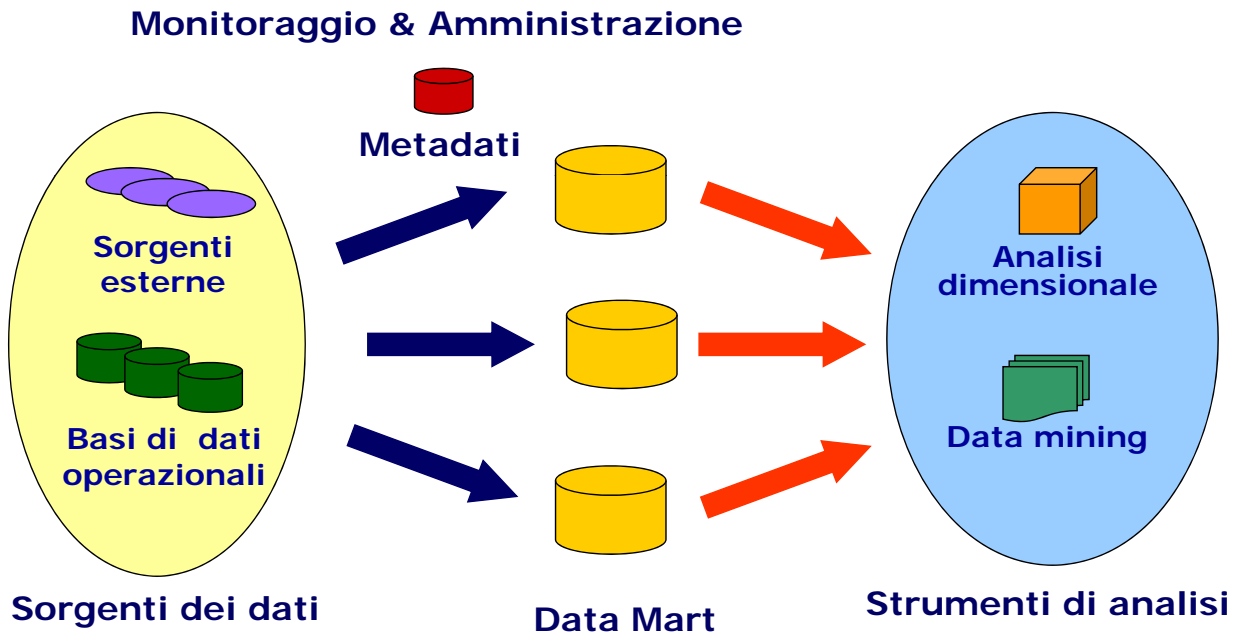
# Architettura generale per il data warehousing



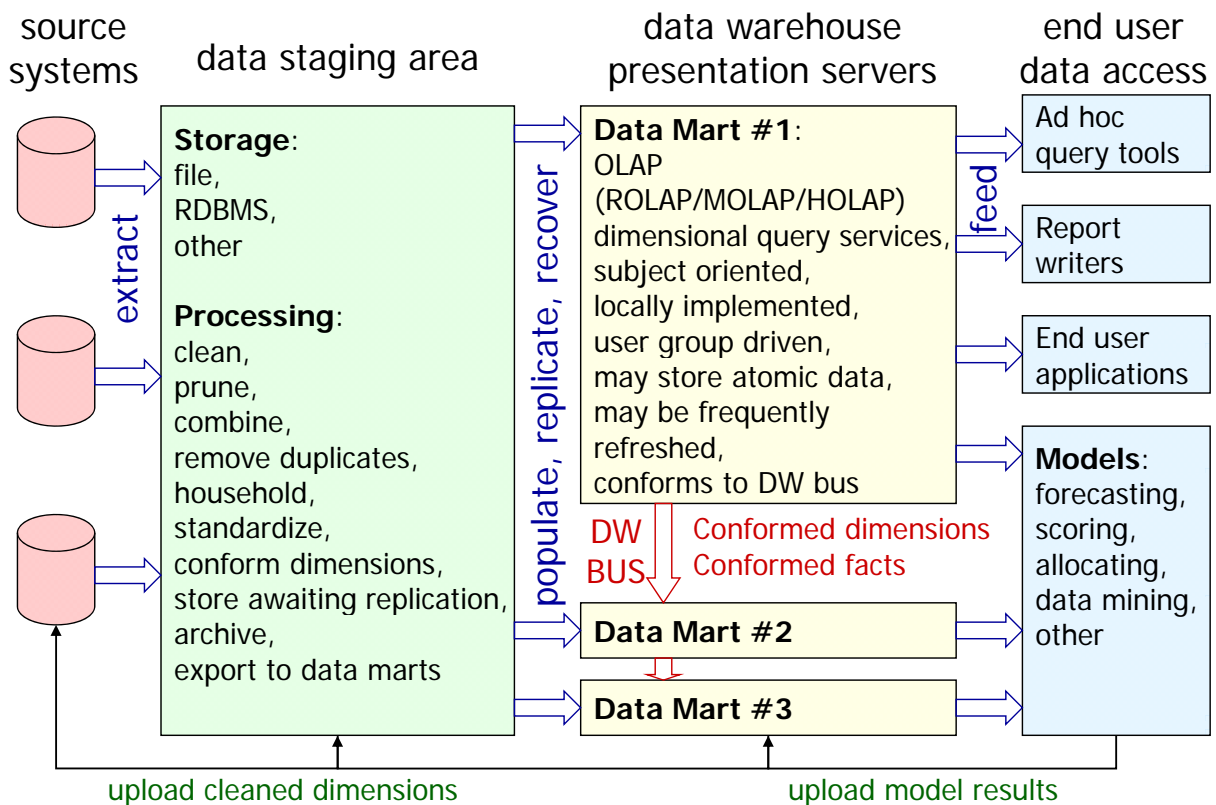
# Architettura per il data warehousing (Inmon)



# Architettura per il data warehousing (Kimball)



# Elementi di un data warehouse



## Dati multidimensionali

L'analisi dei dati avviene su dati rappresentati in forma multidimensionale, ovvero organizzati mediante i seguenti concetti

- **fatto** (o processo)
  - un concetto sul quale centrare l'analisi
- **misura**
  - una proprietà atomica o misura di un fatto da analizzare
  - le misure sono solitamente valori numerici e additivi su un dominio continuo
- **dimensione**
  - una prospettiva rispetto alla quale effettuare l'analisi
  - le dimensioni descrivono domini discreti, solitamente organizzati in livelli di aggregazione

## Dati multidimensionali - esempi

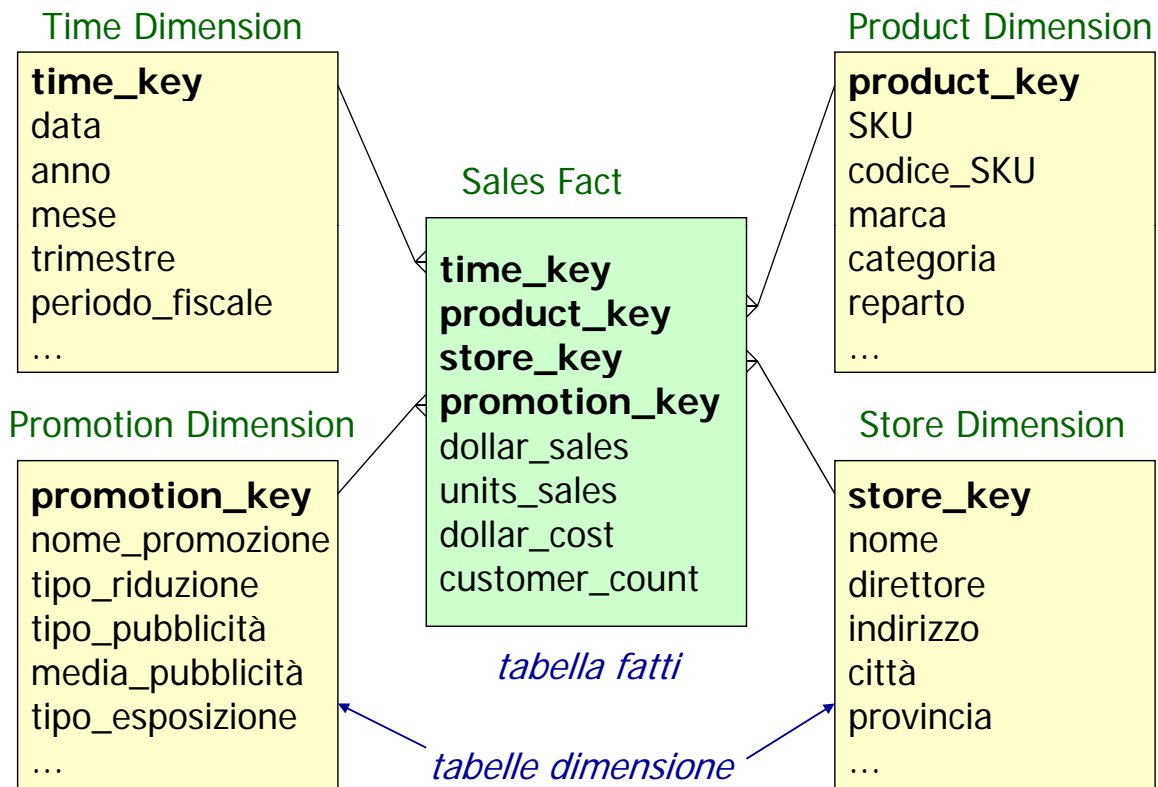
Data mart delle vendite

- fatto: vendite dei prodotti, giornaliera, per negozio
- dimensioni: prodotto, tempo (giorno), negozio, promozione
- misure: quantità venduta, incasso, costo, conteggio dei clienti

Data mart delle telefonate

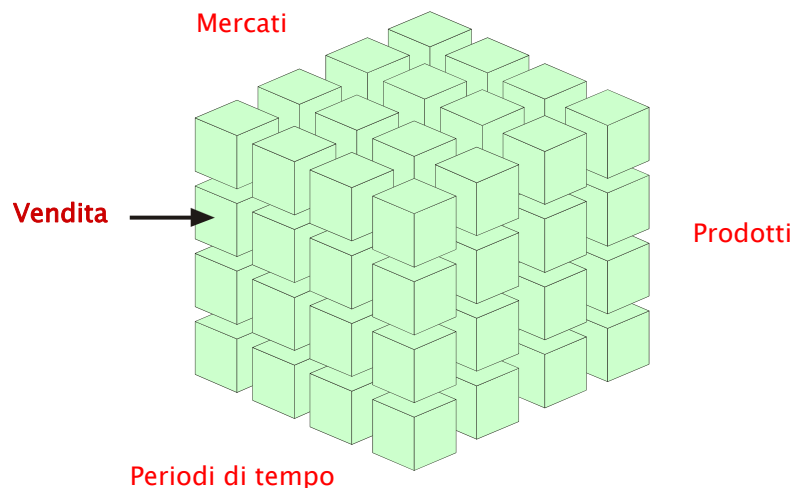
- fatto: telefonata
- dimensioni: chiamante, chiamato, tariffa, tempo (giorno), tempo (ora del giorno)
- misure: durata, costo

## Rappresentazione di dati multidimensionali



## Rappresentazione multidimensionale dei dati

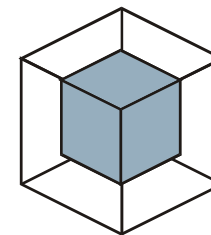
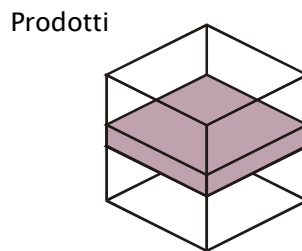
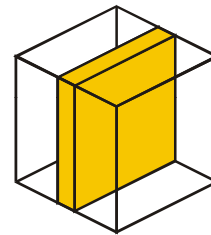
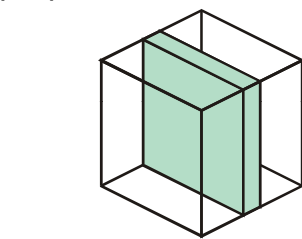
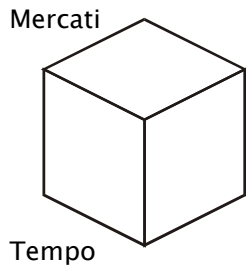
Gli analisti sono abituati a ragionare in termini di dimensioni e misure — non di schemi, tabelle e record



## Viste su dati multidimensionali

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

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



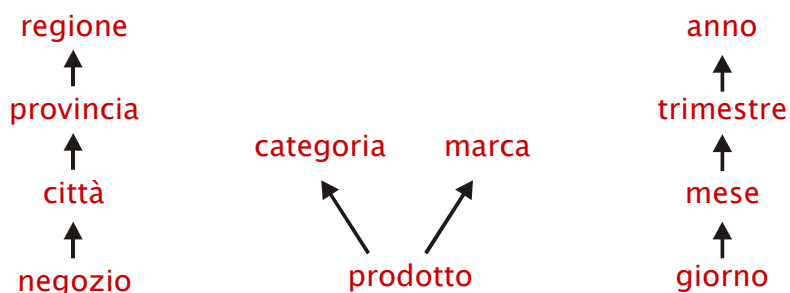
Il manager di prodotto esamina la vendita di un prodotto in tutti i periodi e in tutti i mercati

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

## Dimensioni e gerarchie di livelli

Ciascuna dimensione è organizzata in una gerarchia che rappresenta i possibili **livelli** di aggregazione per i dati

- negozio, città, provincia, regione
- prodotto, categoria, marca
- giorno, mese, trimestre, anno



## Operazioni classiche su dati multidimensionali

**Roll up** — aggrega i dati (rispetto all'interrogazione corrente), ovvero mostra dati a un maggior livello di aggregazione

**Drill down** — disaggrega i dati (rispetto all'interrogazione corrente), ovvero mostra dati a un minor livello di aggregazione

**Drill across** — combina i dati associati a più fatti

**Slice & dice** — seleziona e proietta — solitamente su un piano bidimensionale

**Pivot** — re-orienta il cubo

## Ciclo di vita dimensionale

