

# Data Warehousing

Paolo Atzeni

(con materiale di Luca Cabibbo e Riccardo Torlone)

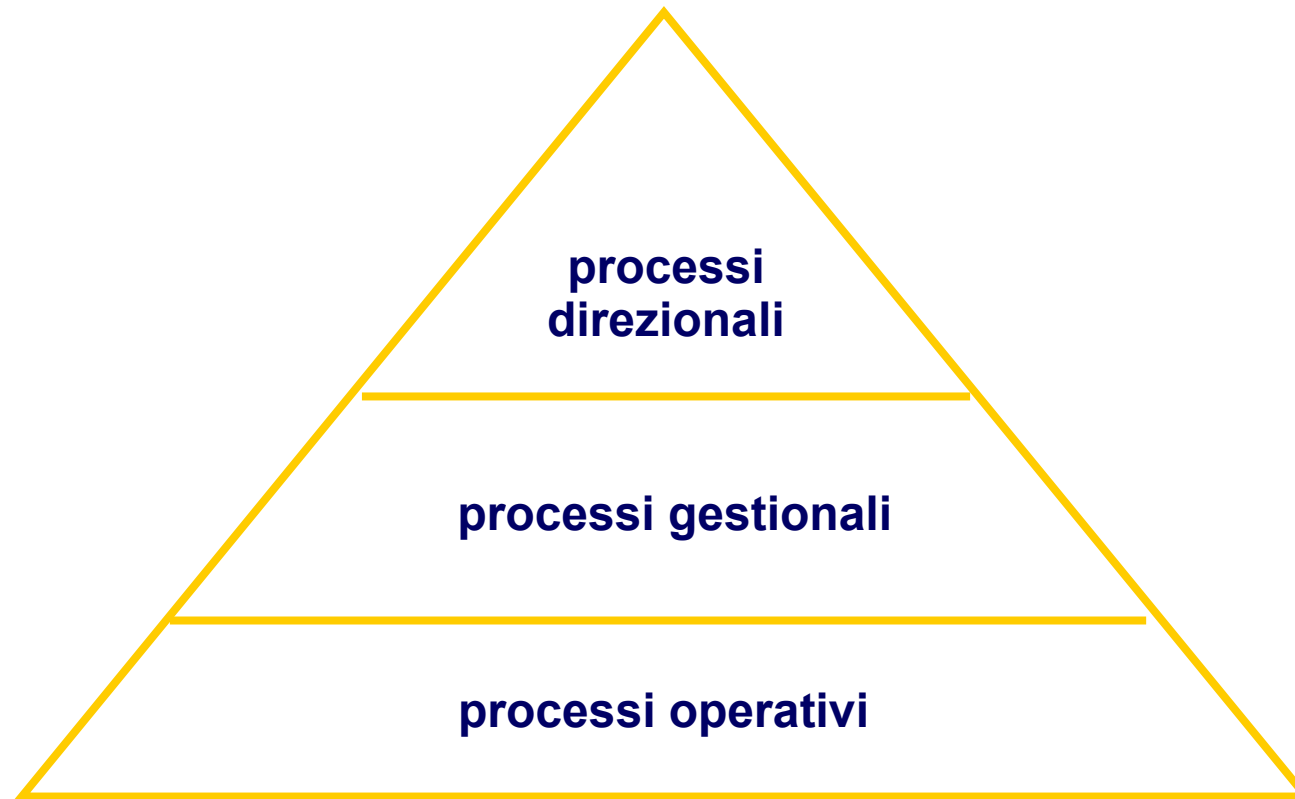
6 giugno 2006

# Sommario

## Introduzione

- Basi di dati integrate, sì, ma ...
- OLTP e OLAP
- Data warehouse e data warehousing
- Dati multidimensionali
- Progettazione di data warehouse
- Studi di caso

# Processi



# Processi presso una banca

- Processi operativi
  - gestione di un movimento su un conto corrente bancario, presso sportello tradizionale o automatico
- Processi gestionali
  - concessione di un fido
  - revisione delle condizioni su un conto corrente
- Processi direzionali
  - verifica dell'andamento dei servizi di carta di credito
  - lancio di una campagna promozionale
  - stipula di accordi commerciali

# Processi presso un'azienda telefonica

- Processi operativi
  - stipula di contratti ordinari
  - instradamento delle telefonate
  - memorizzazione di dati contabili sulle telefonate (chiamante, chiamato, giorno, ora, durata, instradamento,...)
- Processi gestionali
  - stipula di contratti speciali
  - installazione di infrastrutture
- Processi direzionali
  - scelta dei parametri che fissano il costo delle telefonate
  - definizione di contratti diversificati
  - pianificazione del potenziamento delle infrastrutture

# Caratteristiche dei processi dei vari tipi

- Processi operativi
  - su dati dipartimentali e dettagliati
  - operazioni strutturate, basate su regole perfettamente definite
- Processi gestionali
  - su dati settoriali e parzialmente aggregati
  - operazioni semi-strutturate, basate su regole note, ma con un intervento umano con assunzione di responsabilità
- Processi direzionali
  - su dati integrati e fortemente aggregati
  - operazioni non strutturate, senza criteri precisi: capacità personale è essenziale

# Sistemi informatici: una classificazione

- per i processi operativi
  - **Transaction processing systems**
- per i processi gestionali
  - **Management information systems** (di solito settoriali)
- per i processi direzionali  
o meglio, per il supporto ad essi
  - **Decision support systems** (il più possibile integrati)

# Sistemi di supporto alle decisioni

- I sistemi di supporto alle decisioni (DSS) costituiscono la tecnologia che supporta la dirigenza aziendale nel prendere decisioni tattico-strategiche in modo efficace e veloce, mediante particolari tipologie di elaborazione (per esempio OLAP)
- Ma su quali dati?
  - quelli accumulati per i processi operativi e gestionali



# OLTP e OLAP

- OLTP:
  - On-Line Transaction Processing
  - nei sistemi di livello operativo
- OLAP:
  - On-Line Analytical Processing
  - nei sistemi di supporto alle decisioni

# OLTP

- Tradizionale elaborazione di transazioni, che realizzano i processi operativi dell'azienda-ente
  - Operazioni predefinite, brevi e relativamente semplici
  - Ogni operazione coinvolge “pochi” dati
  - Dati di dettaglio, aggiornati
  - Le proprietà “**acide**” (atomicità, correttezza, isolamento, durabilità) delle transazioni sono essenziali

# OLAP

- Elaborazione di operazioni per il supporto alle decisioni
  - Operazioni complesse e casuali
  - Ogni operazione può coinvolgere molti dati
  - Dati aggregati, storici, anche non attualissimi
  - Le proprietà “acide” non sono rilevanti, perché le operazioni sono di sola lettura

## 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

# OLTP e OLAP

- I requisiti sono quindi contrastanti
- Le applicazioni dei due tipi possono danneggiarsi a vicenda

## Evoluzione dei DSS (idea schematica)

- Anni '60 — rapporti batch
  - difficile trovare e analizzare dati
  - ogni richiesta richiede un nuovo programma
- Anni '70 — DSS basato su terminale
  - accesso ai dati operazionali, molto inefficiente
- Anni '80 — strumenti d'automazione d'ufficio e di analisi
  - fogli elettronici, interfacce grafiche
- Anni '90 — data warehousing
  - strumenti di OLAP

# Base di dati

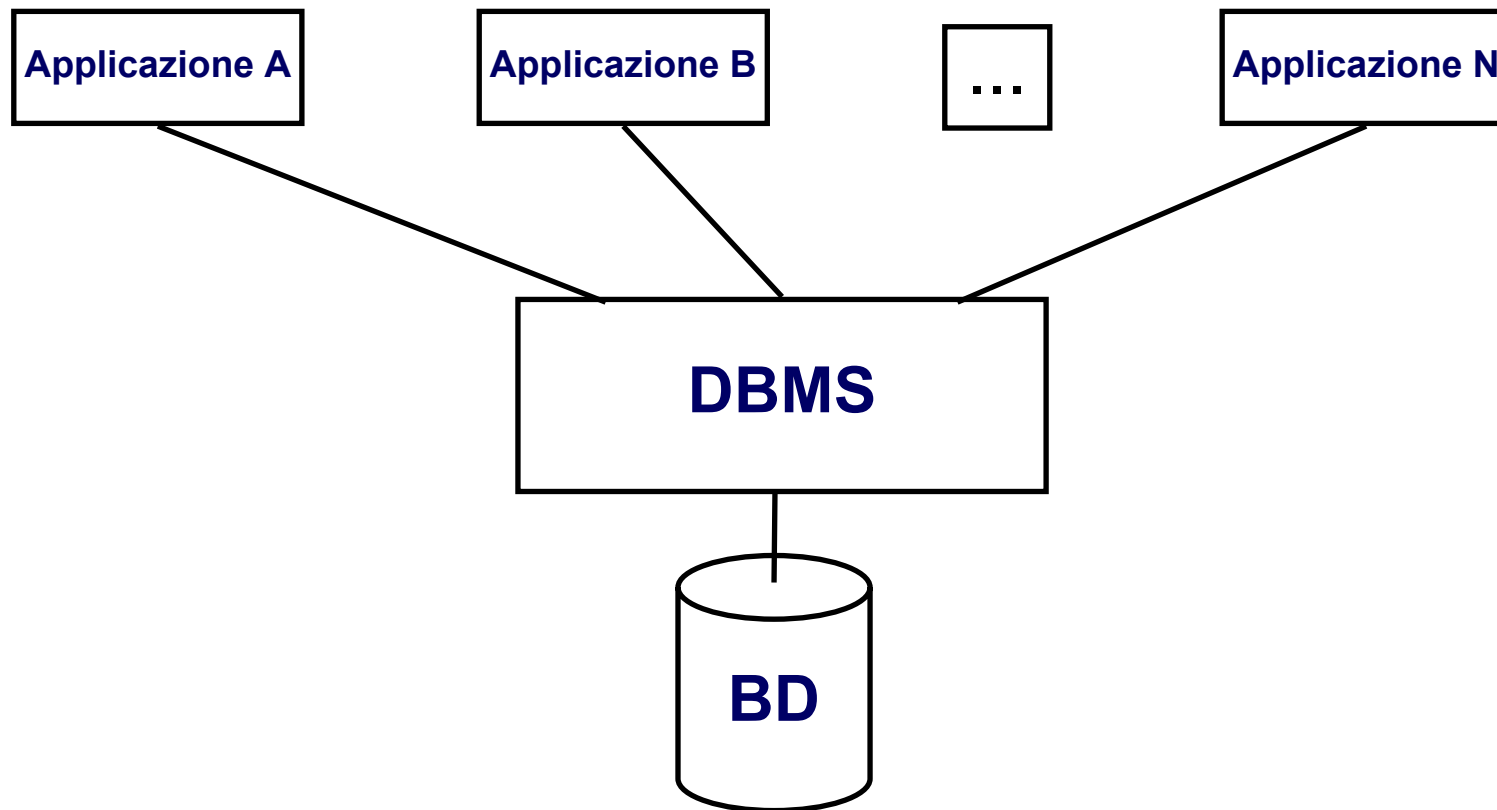
- “Collezione di dati **persistente e condivisa**, gestita in modo **efficace, efficiente e affidabile** (da un **DBMS**)”
- il concetto di base di dati nasce per rispondere alle esigenze di “gestione di una risorsa pregiata”, condivisa da più applicazioni

## Basi di dati: "le magnifiche sorti e progressive"

- “ogni organizzazione ha **una** base di dati, che organizza tutti i dati di interesse in forma integrata e non ridondante”
- “ciascuna applicazione ha accesso a tutti i dati di proprio interesse, in tempo reale e senza duplicazione, riorganizzati secondo le proprie necessità”
- “bla bla bla ...”



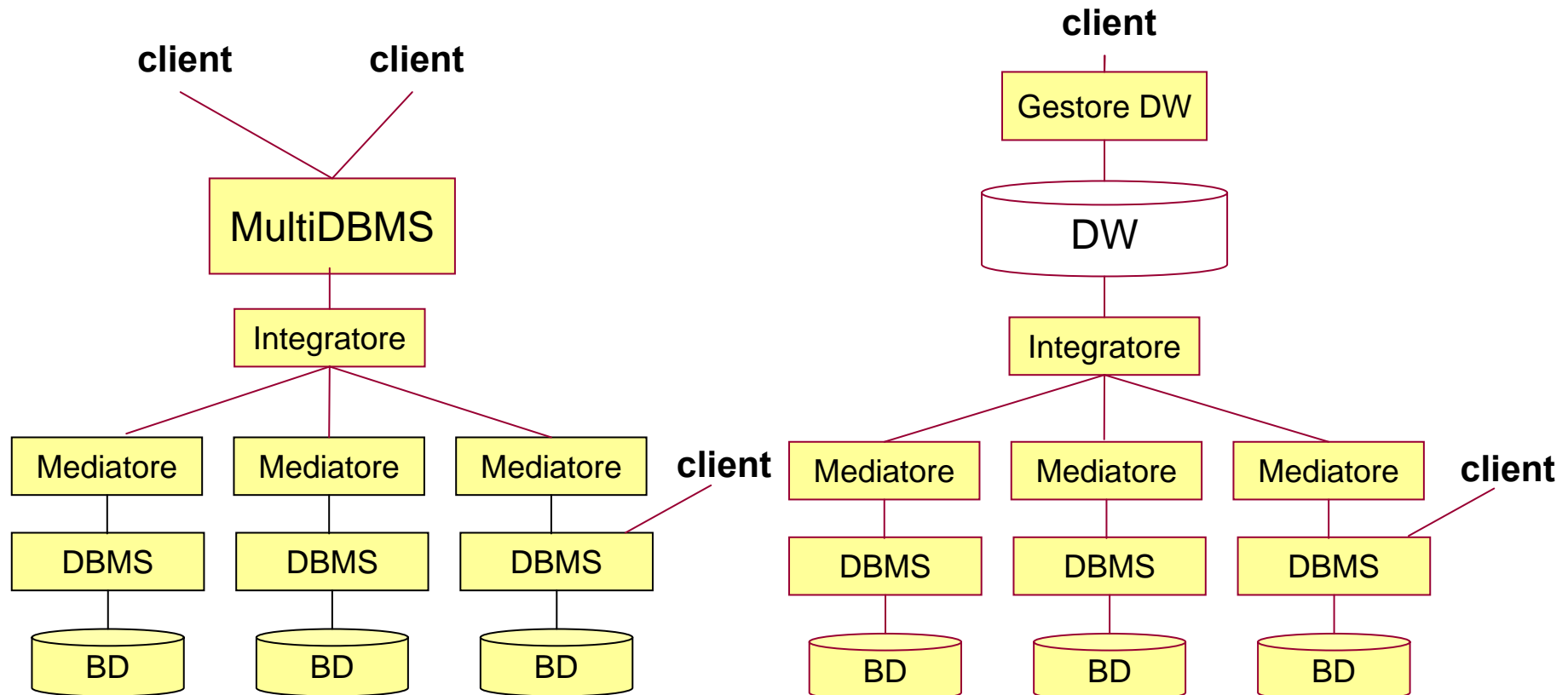
## La base di dati “ideale”



# L'obiettivo ideale è sensato e praticabile?

- La realtà è in continua evoluzione, non esiste uno “stato stazionario” (se non nell’iperuranio):
  - cambiano le esigenze
  - cambiano le strutture
  - le realizzazioni richiedono tempo
- Il coordinamento forte fra i vari settori può risultare controproducente
- Ogni organizzazione ha di solito diverse basi di dati **distribuite, eterogenee, autonome**

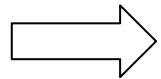
# Multi-database e Data Warehouse (due approcci all'integrazione)



# Sommario

## Introduzione

- Basi di dati integrate, sì, ma ...
- OLTP e OLAP



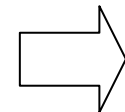
## Data warehouse e data warehousing

- Dati multidimensionali
- Progettazione di data warehouse
- Studi di caso

# Data warehouse

Una base di dati

- utilizzata principalmente per il supporto alle decisioni direzionali (OLAP e non OLTP)
- integrata — aziendale e non dipartimentale
- orientata ai dati — non alle applicazioni
- con dati storici — con un ampio orizzonte temporale, e indicazione (di solito) di elementi di tempo
- con dati aggregati (di solito) — per effettuare stime e valutazioni
- fuori linea — i dati sono aggiornati periodicamente
- separata dalle basi di dati operazionali

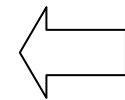


## OLTP e OLAP

	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

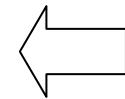
## ... integrata ...

- I dati di interesse provengono da tutte le sorgenti informative — ciascun dato proviene da una o più di esse
- Il data warehouse rappresenta i dati in modo univoco — riconciliando le eterogeneità dalle diverse rappresentazioni
  - nomi
  - struttura
  - codifica
  - rappresentazione multipla



## ... orientata ai dati ...

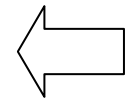
- Le basi di dati operazionali sono costruite a supporto dei singoli processi operativi o applicazioni
  - produzione
  - vendita
- Il data warehouse è costruito attorno alle principali entità del patrimonio informativo aziendale
  - prodotto
  - cliente





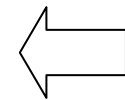
## ... dati storici ...

- Le basi di dati operazionali mantengono il valore corrente delle informazioni
  - L'orizzonte temporale di interesse è dell'ordine dei pochi mesi
- Nel data warehouse è di interesse l'evoluzione storica delle informazioni
  - L'orizzonte temporale di interesse è dell'ordine degli anni



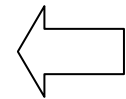
## ... dati aggregati ...

- Nelle attività di analisi dei dati per il supporto alle decisioni
  - non interessa “chi” ma “quanti”
  - non interessa un dato ma
    - la somma,
    - la media,
    - il minimo e il massimo, ...di un insieme di dati.
- Le operazioni di aggregazione sono quindi fondamentali nel warehousing e nella costruzione/mantenimento di un data warehouse.



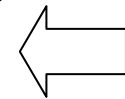
## ... fuori linea ...

- In una base di dati operativa, i dati vengono
  - acceduti
  - inseriti
  - modificati
  - cancellatipochi record alla volta
- Nel data warehouse, abbiamo
  - operazioni di accesso e interrogazione — “diurne”
  - operazioni di caricamento e aggiornamento dei dati — “notturne”che riguardano milioni di record



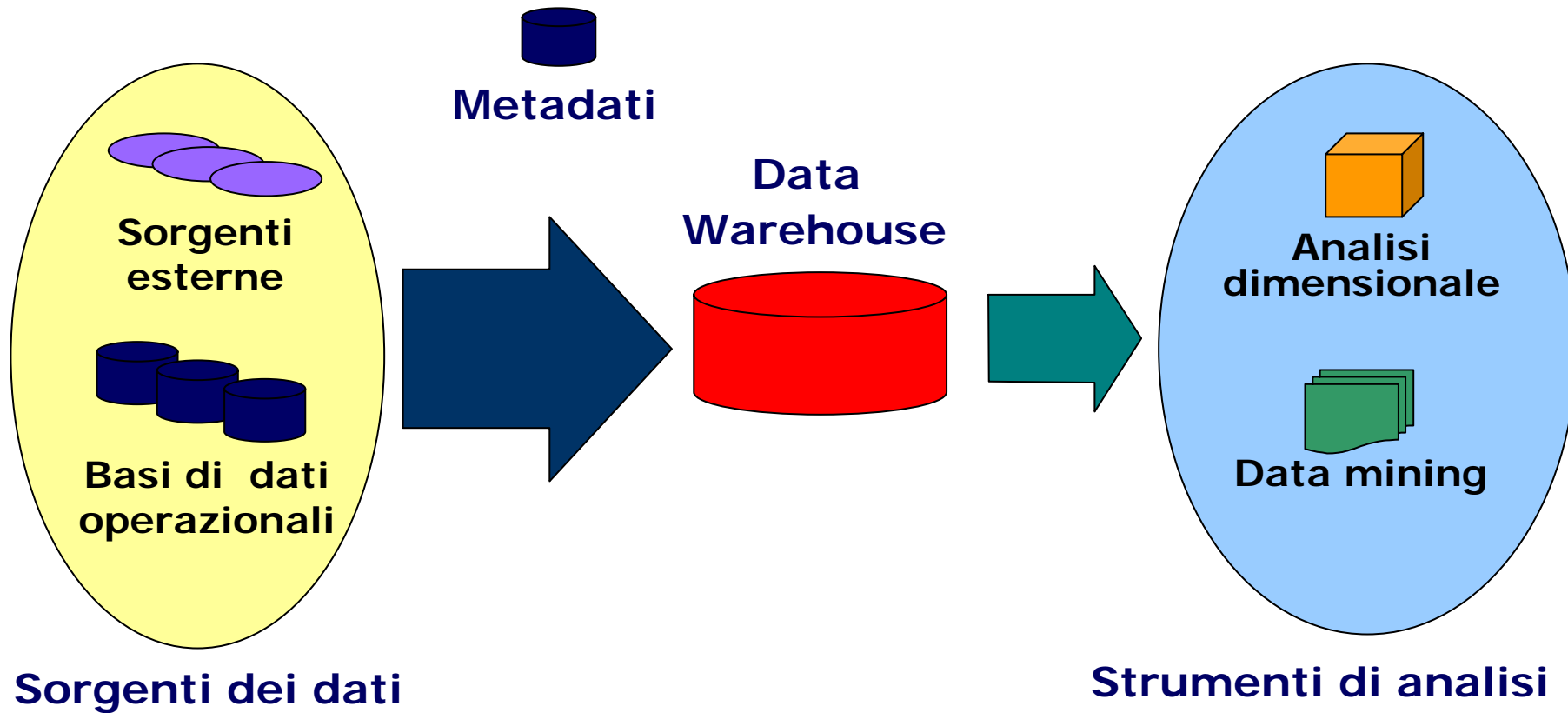
## ... una base di dati separata ...

- Un data warehouse viene mantenuto separatamente dalle basi di dati operazionali perché
  - non esiste un'unica base di dati operazionale che contiene tutti i dati di interesse
  - la base di dati deve essere integrata
  - non è tecnicamente possibile fare l'integrazione in linea
  - i dati di interesse sarebbero comunque diversi
    - devono essere mantenuti dati storici
    - devono essere mantenuti dati aggregati
  - l'analisi dei dati richiede per i dati organizzazioni speciali e metodi di accesso specifici
  - degrado generale delle prestazioni senza la separazione



# Architettura per il data warehousing

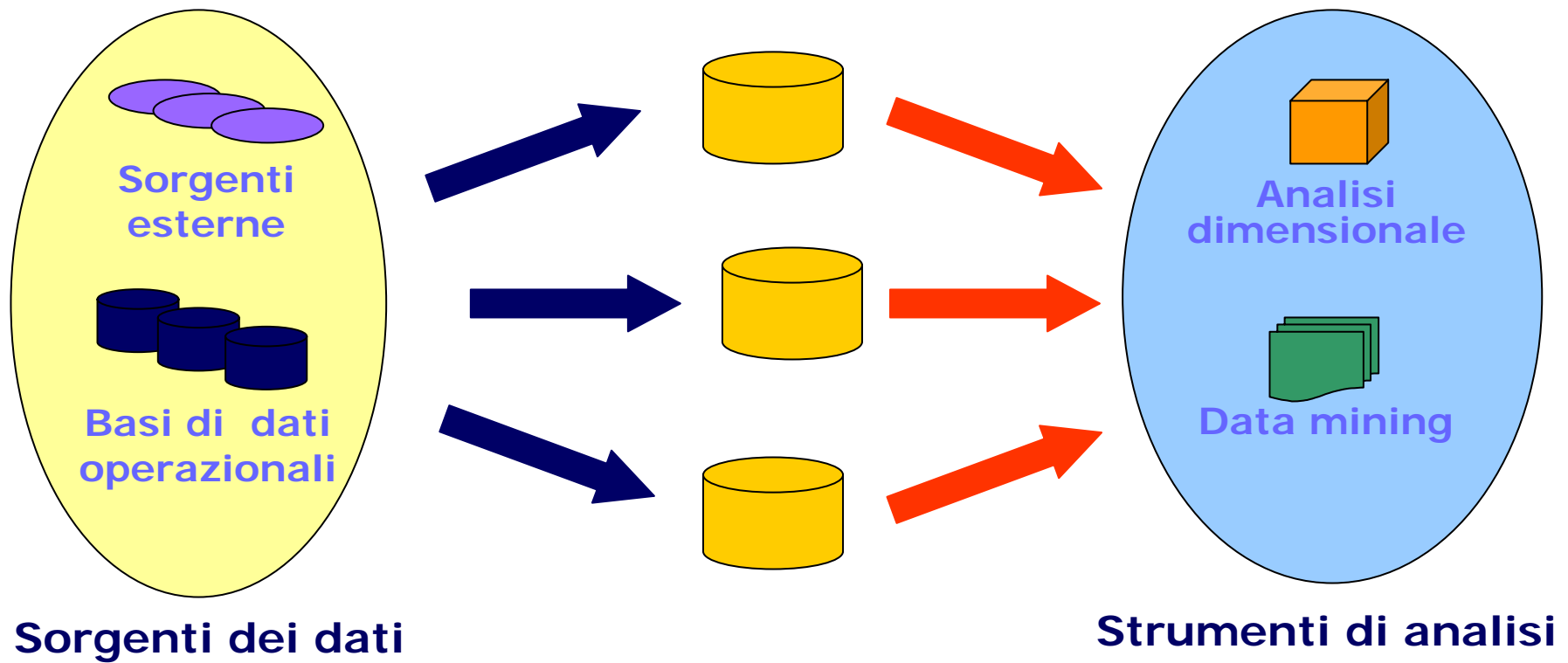
## Monitoraggio & Amministrazione



# Esigenze di analisi e integrazione

- Molto spesso:
  - l'analisi è mirata a specifici processi della azienda o ente
  - un vero e proprio DW integrato
    - non interessa
    - non “viene in mente”
    - non si riesce a fare (per urgenza, mancanza di risorse, o mancanza di “competenza e responsabilità”)
  - può essere utile o necessario concentrarsi (almeno temporaneamente) su un suo sottoinsieme

# Architettura “realistica”

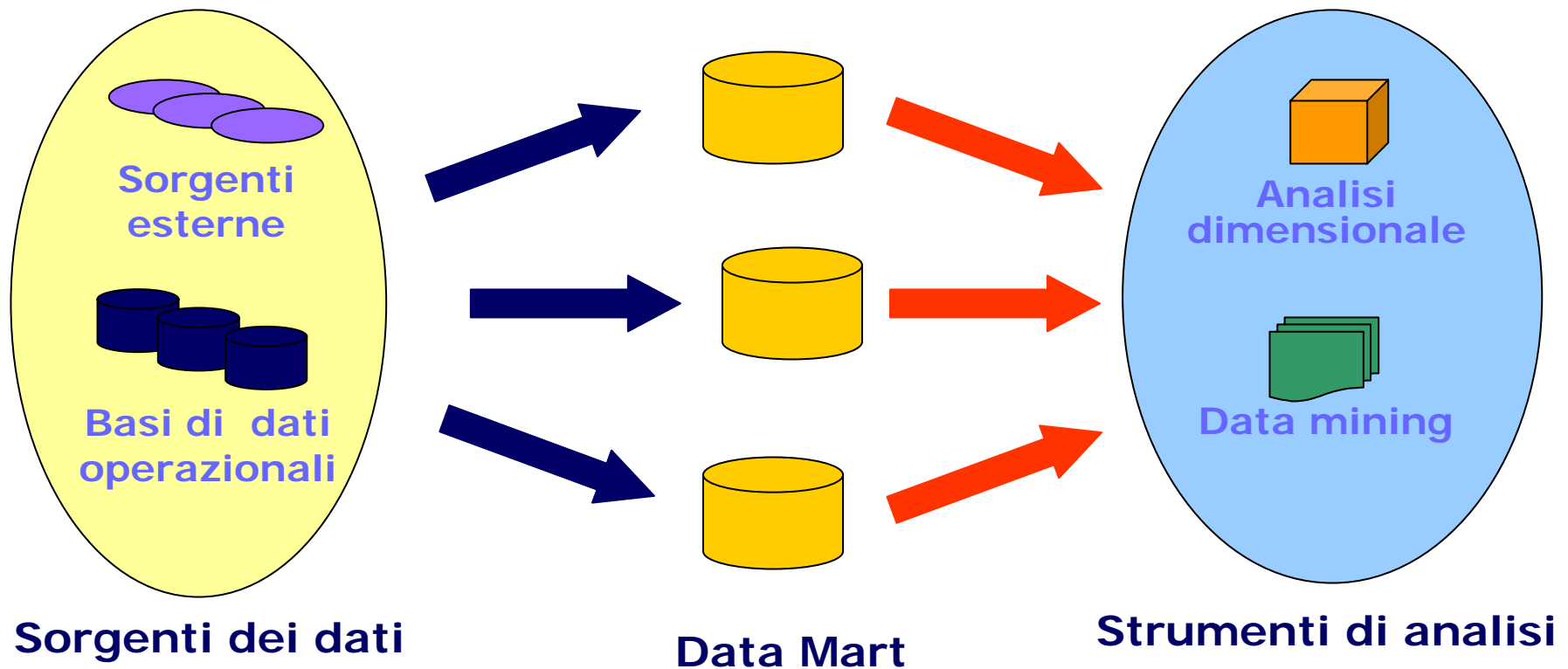


# Data mart

- Un sottoinsieme logico dell'intero data warehouse
  - un data mart è la restrizione del data warehouse a un singolo processo
  - un data warehouse è l'unione di tutti i suoi data mart



# Variante dell'architettura



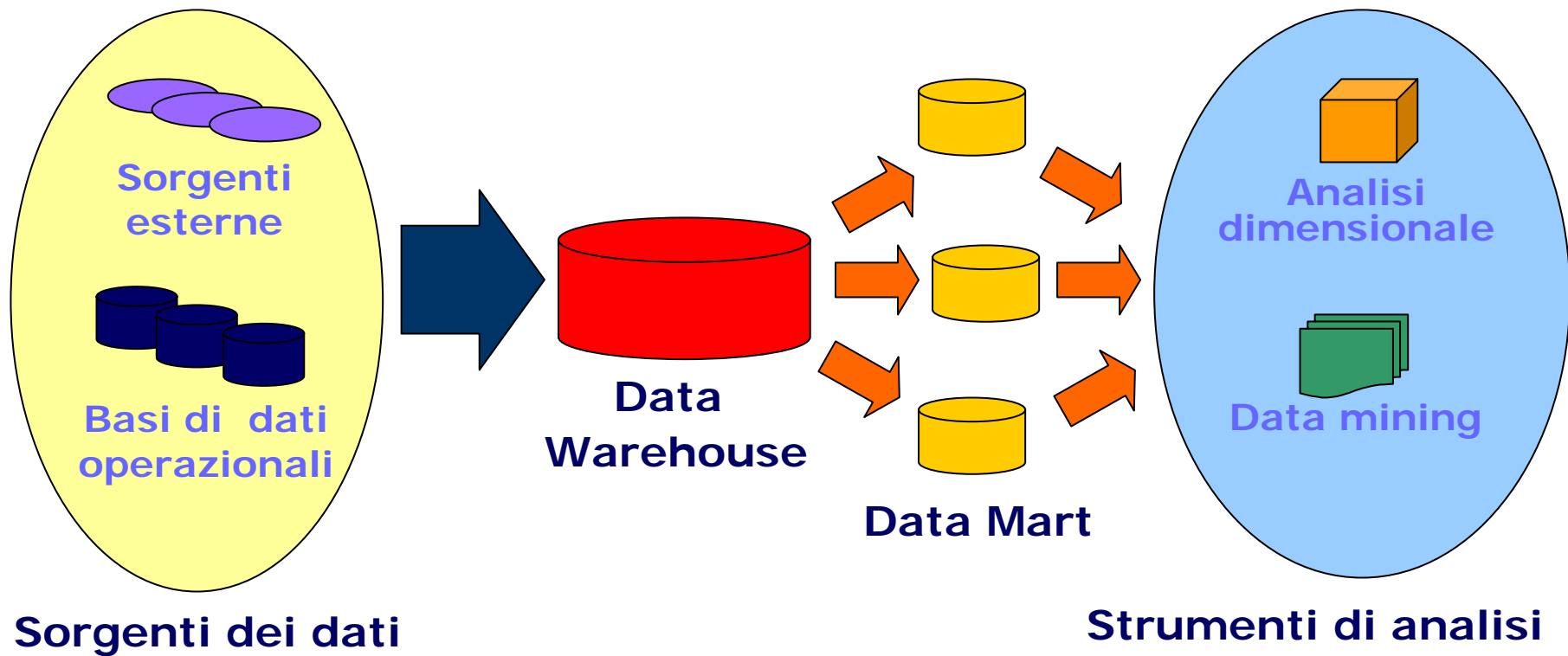
# Data mart e DW

- Pro e contro dei data mart
  - un data mart rappresenta un progetto solitamente fattibile
    - la realizzazione diretta di un data warehouse completo non è invece solitamente fattibile
  - tuttavia, la realizzazione di un insieme di data mart non porta necessariamente alla realizzazione del data warehouse

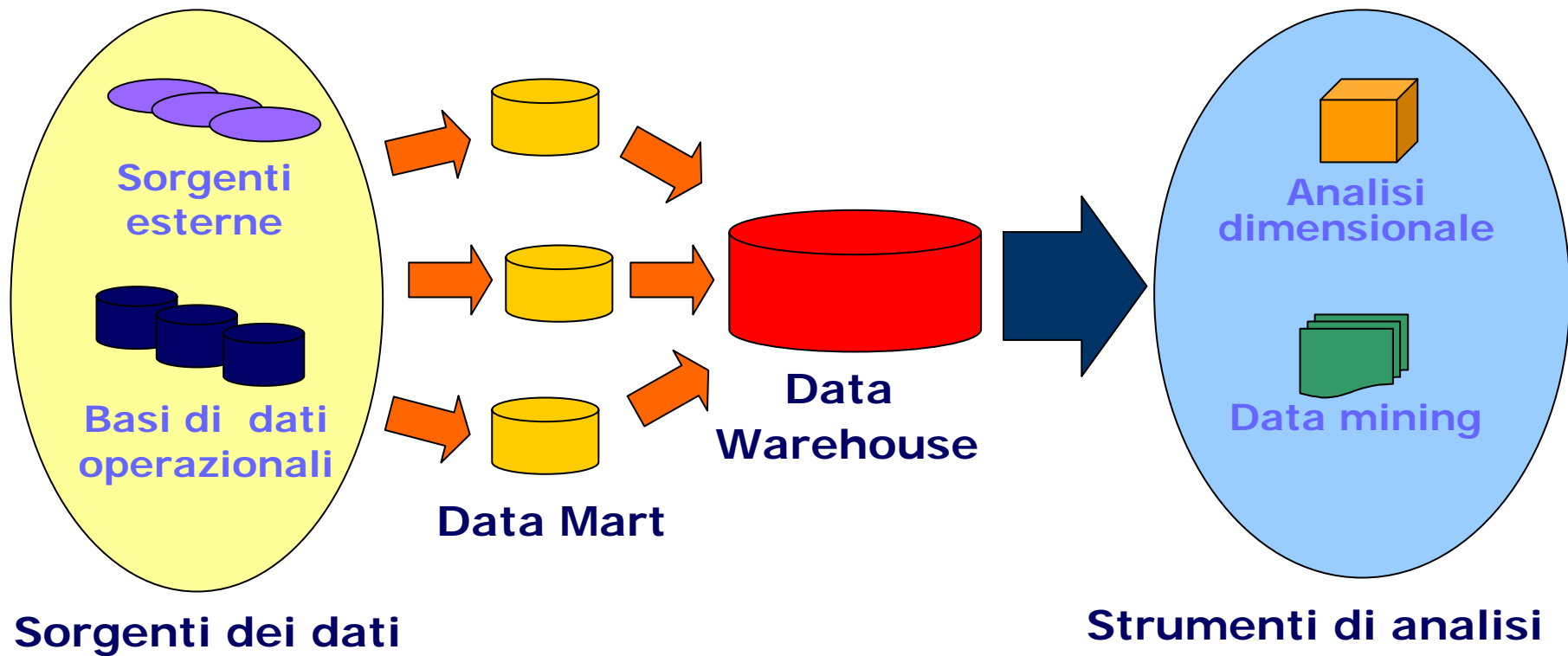
# Top-down o bottom-up?

- Prima il data warehouse o prima i data mart?

# DW e DM



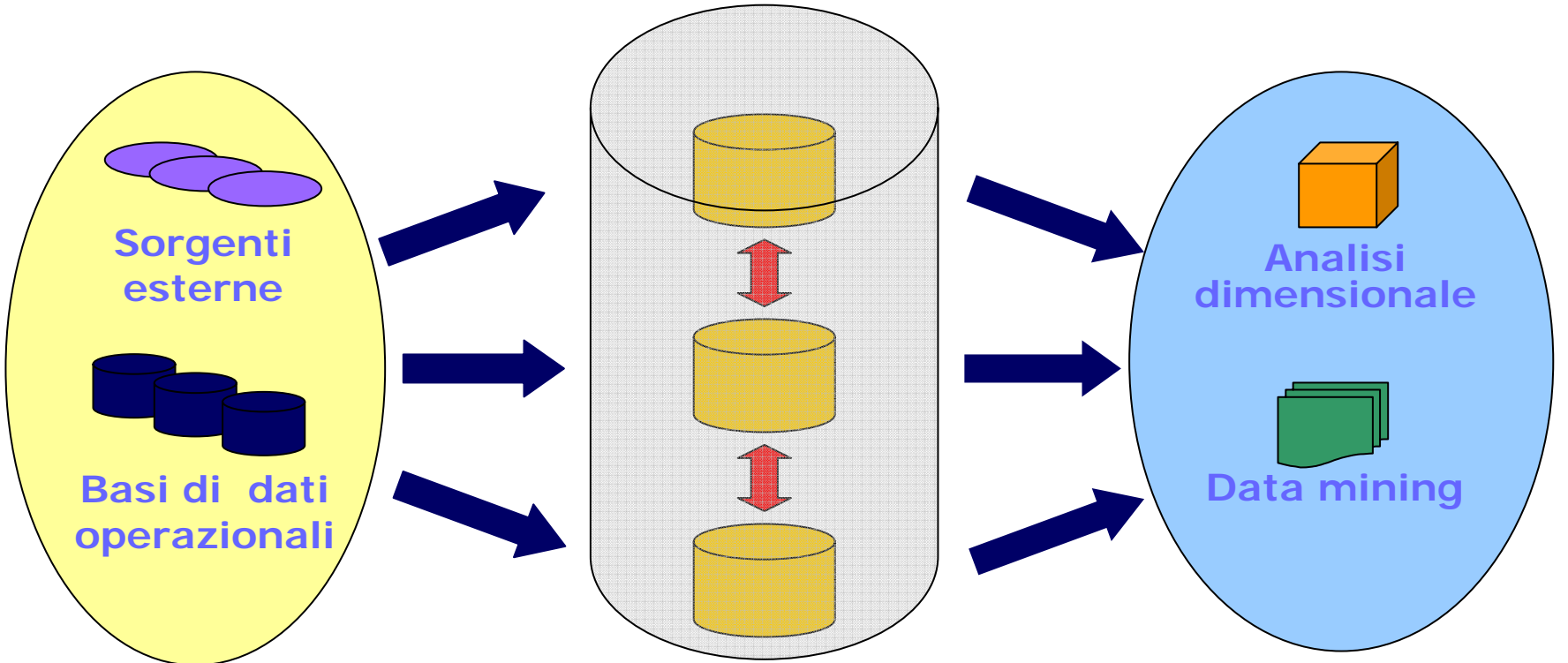
# DW e DM



## Top-down o bottom-up?

- Prima il data warehouse o prima i data mart?
- Non c'è risposta, o meglio: nessuno dei due!
- Infatti:
  - l'approccio è spesso incrementale
- Ma
  - è necessario coordinare i data mart:
    - dimensioni conformi e “DW bus”

# DM e DW



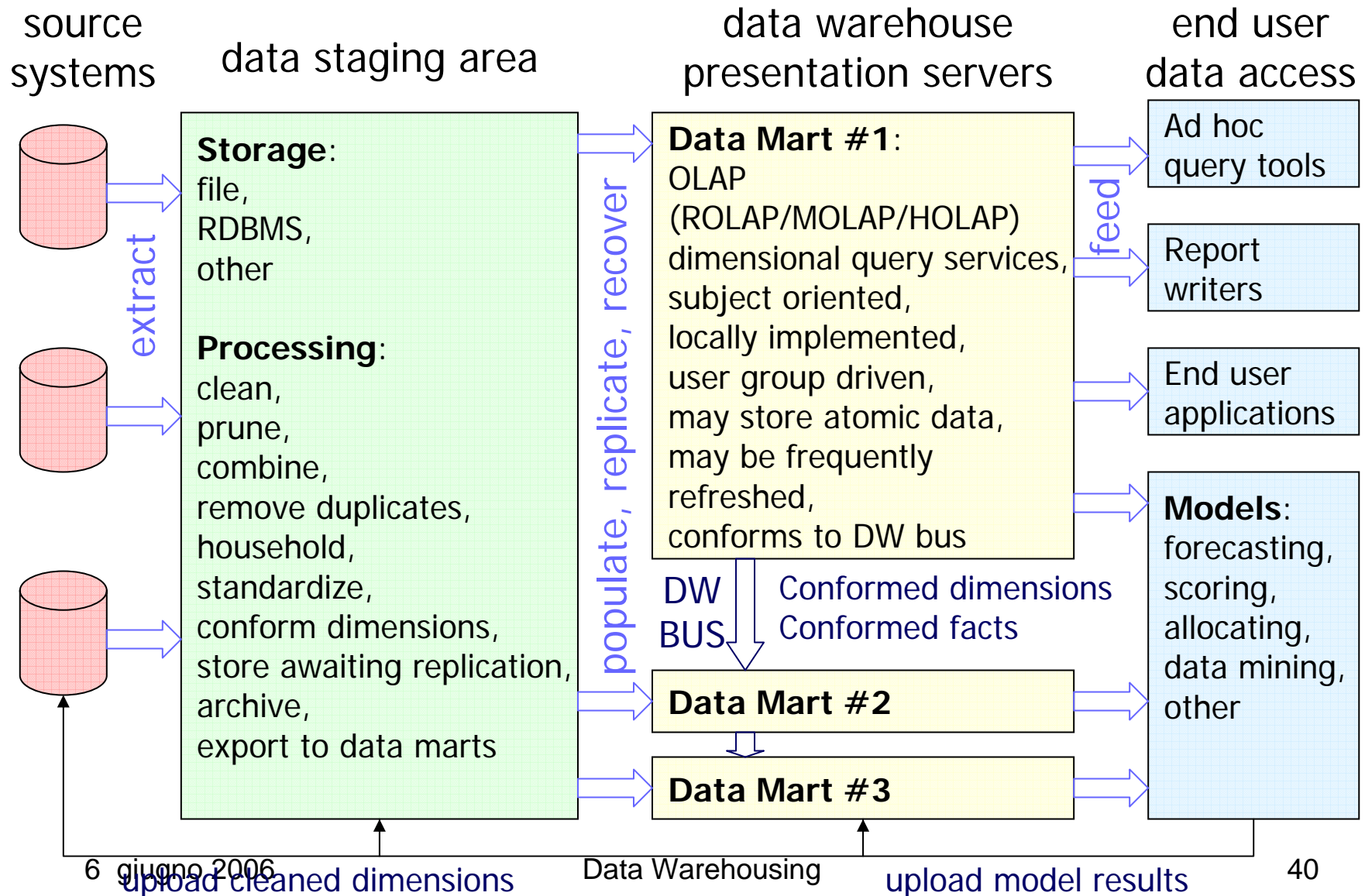
**Sorgenti dei dati**

**Data Mart**

**Strumenti di analisi**

**DW**  
**con "bus"**  
Data Warehousing

# Elementi di un data warehouse





# Sorgenti informative

- i sistemi operazionali dell'organizzazione
  - sono sistemi transazionali (OLTP) orientati alla gestione dei processi operazionali
  - non mantengono dati storici
  - ogni sistema gestisce uno o più soggetti (ad esempio, prodotti o clienti)
    - nell'ambito di un processo
    - ma non in modo conforme nell'ambito dell'organizzazione
  - sono sistemi “legacy”
- sorgenti esterne
  - ad esempio, dati forniti da società specializzate di analisi

# Area di preparazione dei dati

- L'**area di preparazione** dei dati (**data staging**) è usata per il transito dei dati dalle sorgenti informative al data warehouse
  - comprende ogni cosa tra le sorgenti informative e i server di presentazione
    - aree di memorizzazione dei dati estratti dalle sorgenti informative e preparati per il caricamento nel data warehouse
    - processi per la preparazione di tali dati
      - pulizia, trasformazione, combinazione, rimozione di duplicati, archiviazione, preparazione per l'uso nel data warehouse
  - richiede un insieme complesso di attività semplici
  - è distribuita su più calcolatori e ambienti eterogenei
  - gestisce i dati prevalentemente con formati di varia natura (spesso semplici file)

# Server di presentazione

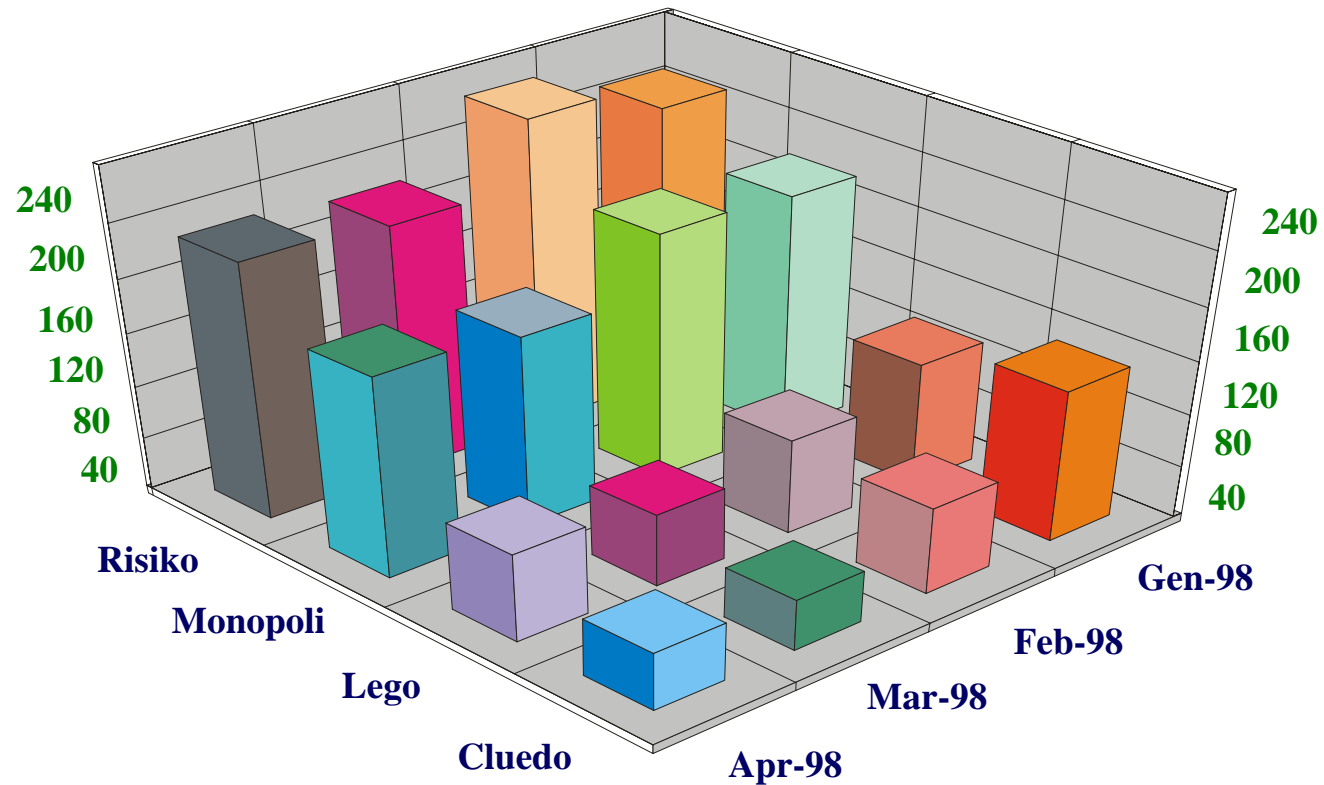
- Un **server di presentazione** è un sistema in cui i dati del data warehouse sono organizzati e memorizzati per essere interrogati direttamente da utenti finali, report writer e altre applicazioni
  - i dati sono rappresentati in forma **multidimensionale** (secondo i concetti di fatto e dimensione, vediamo fra poco)
  - tecnologie che possono essere adottate
    - RDBMS: ROLAP
    - tecnologia OLAP esplicita: MOLAP
      - i concetti di fatto e dimensione sono espliciti

# Visualizzazione dei dati

- I dati vengono infine visualizzati in veste grafica, in maniera da essere facilmente comprensibili.
- Si fa uso di:
  - tabelle
  - istogrammi
  - grafici
  - torte
  - superfici 3D
  - bolle
  - area in pila
  - forme varie
  - ...

# Visualizzazione finale di un'analisi

## *Vendite mensili giocattoli a Roma*

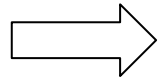


# Sommario

## Introduzione

- Basi di dati integrate, sì, ma ...
- OLTP e OLAP

- Data warehouse e data warehousing



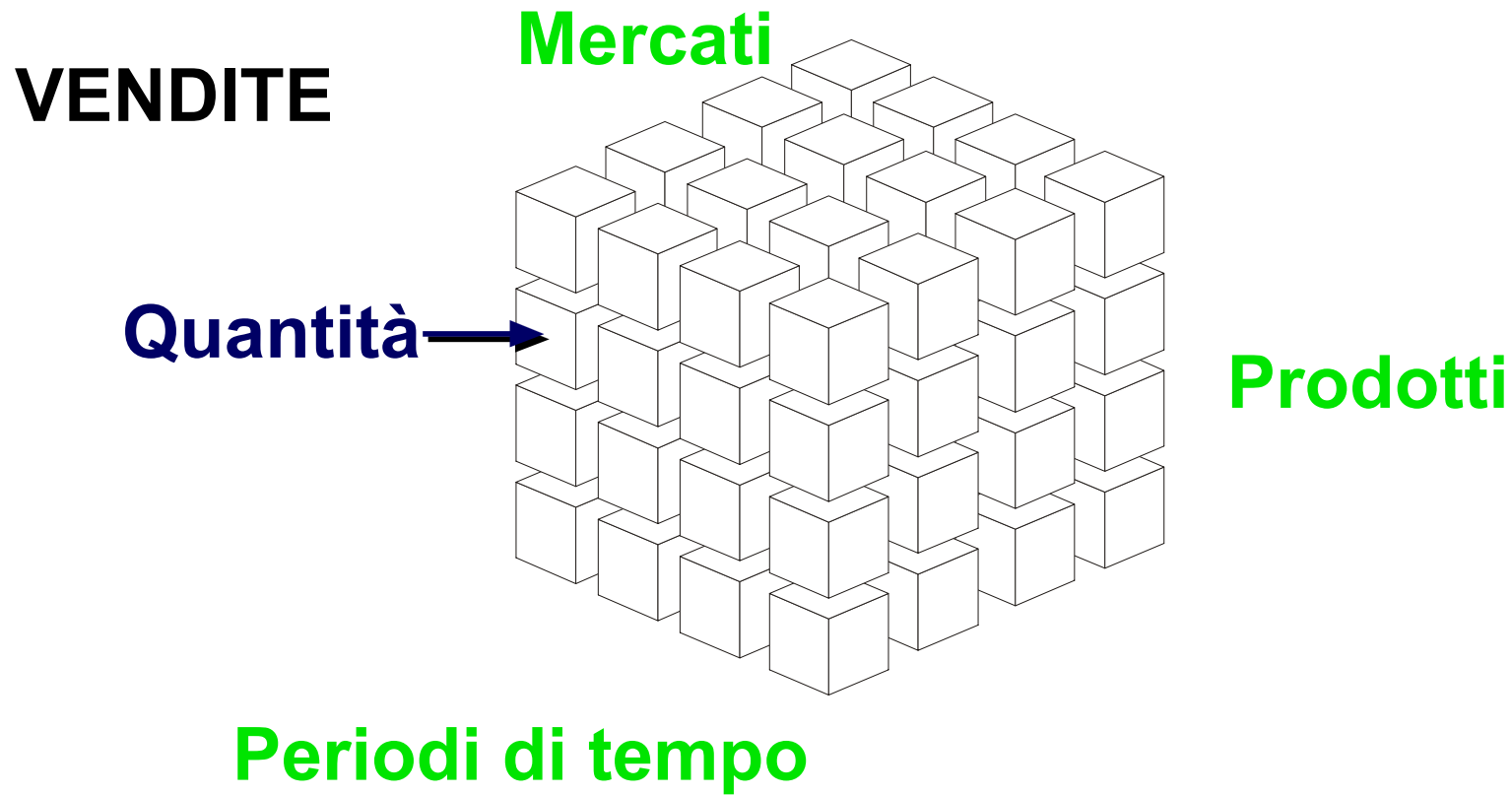
## Dati multidimensionali

- Progettazione di data warehouse
- Studi di caso

# Modello “logico” per DW

- L’analisi dei dati avviene rappresentando i dati in forma **multidimensionale**
- Concetti rilevanti:
  - **fatto** — un concetto sul quale centrare l’analisi
  - **misura** — una proprietà atomica di un fatto da analizzare
  - **dimensione** — descrive una prospettiva lungo la quale effettuare l’analisi
- Esempi di fatti/misure/dimensioni
  - vendita / quantità venduta, incasso / prodotto, tempo
  - telefonata / costo, durata / chiamante, chiamato, tempo

# Rappresentazione multidimensionale dei dati

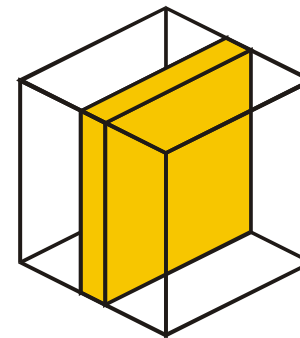
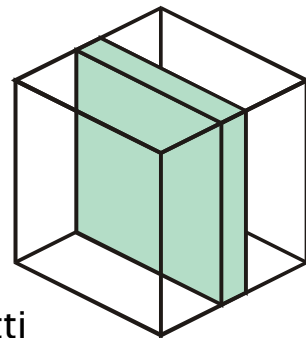
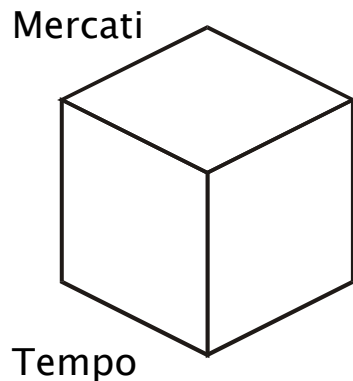




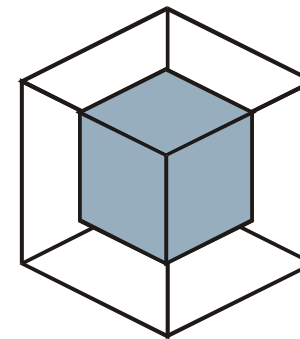
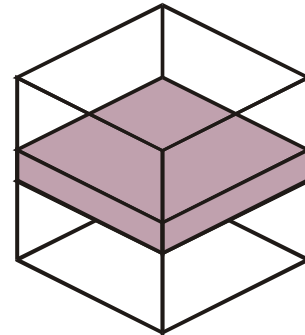
# Viste su dati multidimensionali

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



Prodotti



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 e un orizzonte temporale

# Operazioni su dati multidimensionali

- **Roll up** (o drill up)— aggrega i dati
  - volume di vendita totale dello scorso anno per categoria di prodotto e regione
- **Drill down** — disaggrega i dati
  - per una particolare categoria di prodotto e regione, mostra le vendite giornaliere dettagliate per ciascun negozio
- **Slice & dice** — seleziona e proietta
- (**Pivot** — re-orienta il cubo)

	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze 1	21	4	10	4	6	7
Firenze 2	4	4	4	6	6	3
Roma 1	15	5	8	3	5	20
Roma 2	12	4	7	5	2	4
Roma 3	23	4	9	10	5	5
Latina	3	3	5	1	2	4

	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze 1	21	4	10	4	6	7
Firenze 2	4	4	4	6	6	3
Roma 1	15	5	8	3	5	20
Roma 2	12	4	7	5	2	4
Roma 3	23	4	9	10	5	5
Latina	3	3	5	1	2	4

	Gen	Feb	Mar	Apr	Mag	Giu
	90	26	53	32	32	48

	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze 1	21	4	10	4	6	7
Firenze 2	4	4	4	6	6	3
Roma 1	15	5	8	3	5	20
Roma 2	12	4	7	5	2	4
Roma 3	23	4	9	10	5	5
Latina	3	3	5	1	2	4

Pisa	38
Firenze 1	52
Firenze 2	27
Roma 1	56
Roma 2	34
Roma 3	56
Latina	18

## 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

regione  
↑  
provincia  
↑  
città  
↑  
negozio

categoria    marca  
↑            ↑  
prodotto

anno  
↑  
trimestre  
↑  
mese  
↑  
giorno

	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze 1	21	4	10	4	6	7
Firenze 2	4	4	4	6	6	3
Roma 1	15	5	8	3	5	20
Roma 2	12	4	7	5	2	4
Roma 3	23	4	9	10	5	5
Latina	3	3	5	1	2	4

	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze	25	8	14	10	12	10
Roma	50	13	24	18	12	29
Latina	3	3	5	1	2	4

	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze 1	21	4	10	4	6	7
Firenze 2	4	4	4	6	6	3
Roma 1	15	5	8	3	5	20
Roma 2	12	4	7	5	2	4
Roma 3	23	4	9	10	5	5
Latina	3	3	5	1	2	4

	Gen	Feb	Mar	Apr	Mag	Giu
Toscana	37	10	24	13	18	15
Lazio	53	16	29	19	14	33



	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze 1	21	4	10	4	6	7
Firenze 2	4	4	4	6	6	3
Roma 1	15	5	8	3	5	20
Roma 2	12	4	7	5	2	4
Roma 3	23	4	9	10	5	5
Latina	3	3	5	1	2	4

	I trim	II trim
Pisa	24	14
Firenze 1	35	17
Firenze 2	12	15
Roma 1	28	28
Roma 2	23	11
Roma 3	36	20
Latina	11	7

	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze 1	21	4	10	4	6	7
Firenze 2	4	4	4	6	6	3
Roma 1	15	5	8	3	5	20
Roma 2	12	4	7	5	2	4
Roma 3	23	4	9	10	5	5
Latina	3	3	5	1	2	4

	I trim	II trim
Pisa	24	14
Firenze 1	35	17
Firenze 2	12	15
Roma 1	28	28
Roma 2	23	11
Roma 3	36	20
Latina	11	7

	Gen	Feb	Mar	Apr	Mag	Giu
Pisa	12	2	10	3	6	5
Firenze	25	8	14	10	12	10
Roma	50	13	24	18	12	29
Latina	3	3	5	1	2	4

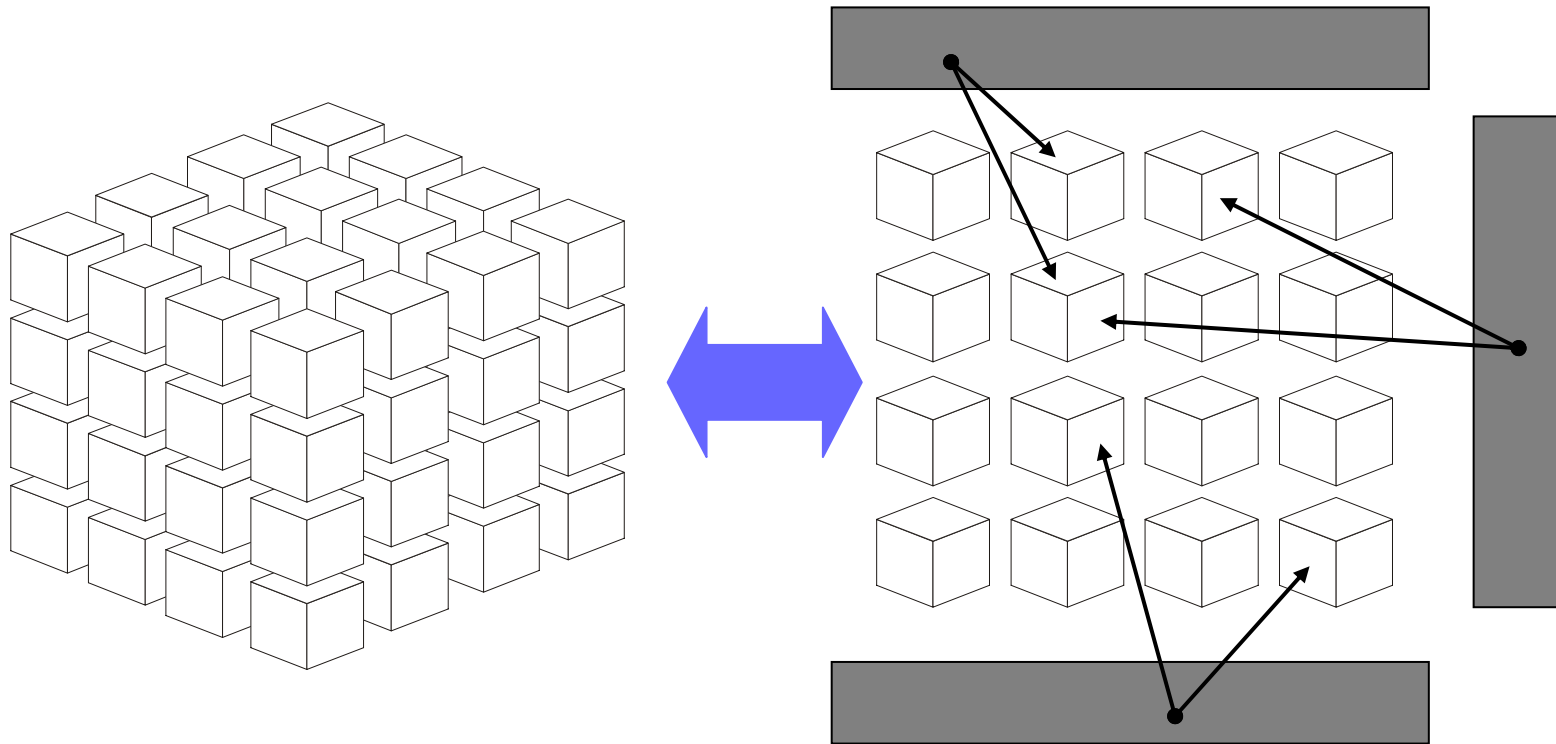
	I trim	II trim
Pisa	24	14
Firenze	47	32
Roma	87	59
Latina	11	7

# Implementazione per dati multidimensionali

- MOLAP
  - M = multidimensional
- ROLAP
  - R = relational

# Implementazione MOLAP

- I dati sono memorizzati direttamente in un formato dimensionale (proprietario). Le gerarchie sui livelli sono codificate in indici di accesso alle matrici



# Implementazione ROLAP: schemi dimensionali

- Uno **schema dimensionale** (**schema a stella**) è composto da
  - una tabella principale, chiamata **tabella fatti**
    - la tabella fatti memorizza le misure di un processo
      - i fatti più comuni hanno misure numeriche e additive
  - due o più tabelle ausiliarie, chiamate **tabelle dimensione**
    - una tabella dimensione rappresenta una prospettiva, un aspetto rispetto a cui è interessante analizzare i fatti
      - gli attributi sono solitamente testuali, discreti e descrittivi

## Schema dimensionale: dimensioni semplici

CodN	Nome
PI	Pisa
FI1	Firenze 1
FI2	Firenze 2
RM1	Roma 1
RM2	Roma 2
RM3	Roma 3
LT	Latina

Negozi	Mese	Vendite
PI	Gen	12
PI	Feb	2
PI	Mar	10
PI	Apr	3
PI	Mag	6
PI	Giu	5
FI1	Gen	21
FI1	Feb	4
FI1	Mar	10
FI1	Apr	4
FI1	Mag	6
FI1	Giu	7
...	...	...

CodM	Mese
Gen	gennaio
Feb	febbraio
Mar	marzo
Apr	aprile
Mag	maggio
Giu	giugno

## Schema dimensionale: dimensioni con livelli

CodN	...	Città	Regione	...
PI	...	Pisa	Toscana	...
FI1	...	Firenze	Toscana	...
FI2	...	Firenze	Toscana	...
RM1	...	Roma	Lazio	...
RM2	...	Roma	Lazio	...
RM3	...	Roma	Lazio	...
LT	...	Latina	Lazio	...

Negozi	Mese	Vendite
PI	Gen	12
PI	Feb	2
PI	Mar	10
PI	Apr	3
PI	Mag	6
PI	Giu	5
FI1	Gen	21
FI1	Feb	4
FI1	Mar	10
FI1	Apr	4
FI1	Mag	6
FI1	Giu	7
...	...	...

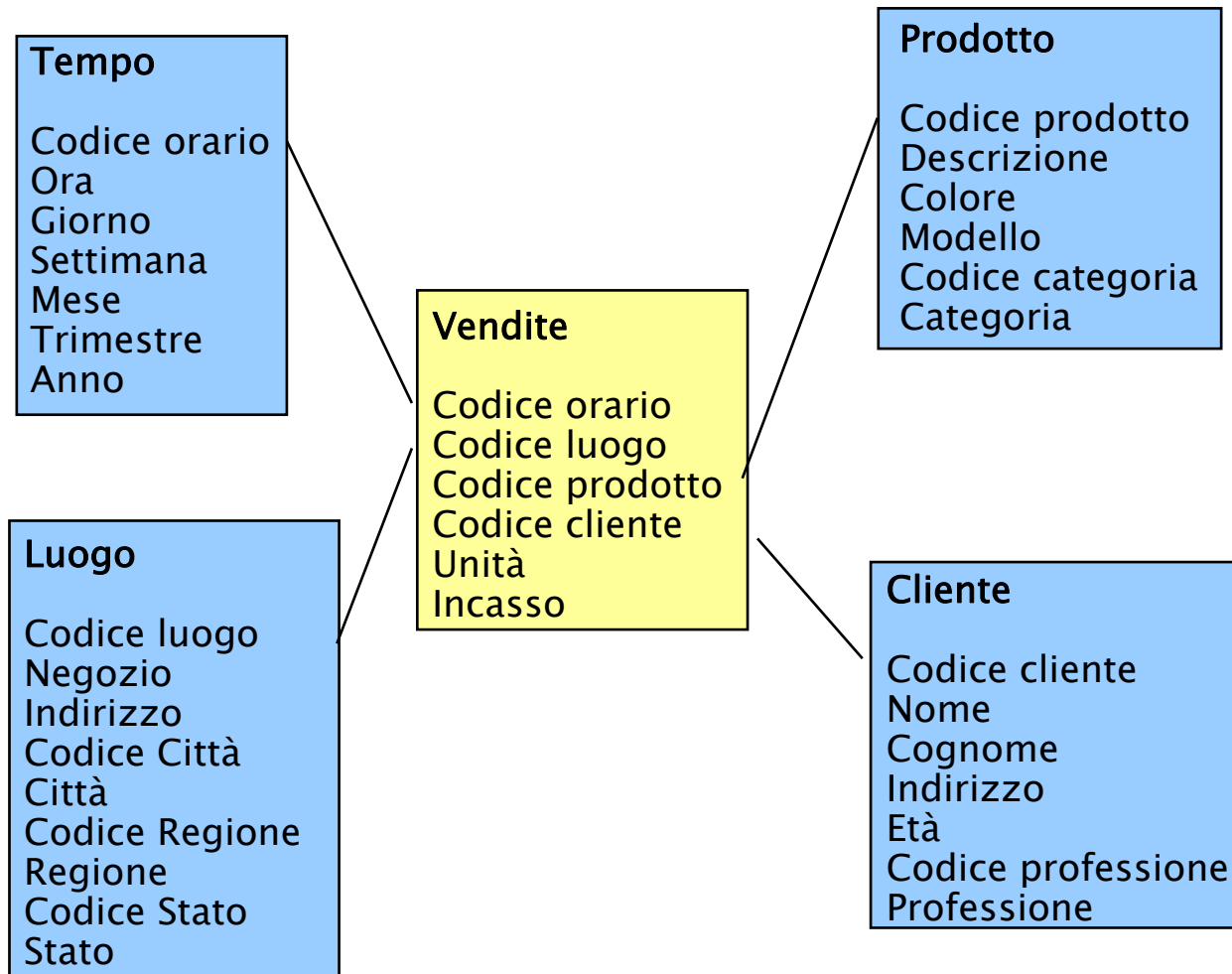
CodM	Mese	Trimestre
Gen	gennaio	I trim
Feb	febbraio	I trim
Mar	marzo	I trim
Apr	aprile	II trim
Mag	maggio	II trim
Giu	giugno	II trim

# Data warehouse dimensionale

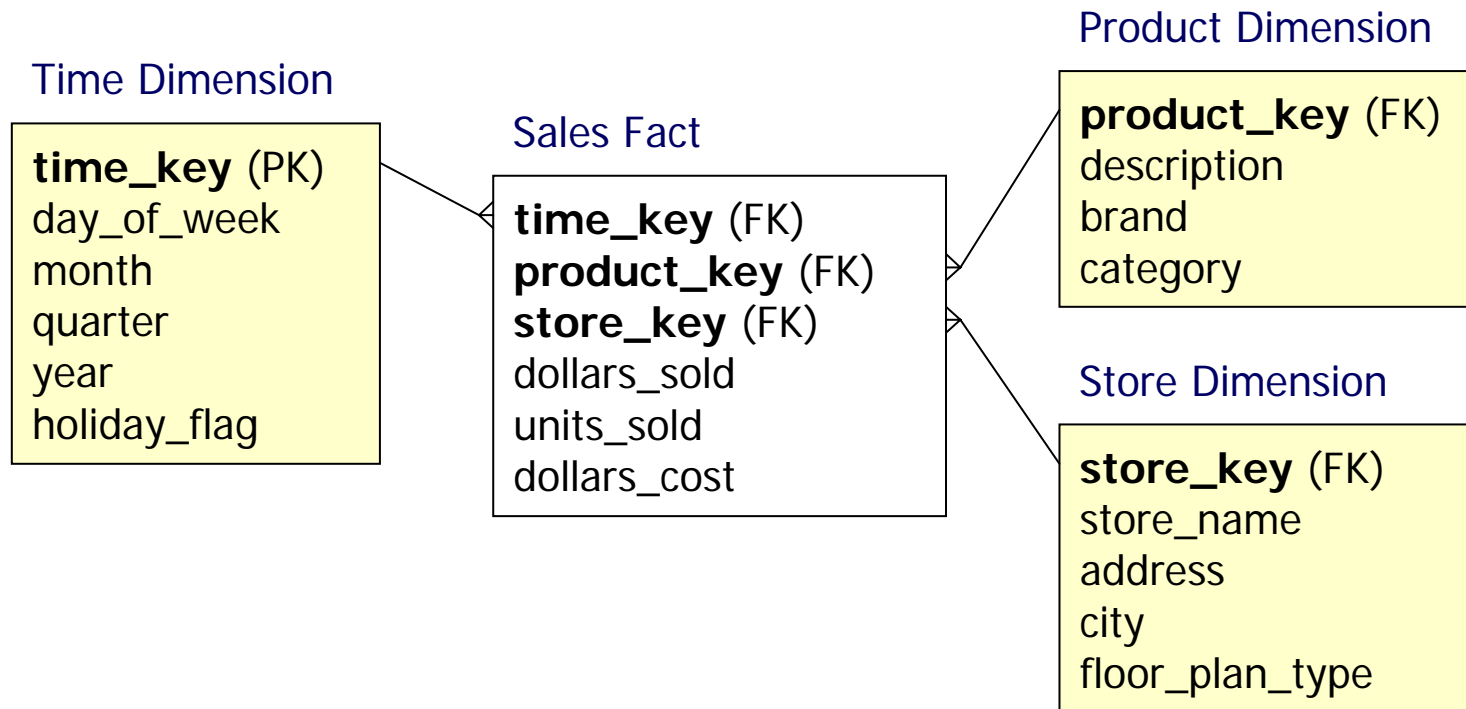
- lo schema di un data warehouse è un insieme di schemi dimensionali
  - ogni data mart è un insieme di schemi dimensionali
  - tutti i data mart vengono costruiti usando il “DW bus”
    - dimensioni conformi
      - ogni dimensione ha lo stesso significato in ciascuno schema dimensionale e data mart
    - fatti conformi
      - anche i fatti hanno interpretazione uniforme



# Uno schema dimensionale



# Un altro schema dimensionale



- i dati delle vendite di prodotti in un certo numero di negozi nel corso del tempo
  - memorizza i totali giornalieri delle vendite dei prodotti per negozio

# Schemi dimensionali, dettagli

- Dimensioni
  - tabelle dimensione, caratteristiche
  - chiavi
  - "snowflaking"
- Fatti
  - tabelle fatti, caratteristiche
  - additività

# Tablelle dimensione

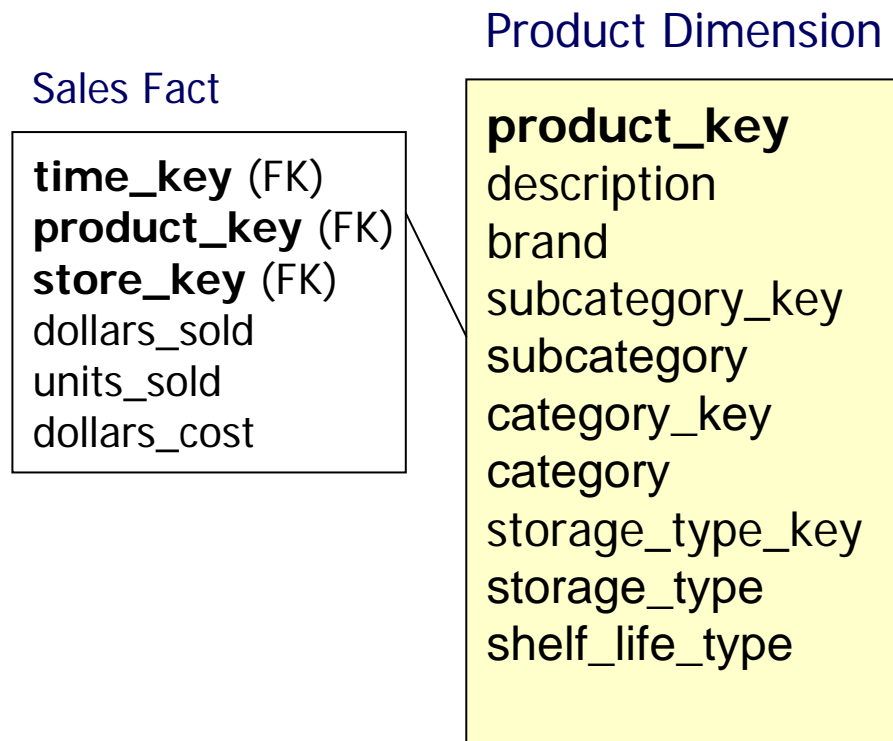
- Memorizza gli elementi (o membri) di una dimensione rispetto alla quale è interessante analizzare un processo (e le relative descrizioni)
  - ciascun record di una tabella dimensione descrive esattamente un elemento della rispettiva dimensione
    - un record di Time Dimension descrive un giorno (nell'ambito dell'intervallo temporale di interesse)
    - un record di Product Dimension descrive un prodotto in vendita nei negozi
  - i campi (non chiave) memorizzano gli attributi dei membri
    - gli attributi sono le proprietà dei membri, che sono solitamente testuali, discrete e descrittive

# Chiavi nei DW

- Negli schemi dimensionali, si preferiscono di solito chiavi semplici (numeriche) e “locali” (progressive), per vari motivi
  - sono piccole (e evitano le chiavi composte)
  - permettono di gestire casi speciali (ad esempio, la “non appartenenza” ad una categoria)
  - evitano problemi dovuti al riuso (esempio, le matricole dei laureati, oppure le fatture che ricominciano da 1 ogni anno)
  - evitano i cambi di tipo (esempio, le targhe auto) o i problemi dovuti alle fusioni aziendali

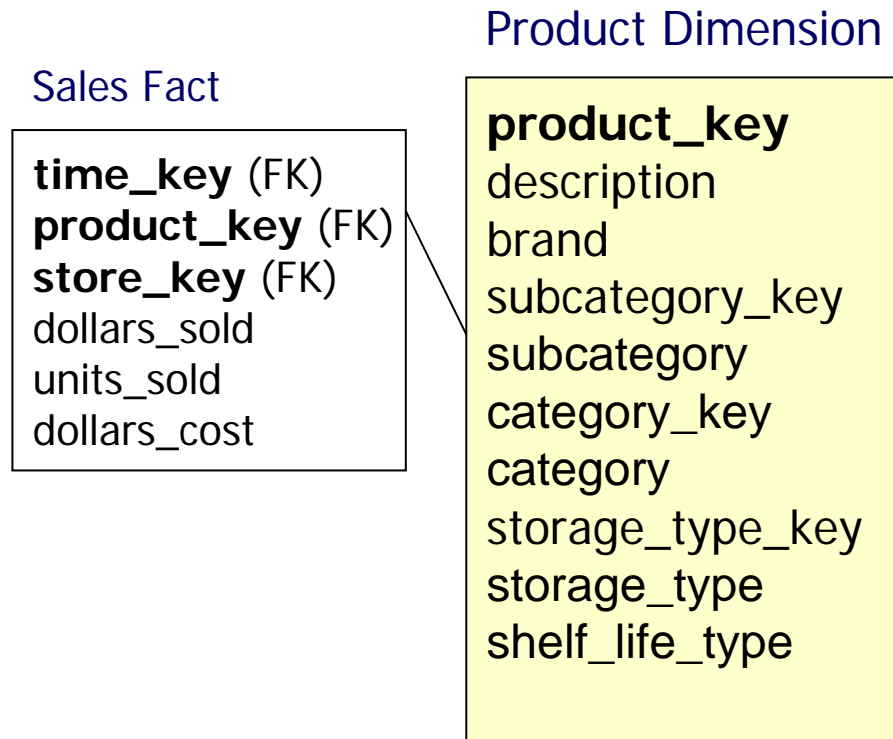
# Un inciso

- Le dimensioni sono spesso "**non normalizzate**"



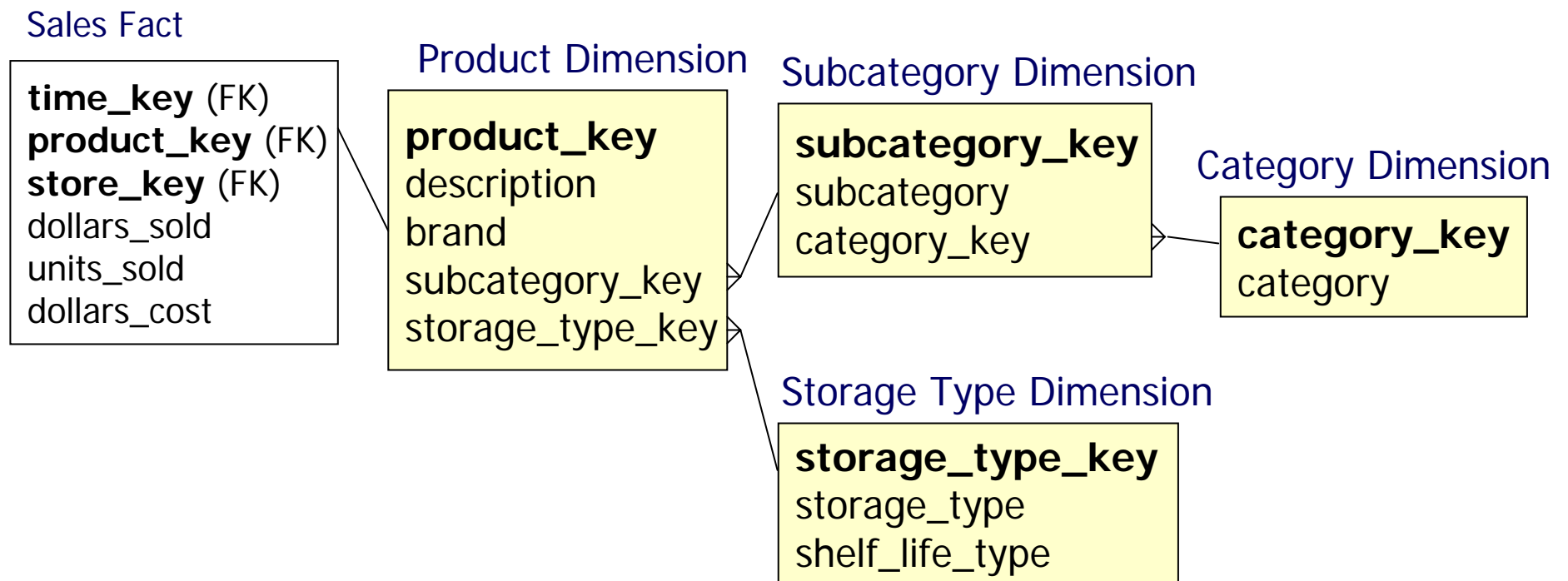
# Snowflaking

- Normalizzazione di una tabella dimensione, che evidenzia “gerarchie di attributi”



# Snowflaking

- Normalizzazione di una tabella dimensione, che evidenzia “gerarchie di attributi”





# Occupazione di memoria

- Stima dell'occupazione di memoria della base di dati dimensionale di esempio
  - Tempo: 2 anni di 365 giorni, ovvero 730 giorni
  - Negozi: 300
  - Prodotti: 30.000
  - Fatti relativi alle vendite
    - ipotizziamo un livello di sparsità del 10% delle vendite giornaliere dei prodotti nei negozi
      - ovvero, che ogni negozio vende giornalmente 3.000 diversi prodotti
    - $730 \times 300 \times 3000 = 630.000.000$  record

# Snowflaking: conviene?

- Lo snowflaking è solitamente svantaggioso
  - inutile per l'occupazione di memoria
    - ad esempio, supponiamo che la dimensione prodotto contenga 30.000 record, di circa 2.000 byte ciascuno
      - occupando quindi 60MB di memoria
    - la tabella fatti contiene invece 630.000.000 record, di circa 10 byte ciascuno
      - occupando quindi 6.3GB di memoria
    - le tabelle fatti sono sempre molto più grandi delle tabelle dimensione associate
      - anche riducendo l'occupazione di memoria della dimensione prodotto del 100%, l'occupazione di memoria complessiva è ridotta di meno dell'1%
  - può peggiorare decisamente le prestazioni

# Tabella fatti

- memorizza le misure numeriche di un processo
  - ogni record della tabella fatti memorizza una ennupla di misure (fatti) relativa a una combinazione degli elementi delle dimensioni ("all'intersezione di tutte le dimensioni")
- Nell'esempio
  - il processo (i fatti) è la vendita di prodotti nei negozi
  - le misure (i fatti) sono
    - l'incasso in dollari (dollars\_sold)
    - la quantità venduta (units\_sold)
    - le spese sostenute a fronte della vendita (dollars\_cost)
  - la grana è il totale per prodotto, negozio e giorno

## Tabella fatti, 2

- I campi della tabella fatti sono partizionati in due insiemi
  - chiave (composta)
    - sono riferimenti alle chiavi primarie delle tabelle dimensione
    - stabiliscono la grana della tabella fatti
  - altri campi: misure
    - talvolta chiamati fatti
    - solitamente valori numerici "comparabili" e additivi (vediamo tra poco)
- Una tabella fatti memorizza una funzione (in senso matematico) dalle dimensioni ai fatti
  - ovvero, una funzione che associa a ciascun fatto un valore per ciascuna possibile combinazione dei membri delle dimensioni

# Additività dei fatti

- Un fatto è **additivo** se ha senso sommarlo rispetto a ogni possibile combinazione delle dimensioni da cui dipende
  - l'incasso in dollari è additivo perché ha senso calcolare la somma degli incassi per un certo intervallo di tempo, insieme di prodotti e insieme di negozi
    - ad esempio, in un mese, per una categoria di prodotti e per i negozi in un'area geografica
  - l'additività è una proprietà importante, perché le applicazioni del data warehouse devono solitamente combinare i fatti descritti da molti record di una tabella fatti
    - il modo più comune di combinare un insieme di fatti è di sommarli (se questo ha senso)
    - è possibile anche l'uso di altre operazioni

# Semi additività e non additività

- I fatti possono essere anche
  - semi additivi
    - se ha senso sommarli solo rispetto ad alcune dimensioni
      - ad esempio, il numero di pezzi in deposito di un prodotto è sommabile rispetto alle categorie di prodotto e ai magazzini, ma non rispetto al tempo
  - non additivi
    - se non ha senso sommarli
  - può avere senso combinare fatti anche non completamente additivi mediante funzioni diverse dalla somma (ad esempio, medie pesate)

# Discussione

- Per il data warehouse, la modellazione dimensionale presenta dei vantaggi rispetto alla modellazione tradizionale (ER-BCNF) adottata nei sistemi operazionali
  - gli schemi dimensionali hanno una forma standardizzata e prevedibile
    - è facilmente comprensibile e rende possibile la navigazione dei dati
    - semplifica la scrittura delle applicazioni
    - ha una strategia di esecuzione efficiente
  - gli schemi dimensionali hanno una struttura simmetrica rispetto alle dimensioni
    - la progettazione può essere effettuata in modo indipendente per ciascuna dimensione
    - le interfacce utente e le strategie di esecuzione sono simmetriche

# Vantaggi della modellazione dimensionale

- gli schemi dimensionali sono facilmente estendibili
  - rispetto all'introduzione di nuovi fatti
  - rispetto all'introduzione di nuovi attributi per le dimensioni
  - rispetto all'introduzione di nuove dimensioni “supplementari”
    - se ogni record della tabella fatti dipende già funzionalmente dai membri della nuova dimensione
- si presta alla gestione e materializzazione di dati aggregati
- sono state già sviluppate numerose tecniche per la descrizione di tipologie fondamentali di fatti e dimensioni:
  - una sorta di “pattern” noti e documentati



# Interrogazioni di schemi dimensionali

- Gli attributi delle tabelle dimensione sono il principale strumento per l'interrogazione del data warehouse
  - gli attributi delle dimensioni vengono usati per
    - selezionare un sottoinsieme dei dati di interesse
      - vincolando il valore di uno o più attributi
      - ad esempio, le vendite nel corso dell'anno 2000
    - raggruppare i dati di interesse
      - usando gli attributi come intestazioni della tabella risultato
      - ad esempio, per mostrare le vendite per ciascuna categoria di prodotto in ciascun mese

# Attributi e interrogazioni

- Dati restituiti dall'interrogazione
  - somma degli incassi in dollari e delle quantità vendute
  - per ciascuna categoria di prodotto in ciascun mese
  - nel corso dell'anno 2000

(product) category	(time) month	(sum of) dollars_sold	(sum of) units_sold
Drinks	gennaio 2000	21.509,05	23.293
Drinks	febbraio 2000	19.486,93	22.216
Drinks	marzo 2000	21.986,43	23.532
Food	gennaio 2000	86.937,77	55.135
Supplies	gennaio 2000	21.554,17	13.541

6 giugno 2006 Data Warehousing

## Formato delle interrogazioni

- Le interrogazioni assumono solitamente il seguente formato standard

```
select p.category, t.month,  
       sum(f.dollars_sold), sum (f.items_sold)  
from sales_fact f, product p, time t  
where f.product_key = p.product_key  
      and f.time_key = t.time_key  
      and t.year = 2000  
group by p.category, t.month
```

# Formato delle interrogazioni

- Le interrogazione assumono solitamente il seguente formato standard

attributi di raggruppamento

fatti di interesse, aggregati

```
select p.category, t.month,
       sum(f.dollars_sold), sum (f.items_sold)
from sales_fact f, product p, time t
where f.product_key = p.product_key
      and f.time_key = t.time_key
      and t.year = 2000
group by p.category, t.month
```

tabella fatti e tabelle dimensione di interesse

condizioni di join imposte dallo schema dimensionale

condizioni di selezione

# Formato delle interrogazioni

- Le interrogazione assumono solitamente il seguente formato standard

attributi di raggruppamento

fatti di interesse, aggregati

```
select p.category, t.month,
       sum(f.dollars_sold), sum (f.items_sold)
from sales_fact f join product p
  on f.product_key = p.product_key
  join time t on f.time_key = t.time_key
where t.year = 2000
group by p.category, t.month
```

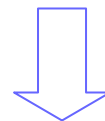
join tra fatti e dimensioni di interesse

condizioni di selezione

# Drill down

- L'operazione di drill down aggiunge dettaglio ai dati restituiti da una interrogazione
  - il drill down avviene aggiungendo un nuovo attributo nell'intestazione di una interrogazione e nel raggruppamento
  - diminuisce la grana dell'aggregazione

(product) category	(time) month	(sum of) dollars_sold	(sum of) units_sold
-----------------------	-----------------	--------------------------	------------------------



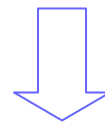
drill down

(product) category	(time) month	(store) city	(sum of) dollars_sold	(sum of) units_sold
-----------------------	-----------------	-----------------	--------------------------	------------------------

# Roll up

- L'operazione di roll up riduce il dettaglio dei dati restituiti da una interrogazione
  - il roll up avviene rimuovendo un attributo dall'intestazione di una interrogazione e dal raggruppamento
  - aumenta la grana dell'aggregazione

(product) category	(time) month	(sum of) dollars_sold	(sum of) units_sold
-----------------------	-----------------	--------------------------	------------------------



roll up

(product) category	(sum of) dollars_sold	(sum of) units_sold
-----------------------	--------------------------	------------------------

# Modello dimensionale, approfondimenti

- Tabelle fatti senza fatti
- Dimensioni supplementari

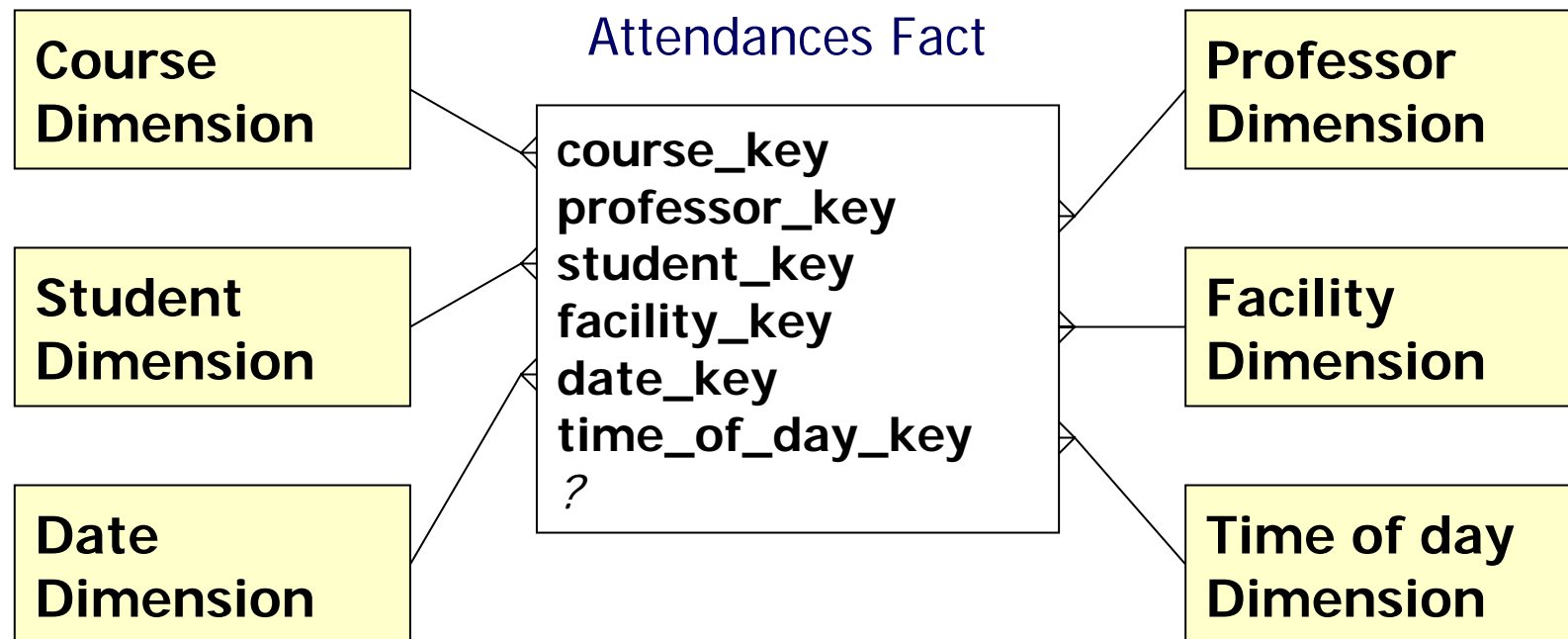


## Tabelle fatti senza fatti

- In tutti gli esempi finora, le tabelle fatti hanno la struttura
  - due o più chiavi esterne, riferimenti alle chiavi delle dimensioni
  - una o più misure
    - numeriche prese all'intersezioni delle dimensioni
- Alcuni processi interessanti sono caratterizzati da “fatti” che (apparentemente) non hanno proprietà misurabili
  - **tabelle fatti senza fatti (factless fact tables)**
    - Vediamo due casi

# Eventi

- In diverse situazioni bisogna memorizzare un grande numero di eventi, che si verificano all'intersezione di un certo numero di dimensioni
  - ad esempio, la presenza giornaliera di studenti nei corsi di una università

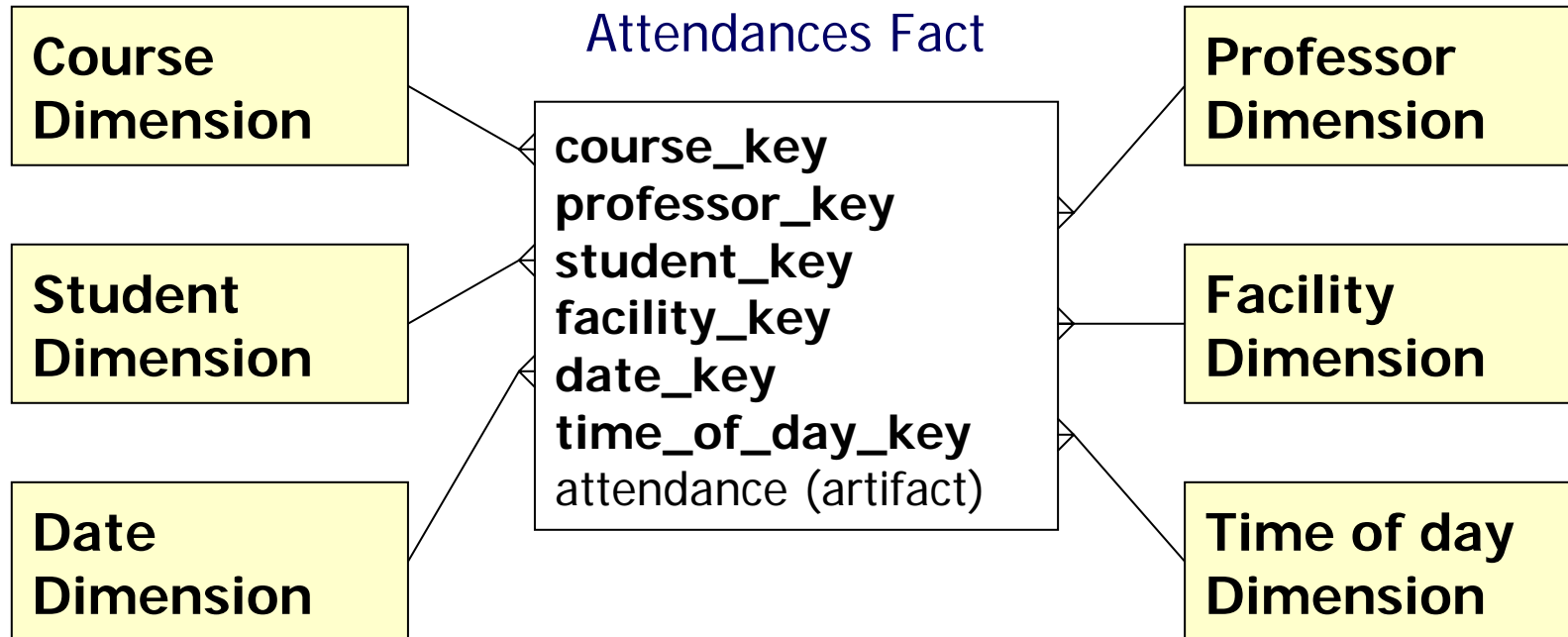


# Rappresentazione di eventi

- Un insieme di eventi (senza fatti) può essere rappresentato da una tabella fatti senza fatti e da un insieme delle dimensioni di interesse
  - analisi
    - quali sono stati i corsi più frequentati?
    - quali sono state le aule più utilizzate?
    - qual è stata l'occupazione media delle aule in funzione dell'ora del giorno?
- Molte di queste analisi richiedono di contare il numero di occorrenze distinte di uno certo insieme di attributi rispetto a un insieme di eventi
  - non possono essere sempre calcolate solo con la funzione COUNT di SQL
  - è spesso necessario scrivere COUNT(DISTINCT ...)

# Rappresentazione di eventi

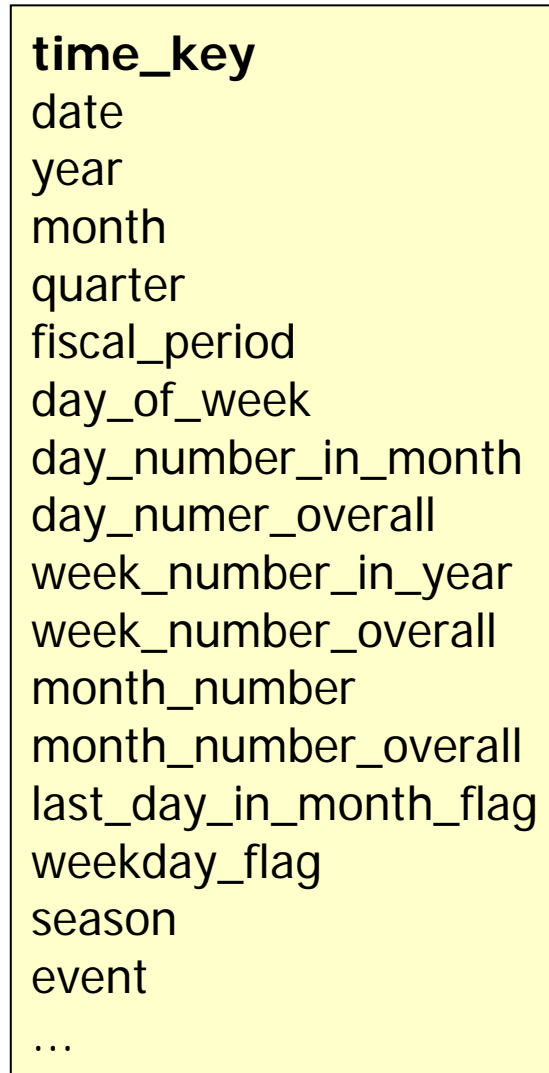
- Fatto numerico fittizio a cui viene assegnato valore 1



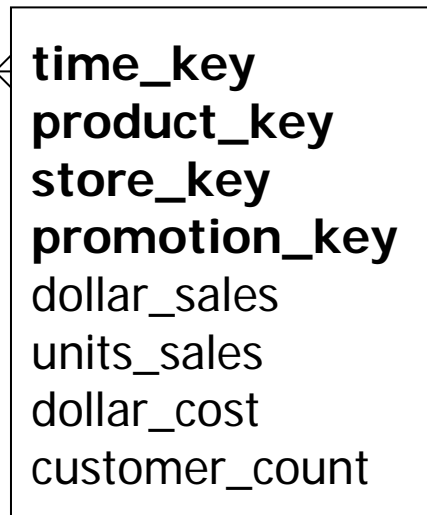
- è possibile scrivere interrogazioni corrette usando la funzione COUNT o SUM
  - le interrogazioni risultano più comprensibili

# Un'altra esigenza per tabelle senza fatti

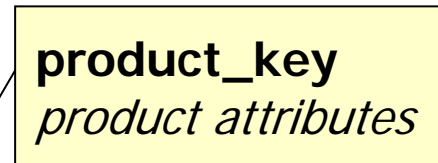
Time Dimension



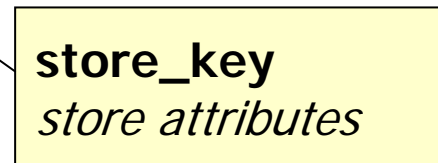
Sales Fact



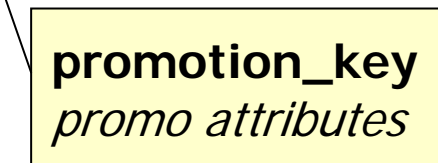
Product Dimension



Store Dimension



Promotion Dimension



## Tabelle "di copertura"

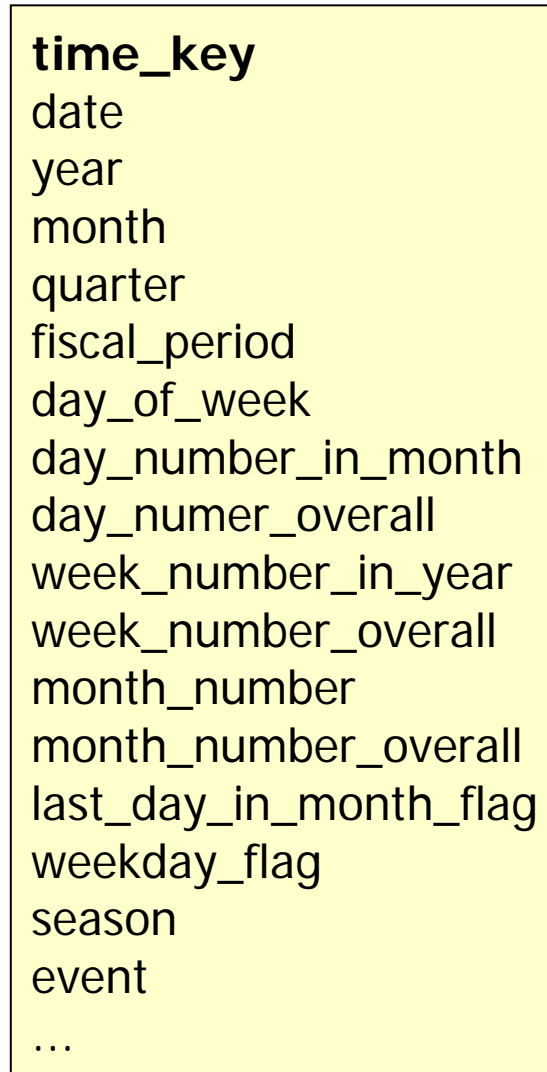
- Rappresentazione di eventi che non sono accaduti
  - nel processo delle vendite, lo schema dimensionale proposto non permette di effettuare la seguente analisi
    - quali prodotti in promozione non sono stati venduti?
  - è possibile usare una tabella "di copertura" per rappresentare i prodotti in promozione nei vari giorni e negozi
    - all'evento "essere in promozione" potrebbe non essere associato nessun particolare fatto misurabile
    - i prodotti in promozione non venduti possono essere calcolati per differenza insiemistica

# Tablelle di copertura

- In questo caso, la tabella di copertura delle promozioni è densa (rispetto ai prodotti in promozione)
  - deve però memorizzare solo i prodotti in promozione
    - e non i prodotti che non sono in promozione
  - anche in questo caso, può essere opportuno introdurre un fatto fittizio **existence**, di valore costante 1
  - se le promozioni sono settimanali, la grana della dimensione tempo può essere la settimana anziché il giorno

# Una tabella di copertura senza fatti

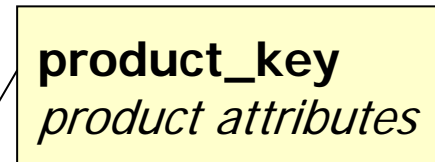
Time Dimension



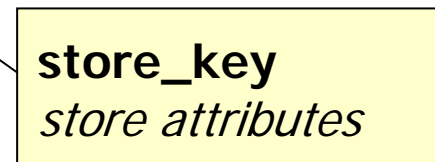
Promotion Coverage



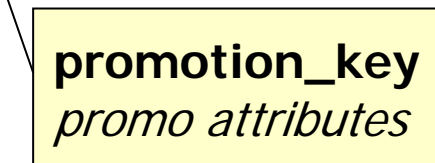
Product Dimension



Store Dimension



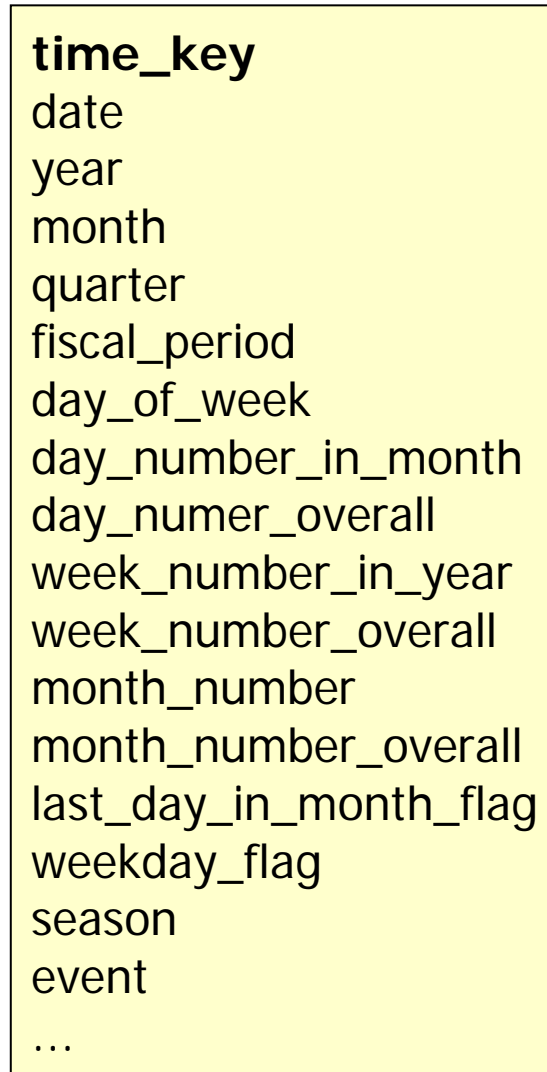
Promotion Dimension



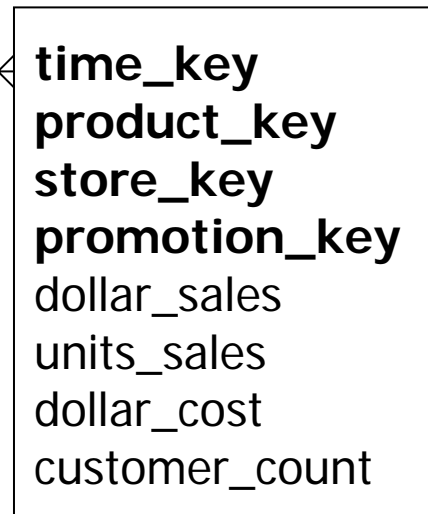


# Dimensioni primarie e secondarie

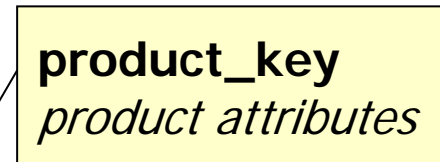
## Time Dimension



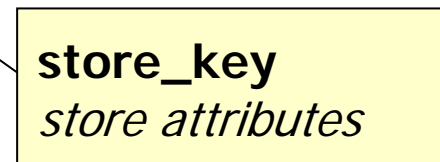
## Sales Fact



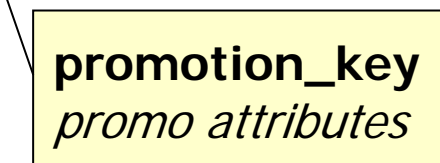
## Product Dimension



## Store Dimension



## Promotion Dimension



# Dimensioni

- Fissati il processo (vendite giornaliere dei prodotti) e la grana (unità di vendita per negozio per promozione per giorno) bisogna scegliere le dimensioni
  - in questo caso, la scelta delle dimensioni tempo, prodotto e negozio è immediata
    - tempo, prodotto e negozio sono **dimensioni primarie** nel senso che le misure relative ai movimenti giornalieri dei prodotti dipendono funzionalmente dal tempo, dal prodotto e dal negozio
  - un'altra dimensione è la dimensione promozione
    - ogni membro della dimensione promozione rappresenta le condizioni di promozione che si applicano alle vendite di una unità di vendita in un giorno in un negozio

# Dimensioni supplementari

- Promozione è una **dimensione supplementare**, nel senso che per ogni possibile combinazione delle dimensioni primarie è univoca la scelta del valore per questa dimensione
  - ovvero, la promozione dipende funzionalmente dalla data, dal prodotto e dal negozio
- Se una dimensione supplementare non fosse conforme alla grana della tabella fatti (richiedendo maggior dettaglio nei dati) allora la scelta della grana dovrebbe essere corretta (e la dimensione potrebbe essere primaria)

# Sommario

## Introduzione

- Basi di dati integrate, sì, ma ...
- OLTP e OLAP

- Data warehouse e data warehousing
- Dati multidimensionali

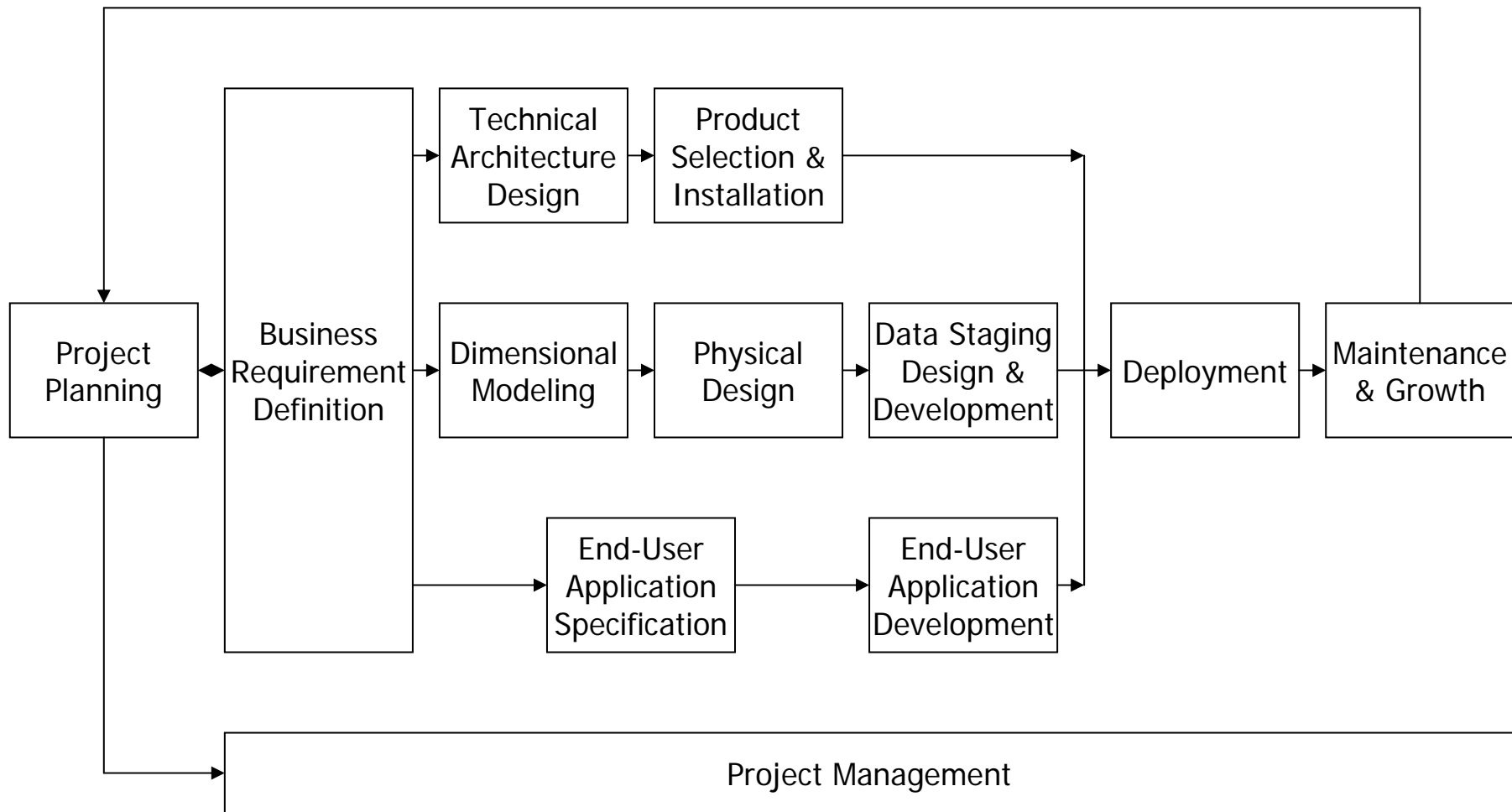
 Progettazione di data warehouse

- Studi di caso

# Ciclo di vita dimensionale

- **Il ciclo di vita dimensionale (Business Dimensional Lifecycle)** è una metodologia completa di progettazione e realizzazione di data warehouse (Kimball et al.)
  - fornisce il contesto di riferimento per la progettazione e realizzazione di data warehouse dimensionali
  - mediante un insieme di attività e di relazioni tra attività

# Ciclo di vita dimensionale



# Semplificando

- Progettazione di uno schema dimensionale
- Progettazione di un DW dimensionale

# Progettazione di uno schema dimensionale

- Una metodologia per la progettazione di uno schema dimensionale
  - uno schema dimensionale è composto da una singola tabella fatti e da un insieme di tabelle dimensione



# Progettazione di uno schema dimensionale

- La progettazione di uno schema dimensionale richiede lo svolgimento (in sequenza o quasi) dei seguenti quattro passi
  - scelta del **processo** (business process) da modellare
  - scelta della **grana** del processo
  - scelta delle **dimensioni** da cui dipende ciascun record della tabella fatti
  - scelta dei **fatti** misurabili che popoleranno ogni record della tabella fatti
- Queste scelte devono essere guidate
  - dai requisiti
  - dalle sorgenti informative disponibili

# Data-driven vs requirements-driven DW design

- Un DW va progettato con riferimento alle esigenze aziendali, altrimenti le probabilità di fallimento sono molto alte
- Dal punto di vista tecnico, possiamo anche concentrarci solo sui dati, ma sapendo che abbiamo una prospettiva limitata

# Progettazione di uno schema dimensionale

- Scelta del **processo** (business process) da modellare
  - per processo si intende un processo operativo, supportato da uno o più sistemi operazionali i cui dati possono essere utilizzati per popolare lo schema dimensionale
    - ad esempio, ordini, fatturazione, consegne, magazzino, vendite, ...
- Scelta della **grana** del processo
  - per grana si intende il livello di dettaglio atomico che deve essere rappresentato nella tabella fatti per il processo
    - livelli tipici per la grana sono le transazioni individuali, l'istantanea (snapshot) giornaliera individuale, l'istantanea mensile individuale, ...

# Progettazione di uno schema dimensionale

- Scelta delle **dimensioni** da cui dipende ciascun record della tabella fatti
  - una dimensione è un insieme di membri, di cui bisogna descrivere tutti gli attributi (solitamente testuali, discreti e descrittivi) necessari nelle selezioni e nei raggruppamenti
    - esempi di dimensioni sono il tempo, il prodotto, il cliente, la promozione, il magazzino, il tipo di transazione, lo stato, ...
- Scelta dei **fatti** misurabili che popoleranno ogni record della tabella fatti
  - i fatti sono misure (solitamente numeriche, continue e additivi) del processo selezionato
    - esempi di fatti sono la quantità venduta, l'incasso della vendita in dollari, ...

## Dall'ER al dimensionale (spunti)

- Individuare sottoschemi relativi a singoli processi
- Fatti:
  - nascono soprattutto dai requisiti; sullo schema ER
    - le entità coinvolte in diverse relationship 1:n con cardinalità massima 1
    - le relationship molti a molti con attributi non chiave numerici e additivi (o “da contare”):
- Dimensioni
  - dalle relationship o entità collegate ai fatti (o loro catene "denormalizzate")

# Progettazione di un DW dimensionale

- La progettazione dimensionale è la progettazione logica dei dati del data warehouse, basata sull'architettura a bus
  - progettazione di un insieme di dimensioni conformi
  - progettazione degli schemi dimensionali
  - analisi delle sorgenti informative
    - comprensione delle sorgenti informative disponibili e delle loro qualità
    - progettazione preliminare del mapping dei dati dalle sorgenti informative al data warehouse
  - piano preliminare delle aggregazioni

# Progettazione dei data mart

- Un data warehouse dimensionale viene progettato come un insieme coerente di data mart ognuno dei quali è
  - un sottoinsieme logico dell'intero data warehouse
    - è la restrizione del data warehouse a un singolo processo dell'organizzazione, o a un insieme di attività correlate
  - una collezione di fatti correlati che devono essere analizzati insieme
    - un insieme di schemi dimensionali correlati
- Un insieme di data mart è “coerente” se è organizzato secondo una architettura a bus basata su dimensioni conformi e fatti conformi
  - cioè con significato uniforme in tutto il data warehouse

## Selezione dei data mart

- La progettazione dimensionale di un data warehouse inizia con la selezione ed elencazione dei data mart
  - il criterio principale è
    - un data mart deve rappresentare una collezione di fatti correlati che devono essere analizzati insieme
  - inizialmente, ciascun data mart dovrebbe avere origine in un singolo processo dell'organizzazione e in una singola sorgente informativa
    - successivamente, sarà possibile identificare data mart relativi a più processi e/o con dati derivanti da più sorgenti informative
  - i data mart possono essere (parzialmente) sovrapposti
  - in una grande organizzazione ha (secondo gli esperti) da 10 a 30 data mart



## Esempio — una grande azienda telefonica

- Data mart a sorgente singola
  - fatturazione clienti (residenziali e commerciali)
  - gestione ordini
  - gestione dei malfunzionamenti
  - pubblicità sulle pagine gialle
  - servizio clienti e informazioni sulle fatture
  - offerte promozionali e comunicazioni ai clienti
  - dettaglio delle chiamate dal punto di vista della fatturazione
  - dettaglio delle chiamate dal punto di vista del carico della rete telefonica
  - inventario clienti
  - inventario della rete telefonica
  - ...

## Selezione dei data mart

- La realizzazione di un data warehouse inizia (di solito) da un data mart
  - significativo
    - ovvero, permette analisi interessanti
  - semplice da realizzare
    - di solito, a sorgente singola
- Successivamente, possono essere realizzati altri data mart, più complessi
  - ad esempio, a sorgente multipla
    - come il data mart della profittabilità dei clienti

# Progettazione delle dimensioni

- Scelti i data mart di interesse, si procede selezionando e elencando le dimensioni di interesse
  - bisogna progettare un insieme di dimensioni da usare in modo conforme (o conformato) in tutti i data mart del data warehouse
  - si può iniziare identificando le dimensioni di interesse per ciascun data mart

# Dimensione conforme

- Una dimensione che ha lo stesso significato in tutti i data mart (e cioè con tutte le tabelle di fatti con cui va in join)
- Di solito, è quindi sempre la stessa
- Dimensioni molto usate (ad esempio quella temporale) diventano standard aziendali

## Esempio — una grande azienda telefonica

- Dimensioni per il data mart della fatturazione clienti
  - tempo (data di fatturazione)
  - cliente (residenziale o commerciale)
  - servizio
  - tariffa (compresa promozione)
  - fornitore di servizi locali
- Dimensioni per il data mart del dettaglio delle chiamate dal punto di vista della fatturazione
  - chiamante
  - chiamato
  - fornitore di servizi non locali

# La matrice dell'architettura a bus

- I data mart e le dimensioni possono essere utilmente correlati in una matrice che descrive l'architettura a bus del data warehouse
  - ciascuna riga della matrice rappresenta un data mart
  - ciascuna colonna della matrice rappresenta una dimensione
  - ciascun elemento della matrice, all'intersezione di un data mart e una dimensione, viene marcato se la dimensione è di interesse per il data mart
- La definizione della matrice che descrive l'architettura a bus del data warehouse è una “pietra miliare” fondamentale nella progettazione dell'intero data warehouse
  - è il luogo dove viene fissato l'insieme delle dimensioni conformi del data warehouse

# Esempio — grande azienda telefonica

	Time	Customer	Service	Rate Category	Local Service Provider	Calling Party	Called Party	Long-Distance Provider	International Organization	Employee	Location	Equipment Type	Supplier	Item Supplied	Weather	Account Status
Customer billing	X	X	X	X	X			X			X					X
Service orders	X	X	X		X			X	X	X	X	X			X	X
Trouble reports	X	X	X		X	X		X	X	X	X	X	X	X	X	X
Yellow Page Ads	X	X		X		X			X	X	X					X
Customer Inquiries	X	X	X	X	X	X		X	X	X	X				X	X
Promotions & Comm'n	X	X	X	X	X	X		X	X	X	X	X	X	X		X
Billing Call Detail	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X
Network Call Detail	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X
Customer Inventory	X	X	X	X	X			X	X		X	X	X	X		X
Network Inventory	X		X						X	X	X	X	X	X		
Real Estate	X								X	X	X	X				
Labor & Payroll	X								X	X	X					
Computer Charges	X	X	X		X			X	X	X	X	X	X	X		
Purchase Orders	X								X	X	X	X	X	X		
Supplier Deliveries	X								X	X	X	X	X	X		
Combined Fields Ops.	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X
Customer Reln. Mgmt.	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Customer	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

6 giugno 2006

Data Warehousing

119

# Progettazione degli schemi dimensionali

- Successivamente, va completato il progetto degli schemi dimensionali
  - selezione degli attributi delle dimensioni
  - scelta della strategia di gestione dei cambiamenti lenti, per ciascuna dimensione
  - altre scelte di rappresentazione
    - minidimensioni, dimensioni e fatti eterogenei, aggregazioni
  - durata storica del data warehouse
    - quanti dati storici devono essere rappresentati nel data warehouse? con quale grana?
  - pianificazione del caricamento incrementale
    - con che periodicità deve essere aggiornato il data warehouse? con che urgenza?



# Discussione

- Come in tutte le attività di progetto, le fasi di una metodologia non vengono mai eseguite in una sequenza perfetta
  - spesso, lo svolgimento di una fase richiede la correzione di scelte fatte nei passi precedenti
    - ad esempio, se la scelta delle dimensioni portasse a una grana diversa per uno schema dimensionale, o se fosse impossibile estrarre dei dati dalle sorgenti informative
  - in alcuni casi, può essere opportuno avviare una fase anche se la fase immediatamente precedente non è stata conclusa
    - ad esempio, iniziare la progettazione di alcuni data mart anche quando la selezione dei data mart non è stata completata

# Progettazione, approfondimenti

# Convenzioni nella progettazione

- Alcune indicazioni stilistiche (e non) da adottare nella progettazione
  - i nomi (etichette) per data mart, dimensioni, attributi e fatti devono essere scelti attentamente nel dominio applicativo del data warehouse
    - devono essere nomi accettabili per gli utenti finali
  - ogni attributo vive in una sola dimensione, un fatto può essere ripetuto in più tabelle fatti
  - se una dimensione deve essere ripetuta, probabilmente indica ruoli diversi della stessa dimensione e, quindi, dimensioni diverse
    - ad esempio, data del servizio e data di scadenza della fattura

# Convenzioni nella progettazione

- i campi significativi delle sorgenti informative corrispondono a uno o più campi del data warehouse
  - ad esempio, un campo prodotto può essere rappresentato dal codice del prodotto, descrizione sintetica, descrizione completa
- ogni fatto dovrebbe essere associato a una modalità di aggregazione di default
  - ad esempio, somma, minimo, massimo, ultimo valore, semi additivo, algoritmo speciale, non additivo, ...
- è opportuno evidenziare nelle dimensioni le eventuali gerarchie di aggregazione significative per l'utente

# Analisi delle sorgenti informative

- Progettato lo schema logico del data warehouse, bisogna
  - descrivere le sorgenti informative a disposizione
    - ovvero, le sorgenti informative individuate nella fase di raccolta e analisi dei requisiti
  - progettare la trasformazione dei dati dalle sorgenti informative al data warehouse

# Selezione delle sorgenti informative

- Il criterio principale per la selezione delle sorgenti informative da cui estrarre i dati per il data warehouse è relativo all'accuratezza dei dati
  - uno stesso dato può attraversare più sistemi, per essere elaborato in più modi
    - il transito dei dati da un sistema all'altro avviene insieme a delle trasformazioni, che arricchiscono o sintetizzano i dati originari
  - in generale, la qualità di un dato può diminuire allontanandosi dal sistema in cui è stato immesso o generato
    - è quindi opportuno catturare i dati quando vengono generati (possibilmente, dopo che sono stati puliti)

# Reverse engineering delle sorgenti informative

- Le sorgenti informative selezionate per l'estrazione devono essere comprese in dettaglio
  - è opportuno l'uso di modelli formali per descrivere la struttura dei dati
    - schema logico dei dati
    - schema concettuale dei dati
    - glossario dei dati
  - gli schemi concettuali, se non sono disponibili, possono essere ottenuti mediante una attività di reverse engineering dei dati
    - orientata appunto alla comprensione e descrizione concettuale delle sorgenti informative

# Progettazione della trasformazione dei dati

- Per ciascun elemento (record e campo) del data warehouse bisogna progettare la trasformazione necessaria a calcolare l'elemento dalle sorgenti informative
  - descrivendo per ciascun dato
    - il ruolo nel data warehouse
    - la sorgente (o le sorgenti) da cui viene estratto
    - le trasformazioni necessarie



# Piano delle aggregazioni

- Ogni data warehouse contiene dati pre-aggregati
  - la disponibilità di dati pre-aggregati è lo strumento singolo più efficace nel controllo delle prestazioni delle attività di interrogazione del data warehouse
  - i dati aggregati devono essere rappresentati in tabelle fatti apposite, separate dalle tabelle fatti da cui sono calcolati
    - usando anche tabelle dimensione aggregate, contratte e conformate
  - i dati effettivamente aggregati possono cambiare nel corso del tempo
    - le aggregazioni deve essere gestite mediante un “navigatore” e metadati opportuni
  - un piano preliminare delle aggregazioni è comunque utile, ad esempio nella stima dell’occupazione di memoria

# Fasi nel ciclo di vita dimensionale

- pianificazione del progetto
- gestione del progetto
- raccolta e analisi dei requisiti
- progettazione del data warehouse
  - progettazione dei dati
    - progettazione dimensionale, progettazione fisica, progetto della preparazione dei dati
  - progettazione tecnologica
    - progettazione dell'architettura tecnica, selezione e installazione dei prodotti
  - progettazione delle applicazioni
    - specifica delle applicazioni, sviluppo delle applicazioni
- installazione e avviamento
- manutenzione e crescita

# Processi in un data warehouse

- I processi di base in un data warehouse comprendono
  - processi nell'area di preparazione dei dati (attività "notturne")
    - estrazione, trasformazione, caricamento e indicizzazione, controllo di qualità
    - aggiornamento del data warehouse
  - processi utente (attività "diurne")
    - interrogazione
  - processi di amministrazione
    - gestione della sicurezza
    - auditing
    - backup e recovery
    - gestione del feedback

# Estrazione

- L'**estrazione** è il primo passo nel transito dei dati dalle sorgenti informative al data warehouse
  - più precisamente, l'attività di estrazione riguarda
    - la comprensione e la lettura delle sorgenti informative
    - la copiatura nell'area di preparazione dei dati delle porzioni di sorgenti informative che sono necessarie al popolamento del data warehouse

# Trasformazione

- I dati estratti dalle sorgenti informative, prima di essere caricati nel data warehouse, sono sottoposti a diverse **trasformazioni**
  - pulizia
    - per risolvere errori, conflitti, incompletezze
    - per riportare i dati in un formato standard
  - eliminazione di campi non significativi
  - combinazione
    - per identificare e correlare i dati associati alla rappresentazione di uno stesso oggetto in più sorgenti informative
  - creazione di chiavi
    - le chiavi usate nel data warehouse sono diverse da quelle usate nelle sorgenti informative
  - creazione di aggregati

## Caricamento e controllo di qualità

- Dopo il processo di trasformazione, i dati sono organizzati per essere caricati direttamente nel data warehouse
  - il caricamento consiste nella concatenazione (e/o aggiornamento) di un insieme di record per ciascuna tabella (fatti o dimensione) del data warehouse
    - durante il caricamento il data warehouse non è solitamente disponibile per l'accesso e l'interrogazione
  - il caricamento dei dati nel data warehouse viene seguito da una verifica della correttezza delle operazioni di preparazione e caricamento, mediante un'analisi di qualità dei dati
    - se il controllo di qualità ha successo, il nuovo data warehouse è pronto per l'accesso e l'interrogazione

# Aggiornamento del data warehouse

- I dati del data warehouse devono essere aggiornati, anche frequentemente
  - aggiornamenti ordinari e periodici
    - caricamento incrementale di nuovi dati nel data warehouse
  - aggiornamenti straordinari
    - correzione di dati (record e/o schemi)
    - sono aggiornamenti orientati al miglioramento della qualità complessiva dei dati

# Processi di amministrazione

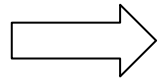
- Auditing
  - sull'origine dei dati (ad esempio, per certificarne la qualità)
  - sull'uso del data warehouse (per l'ottimizzazione del data warehouse)
- Gestione della sicurezza
- Backup e recovery
- Gestione del feedback
  - il transito principale dei dati va dalle sorgenti informative al data warehouse e dal data warehouse agli strumenti di analisi
    - dati “puliti” e risultati di analisi significativi possono transitare nella direzione opposta



# Sommario

## Introduzione

- Basi di dati integrate, sì, ma ...
- OLTP e OLAP
- Data warehouse e data warehousing
- Dati multidimensionali
- Progettazione di data warehouse



Studi di caso

# Studi di caso

- ▶ **Vendite**
- ▶ Inventario
- ▶ Catena del valore

# Il processo delle vendite

- Si consideri il seguente caso di studio, relativo al processo delle vendite in una catena di negozi alimentari (grocery store)
  - lavoriamo nella direzione di una grande catena di alimentari (negli Stati Uniti)
  - la catena comprende 500 grandi negozi di alimentari, distribuiti in tre stati
  - ogni negozio è un supermercato con diversi reparti (department)
    - ad esempio, drogheria, surgelati, latticini, carne, frutta e verdura, pane, pasta, fiori, liquori, ...

## Il processo delle vendite (2)

- ogni negozio ha circa 60.000 prodotti individuali nei suoi scaffali
  - i prodotti individuali sono chiamati **unità di vendita (SKU, stock keeping unit)**
    - ad esempio, una SKU è la lattina di Diet Coke
  - ogni variante di confezionamento dei prodotti costituisce una diversa SKU
    - ad esempio, la confezione da 6 lattine di Diet Coke è una SKU diversa dalla lattina di Diet Coke
- circa 40.000 delle SKU vengono da fornitori esterni, e su di esse è stampato un codice a barre chiamato **codice universale del prodotto (UPC, universal product code)**
  - la grana degli UPC è la stessa delle SKU

## Il processo delle vendite (3)

- le altre 20.000 SKU corrispondono a prodotti come frutta e carne, che non sono confezionati o che sono confezionati localmente, e non hanno UPC
  - anche a questi prodotti è associato un numero (codice) SKU
    - questo codice viene assegnato localmente, ed è condiviso da tutti i negozi della catena

## Il processo delle vendite (4)

- Dove vengono raccolti i dati della catena di negozi alimentari?
  - per i dati relativi alle vendite, sicuramente in ciascuna cassa, mediante dei sistemi **POS** (point of sale)
  - per quanti riguarda gli acquisti
    - alcuni negozi usano un sistema di codici a barre anche alla consegna delle merci
    - altri negozi non registrano le merci consegnate
      - ma vengono raccolte le bolle e le fatture
  - l'inventario è spesso realizzato girando tra gli scaffali e guardando quali prodotti sono assenti

## Il processo delle vendite (5)

- La direzione della catena si occupa della logistica delle ordinazioni, della disposizione delle merci sugli scaffali, della vendita dei prodotti e della massimizzazione del profitto
  - sorgenti del profitto
    - fissare per i prodotti il prezzo più alto possibile
    - ridurre i costi di acquisizione dei prodotti e le spese generali
    - attrarre quanti più clienti è possibile
  - le scelte sotto il controllo della direzione della catena di negozi riguardano
    - i prezzi dei prodotti
    - le promozioni

## Il processo delle vendite (6)

- le promozioni comprendono
  - riduzioni temporanee di prezzo (TPR)
  - pubblicità (su diversi media)
  - esposizione sugli scaffali
  - esposizione alla fine dei corridoi
- Uno degli obiettivi della direzione è la comprensione dell'impatto delle promozioni sulle vendite e, quindi, sui profitti
  - per comprendere l'impatto delle promozioni passate
  - per pianificare e progettare le promozioni future



## Il data mart delle vendite

- La progettazione di un data warehouse (e di ogni singolo schema dimensionale che lo compone) è basata sulla comprensione del processo e dei dati effettivamente disponibili
- Il data warehouse della catena di negozi alimentari riguarda il processo delle vendite dei prodotti nei negozi
  - viene deciso di costruire il data mart delle vendite giornaliere dei prodotti

# Scelta della grana

- La grana scelta per il data mart per il processo delle vendite è
  - unità di vendita (SKU) per negozio per promozione per giorno
- La scelta della grana ha influenza
  - sulle dimensioni usate nel data mart
  - sul tipo di analisi che può essere effettuato
  - sull'occupazione di memoria del data mart

## Altre scelte per la grana

- Scelte alternative per la grana
  - per voce di vendita (transazione individuale)
    - informazioni su ciascuna voce (riga) di ciascuno scontrino di vendita
    - se è nota l'identità del cliente, permette di effettuare interessanti analisi di market basket
    - l'occupazione di memoria del data mart potrebbe essere enorme
  - unità di vendita per negozio per promozione per settimana
    - non permette di distinguere le vendite nei fine settimana da quelle degli altri giorni
  - Prodotto per negozio per promozione
    - Non permette di distinguere l'importanza del confezionamento

## Alcune possibili analisi

- La scelta di grana fatta (unità di vendita per negozio per promozione per giorno) permette ad esempio di effettuare le seguenti analisi
  - è utile vendere più varianti di confezionamento di uno stesso prodotto?
    - possibile solo se la grana riguarda l'unità di vendita
  - di quali prodotti diminuiscono le vendite a fronte della promozione di un certo altro prodotto?
    - possibile solo se la grana riguarda le promozioni
  - quali sono i dieci prodotti più venduti dai miei concorrenti che invece la catena non vende?
    - sulla base di ulteriori dati forniti da società di analisi specializzate

## Altre considerazioni sulle SKU

- Nessuna delle analisi proposte è interessata esplicitamente alle singole SKU
  - non è solitamente interessante presentare l'unità di vendita individuale nel risultato dell'analisi
  - tuttavia, in un data warehouse è necessario memorizzare dati a una grana sufficientemente piccola, per permettere alle interrogazioni di selezionare e raggruppare i dati in modo sufficientemente preciso e mirato

# Scelta delle dimensioni

- Fissati il processo (vendite giornaliere dei prodotti) e la grana (unità di vendita per negozio per promozione per giorno) bisogna scegliere le dimensioni
  - in questo caso, la scelta delle dimensioni primarie tempo, prodotto e negozio è immediata
    - tempo, prodotto e negozio sono **dimensioni primarie** nel senso che le misure relative ai movimenti giornalieri dei prodotti dipendono funzionalmente dal tempo, dal prodotto e dal negozio
  - un'altra dimensione è la dimensione promozione
    - ogni membro della dimensione promozione rappresenta una combinazione delle promozioni che si applica alle vendite di una unità di vendita in un giorno in un negozio

## Dimensioni supplementari

- Promozione è una **dimensione supplementare**, nel senso che per ogni possibile combinazione delle dimensioni primarie è univoca la scelta del valore per questa dimensione
  - ovvero, la promozione dipende funzionalmente dalla data, dal prodotto e dal negozio
- Se una dimensione supplementare non fosse conforme alla grana della tabella fatti (richiedendo maggior dettaglio nei dati) allora la scelta della grana dovrebbe essere corretta (e la dimensione potrebbe essere primaria)
  - promozione sarebbe una dimensione primaria se ogni membro della dimensione rappresentasse una combinazione delle promozioni che è stata effettivamente applicata a una vendita

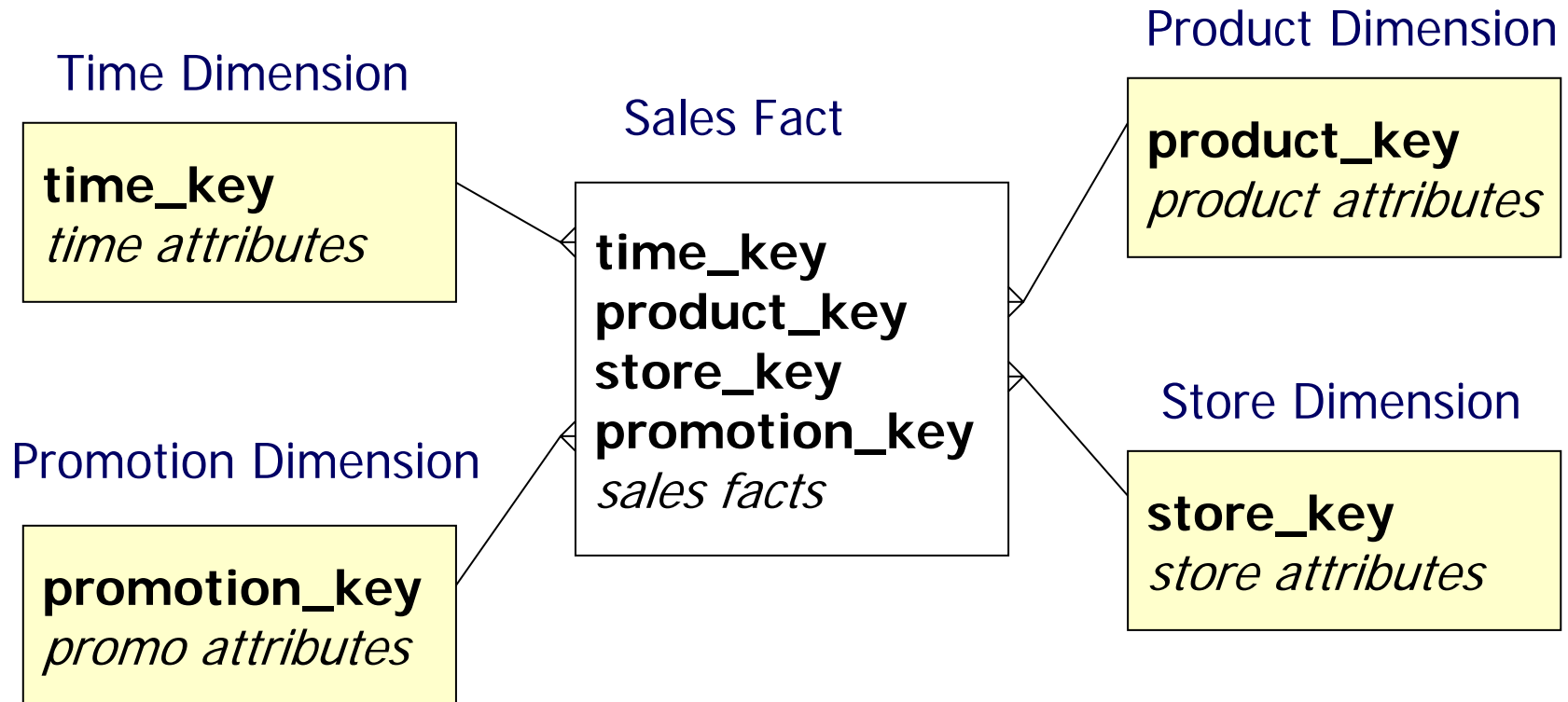
# Scelta delle dimensioni

- Altre ipotetiche dimensioni supplementari (non scelte perché non accessibili dalle sorgenti informative a disposizione)
  - il fornitore che ha fornito il prodotto al negozio
  - il responsabile delle vendite nel negozio nel giorno



# Schema dimensionale

- Versione preliminare dello schema dimensionale per le vendite



– la scelta degli attributi delle dimensioni verrà fatta più avanti

## Scelta dei fatti

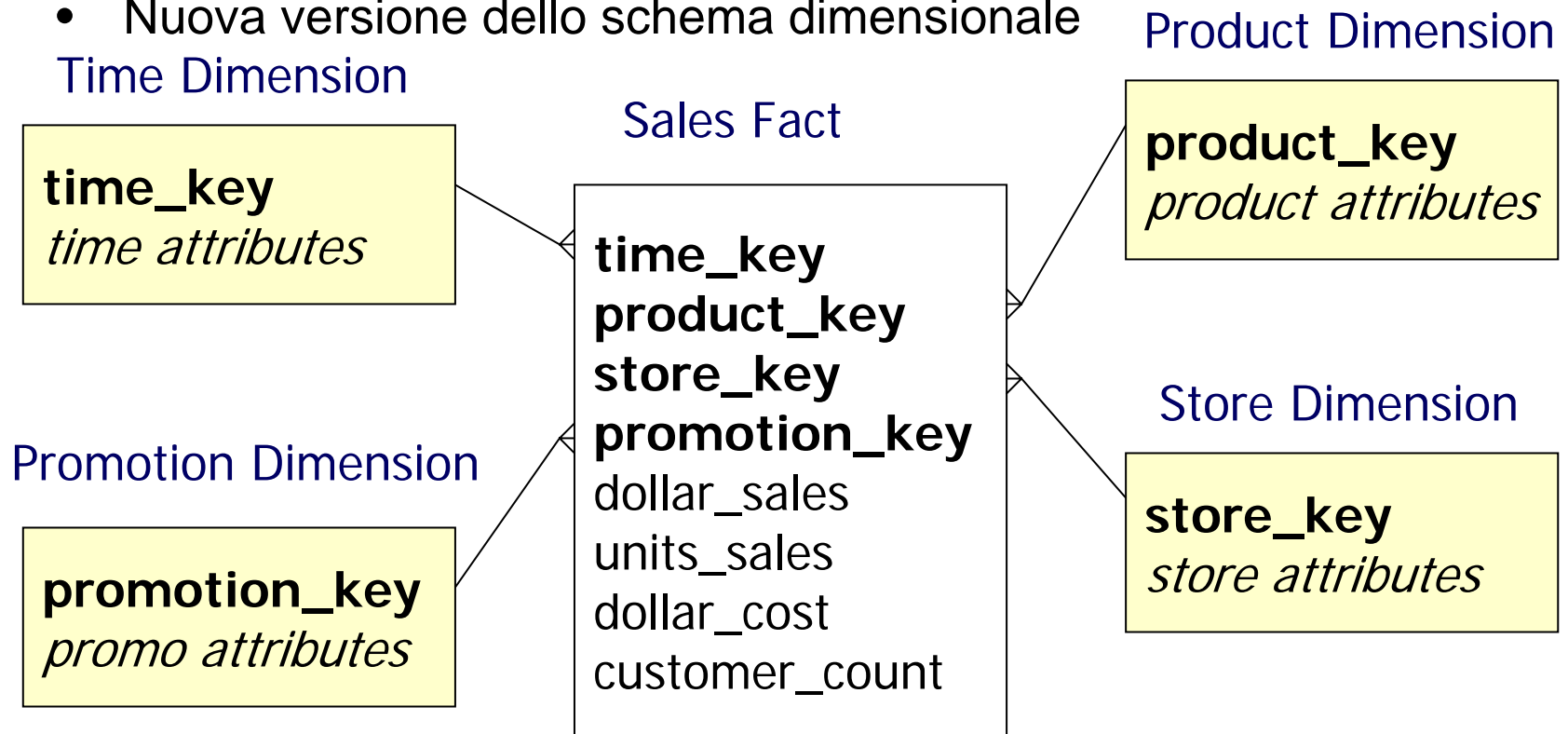
- Le misure disponibili relativamente alle vendite giornaliere dei prodotti (per unità di vendita per negozio per promozione per giorno) sono
  - incasso totale in dollari (**dollar\_sales**)
  - numero totale di unità vendute (**units\_sales**)
  - costo totale in dollari (**dollar\_cost**)
    - relativa al prodotto consegnato dal fornitore al negozio
  - numero di clienti (**customer\_count**)
    - che hanno acquistato il prodotto (SKU)
      - calcolato contando il numero di scontrini in cui è presente l'unità di vendita

## Disponibilità dei fatti

- Le misure relative alle vendite sono ottenute dai POS
  - i POS permettono di esportare tutti i dati relativi agli scontrini emessi giornalmente
  - questi dati possono essere elaborati per fornire le informazioni relative ai fatti scelti alla grana scelta

# Schema dimensionale

- Nuova versione dello schema dimensionale  
Time Dimension



## Stima della taglia dei dati

- Alcune stime relative alla quantità di dati
  - il numero complessivo di voci nelle transazioni individuali può essere calcolato conoscendo l'incasso complessivo della catena ( $\$4 \cdot 10^9$  per anno) e il costo medio della voce di vendita ( $\$2$ )
    - ci sono  $2 \cdot 10^9$  voci nelle transazioni individuali
  - le voci nelle transazioni individuali giornaliere per negozio sono  $2 \cdot 10^9 / (365 \cdot 500) = 11.000$  circa
  - ogni negozio offre 30.000 SKU, e ne vende giornalmente 3.000
    - il trasferimento dei dati dai negozi al data warehouse deve preferibilmente riguardare dati pre-elaborati

# Occupazione di memoria della tabella fatti

- Stima dell'occupazione di memoria della tabella fatti
  - ipotesi
    - la chiave delle tabelle dimensione è un intero
      - di 4 byte per tempo, prodotto e promozione
      - di 2 byte per negozio
    - i quattro campi chiave della tabella fatti occupano 14 byte
    - ogni fatto è rappresentato da un intero di 4 byte
  - ogni record della tabella fatti occupa 30 byte
  - la tabella fatti contiene  $500 \cdot 3.000 \cdot 365 = 547.500.000$  record per anno
  - se vengono mantenuti dati storici relativi a due anni, l'occupazione di memoria della tabella fatti è di circa 30GB (di spazio primario)

# La dimensione tempo

- La dimensione tempo (nel caso in esame) descrive i giorni di un intervallo temporale di interesse
  - i membri della dimensione tempo sono i giorni dell'intervallo di interesse
- La dimensione tempo è presente nella maggior parte degli schemi dimensionali, e praticamente in tutti i data warehouse
  - la realizzazione di una tabella dimensione per il tempo è semplice
    - può essere facilmente pre-calcolata
    - i giorni per dieci anni sono poco più di 3.650

# La dimensione tempo

- È necessaria una tabella dimensione tempo esplicita? Non potrebbe essere invece usato un campo di tipo data?
  - in alcuni (rari) casi, l'uso di un campo di tipo data è una scelta sufficiente
    - ma non c'è solitamente nessun vantaggio evidente per questa scelta
  - i vantaggi di avere una tabella dimensione tempo esplicita comprendono
    - la possibilità di distinguere tra giorni feriali, festivi e prefestivi, di considerare sia intervalli temporali solari che fiscali, di tenere conto delle stagioni di vendita, di eventi (ad esempio, la finale del Super Bowl) e altro



# Chiave e attributi della dimensione tempo

- **time\_key** è la chiave, un numero intero
- **date** è la data del giorno (ad esempio, 25 ottobre 2000)
- **year** è l'anno (2000)
- **month** è il mese (ottobre 2000)
- **quarter** è il numero del trimestre (4)
- **fiscal\_period** è il periodo fiscale (4Q-2000)
- **day\_of\_week** è il giorno della settimana (“mercoledì”)
  - utile, ad esempio, per confrontare le vendite dei mercoledì rispetto ai venerdì
- **day\_number\_in\_month** è il giorno nel mese (25)
  - per confrontare le vendite negli stessi giorni in mesi diversi

# Chiave e attributi della dimensione tempo

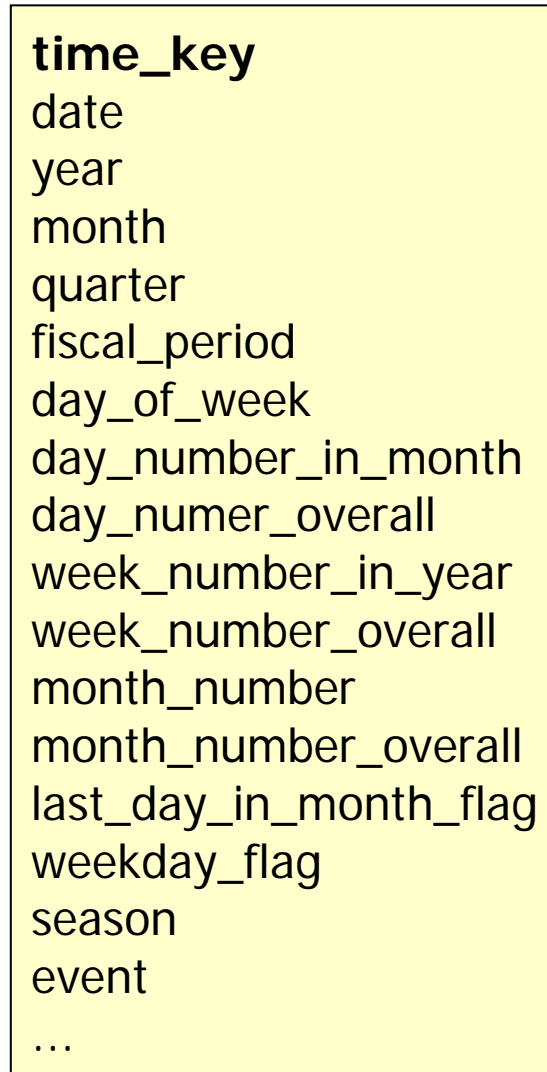
- **day\_number\_overall** assegna una numerazione consecutiva a tutti i giorni di interesse
  - utile per calcoli aritmetici sulle date
- **week\_number\_in\_year**, **week\_number\_overall**, **month\_number**, **month\_number\_overall** hanno un significato analogo
- **last\_day\_in\_month\_flag** permette di selezionare l'ultimo giorno di ciascun mese
- **holiday\_flag** permette di selezionare i giorni feriali/festivi
- **weekday\_flag** permette di selezionare i giorni lavorativi

# Chiave e attributi della dimensione tempo

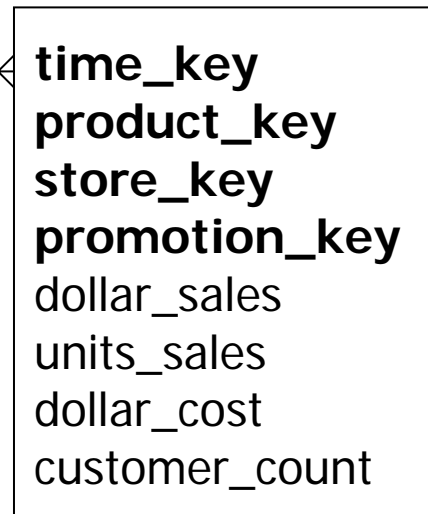
- **season** è la “stagione di vendita”
  - ad esempio, Natale, Pasqua, San Valentino, nessuna stagione, ...
    - è importante scegliere valori “concreti” (come “nessuna stagione”) anche per rappresentare valori apparentemente nulli
    - i valori nulli vanno evitati
  - **event**, simile a **season**, è associata a eventi speciali
    - ad esempio, finale del Super Bowl, Hurricane Hugo
  - altri attributi
- La dimensione tempo non comprende eventi promozionali
  - non dipendono solo dal calendario
  - sono gestiti mediante la dimensione promozione

# La dimensione tempo

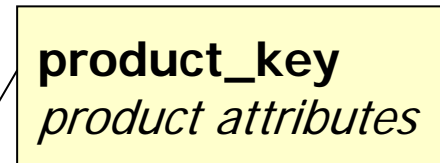
## Time Dimension



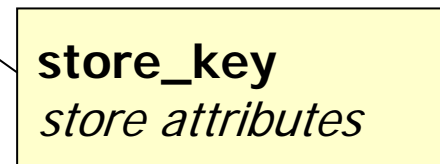
## Sales Fact



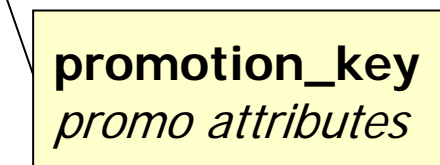
## Product Dimension



## Store Dimension



## Promotion Dimension



# La dimensione prodotto

- La dimensione prodotto descrive le unità di vendita (SKU) della catena di negozi
  - i dati per la dimensione prodotto sono solitamente estratti dal file principale dei prodotti usati per i sistemi POS
    - gestito dalla direzione e trasferito frequentemente dalla direzione ai POS
  - è responsabilità della direzione recepire i nuovi UPC e creare dei nuovi record nel file principale dei prodotti
    - ad ogni nuovo UPC deve essere assegnato un numero di SKU univoco
    - la direzione assegna anche i numeri di SKU per i prodotti “locali”
  - la tabella dimensione per i prodotti deve essere aggiornata in seguito a modifiche nel file dei prodotti

# Attributi dei prodotti

- Il file principale dei prodotti contiene molti attributi descrittivi per ciascuna SKU
  - ad esempio, la gerarchia delle merci (merchandise hierarchy)
    - le **SKU** si raggruppano (roll up) per dimensioni delle confezioni (**package\_size**)
    - le dimensioni delle confezioni si raggruppano in marche (**brand**)
    - le marche si raggruppano in sotto-categorie (**subcategory**)
    - le sottocategorie si raggruppano in categorie (**category**)
    - le categorie si raggruppano in reparti (**department**)

# Attributi dei prodotti

- Ad esempio
  - **SKU**: Green 3-pack Brawny Paper Towels, UPC #...
  - **package\_size**: 3-pack
  - **brand**: Brawny
  - **subcategory**: paper towels
  - **category**: paper
  - **department**: grocery

# Attributi dei prodotti

- Altri attributi non fanno parte della gerarchia delle merci
  - numero di SKU
  - tipo della confezione
  - prodotto dietetico
  - peso (numerico) e unità di misura del peso
  - colore
  - unità per confezione venduta, unità per confezione spedita
  - dimensioni (larghezza, altezza, profondità)
  - molti altri...
    - la dimensione prodotto ha solitamente 50 o più attributi, che possono essere utilmente usati nelle interrogazioni come criteri di selezione e/o di raggruppamento



# La dimensione negozio

- La dimensione negozio descrive i negozi della catena
  - i dati relativi ai negozi possono provenire da un foglio elettronico e/o da altre sorgenti informative
- La dimensione negozio è una dimensione essenzialmente geografica
  - ogni negozio occupa un punto nello spazio
  - i negozi possono essere raggruppati rispetto a ogni possibile geografia
    - ad esempio (negli Stati Uniti) per zip code, città, contea, stato
    - ma anche per distretto di vendita e regione di vendita (nozioni relative alla struttura organizzativa della catena)

# Attributi dei negozi

- nome, numero (codice nella catena), indirizzo, telefono, direttore, ...
- attributi geografici
  - zip code, città, contea, stato
  - distretto e regione di vendita
- informazioni su servizi supplementari
  - stampa foto, servizi finanziari, ...
- aree
  - area del negozio (in sqft), area del reparto surgelati, ...
- date
  - data prima apertura, ultima ristrutturazione, ...
    - rappresentati da date o da riferimenti a sinonimi della tabella dimensione tempo

# Nomi degli attributi

- I nomi degli attributi devono essere il più possibile descrittivi e non ambigui
  - ad esempio, negli schemi dimensionali sono solitamente presenti più dimensioni geografiche
    - come negozio, magazzino, cliente
    - ha senso di parlare della città in cui si trova il negozio o il magazzino, della città di residenza e di nascita del cliente
    - tali attributi (anche se in diverse tabelle)
      - non devono semplicemente chiamarsi **city**
      - ma devono chiamarsi **store\_city**, **warehouse\_city**, **customer\_home\_city**, **customer\_born\_city**
  - inoltre, tutti i termini usati negli schemi devono essere opportunamente descritti in un glossario

## Attributo o fatto?

- Campi come le aree dei negozi sono numerici e additivi (attraverso i negozi)
  - gli attributi sono solitamente descrittivi
- I dati sulle aree dei negozi devono essere rappresentati come fatti?
  - no, perché sono solitamente invarianti nel tempo
    - i fatti interessanti variano al variare delle dimensioni da cui dipendono
  - semmai, potrebbe essere utile introdurre degli ulteriori campi per categorizzare (ovvero, discretizzare) questi valori numerici
    - come piccolo, medio, grande, molto grande, oppure per fasce di aree

# La dimensione promozione

- La dimensione promozione descrive ogni possibile promozione che si applica alla vendita dei prodotti
  - ad esempio, riduzioni temporanee di prezzi, esposizione alla fine dei corridoi, pubblicità sui giornali, buoni sconto, ...
  - la dimensione promozione non descrive la promozione effettivamente applicata alla vendita
- La dimensione promozione è una dimensione causale (non casuale)
  - descrive fattori che sono la causa di potenziali cambiamenti (nelle abitudini dei clienti)
  - la dimensione promozione è la dimensione potenzialmente più interessante del nostro schema dimensionale

# Effetti delle promozioni

- Alcuni possibili effetti delle promozioni
  - aumenti della vendita dei prodotti in promozione
    - misurabili solo se sono noti i livelli base di vendita (senza la promozione)
      - i livelli base di vendita possono essere stimati dalle vendite precedenti e sulla base di modelli matematici sofisticati
  - diminuzione della vendita al termine della promozione
  - riduzione della vendita di altri prodotti
  - aumento complessivo della vendita, considerando il periodo della promozione e periodi immediatamente precedenti e/o successivi
  - profittabilità della promozione
    - tiene conto dei diversi aspetti

# Promozioni

- Le diverse modalità di promozione possono essere applicate contemporaneamente
  - ad esempio, riduzione temporanea del prezzo, pubblicità sui giornali e esposizione alla fine dei corridoi
  - ogni record della tabella dimensione delle promozioni descrive una possibile combinazione delle modalità di promozione
    - anche se in un anno ci possono essere 1.000 pubblicità sui giornali, 1.000 riduzioni temporanee dei prezzi e 200 esposizioni alla fine dei corridoi, le combinazioni effettive sono solitamente limitate (ad esempio, 5.000)

# Promozioni

- ciascuna particolare promozione può essere applicata diversamente nei diversi negozi
  - ad esempio, in alcuni negozi può essere impossibile effettuare le esposizioni alla fine dei corridoi
  - in questo caso, la promozione è rappresentata da due record
    - riduzione, pubblicità e esposizione
    - riduzione e pubblicità



# Attributi delle promozioni

- nome della promozione
- tipo della riduzione di prezzo
  - ad esempio, buono sconto, temporanea, nessuno
- tipo della pubblicità
  - ad esempio, giornale, radio, giornale e radio, posta
- media della pubblicità
- tipo dell'esposizione
- tipo del buono sconto
- costo della promozione
- date di inizio e fine della promozione
- altri attributi

## Una dimensione o più dimensioni?

- Le promozioni sono basate su quattro meccanismi causali
  - riduzione di prezzo, pubblicità, esposizione, buoni sconto
- La promozione è una sola dimensione
  - o deve essere rappresentata da quattro diverse dimensioni?
  - la decomposizione in quattro dimensioni è possibile
    - dipende dai requisiti e dalle esigenze di analisi dell'utente finale
    - se l'utente pensa separatamente (indipendentemente) a questi quattro meccanismi, allora è forse opportuno definire quattro diverse dimensioni

## Tabelle fatti senza fatti

- Lo schema dimensionale che è stato costruito è in grado di rispondere a molte interrogazioni
  - tuttavia, non è in grado di calcolare i prodotti in promozione che non sono stati venduti
    - più avanti sarà studiata una tecnica (tabelle fatti senza fatti) per poter gestire anche questo tipo di informazioni

## Additività dei fatti

- Lo schema dimensionale della catena di negozi memorizza i seguenti fatti relativi alle vendite
  - incasso totale in dollari (dollar\_sales), numero totale di unità vendute (units\_sales), costo totale in dollari (dollar\_cost), numero di clienti (customer\_count)
    - i primi tre fatti sono additivi rispetto a tutte le dimensioni

## Fatti calcolati e additività

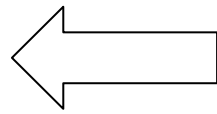
- Il profitto lordo (per unità di vendita, giorno e negozio) può essere calcolato sottraendo il costo totale dall'incasso totale
  - anche questo fatto, calcolato, è additivo rispetto a tutte le dimensioni
- Il margine lordo è calcolato dividendo il profitto lordo per l'incasso totale
  - per ogni possibile aggregazione, il margine lordo può essere calcolato prima sommando tutti gli incassi e i costi e poi dividendo
    - alcuni fatti non additivi (calcolati da fatti additivi) possono essere aggregati ricordandosi di distribuire correttamente le operazioni

# Fatti non additivi

- Il numero di clienti è un fatto semi-additivo
  - non è additivo rispetto alla dimensione prodotto
    - se un prodotto A è stato acquistato da 20 clienti e un prodotto B da 30 clienti, quanti clienti hanno comprato A o B?
  - tuttavia, è additivo rispetto alle altre dimensioni
- I conteggi sono solitamente fatti semi-additivi
  - possono essere sommati correttamente restringendo le chiavi nelle dimensioni in cui non sono additivi a valori singoli

## Fatti non additivi

- Se la promozione indicasse la combinazione di promozioni effettivamente applicata alla vendita, allora il numero di clienti non sarebbe completamente additivo rispetto alle promozioni
  - perché un cliente potrebbe comprare, in una stessa transazione, una unità di vendita con un buono sconto e la stessa unità di vendita senza buono sconto
    - se questa situazione è considerata infrequente, può essere trascurata
    - se invece è frequente e vuole essere analizzata, la dimensione “buono sconto” deve essere scorporata dalla dimensione promozione





# Studi di caso

- Vendite
- **Inventario**
- Catena del valore

# Modelli di inventario

- Una catena di magazzini di cui si vogliono analizzare i livelli di inventario dei prodotti
  - nel caso della catena di negozi alimentari sono stati rappresentati flussi di prodotti
    - sono stati misurati i prodotti effettivamente venduti
    - i flussi sono solitamente additivi (perché una volta “usciti” non possono essere contati nuovamente)
  - nel caso dei magazzini è interessante rappresentare i livelli di inventario dei prodotti
    - sono possibili diversi modelli di rappresentazione dei livelli di inventario

# Livelli di inventario e additività

- I livelli di inventario rappresentano delle istantanee (snapshot) di livelli
  - ad esempio, la disponibilità di un prodotto nel magazzino
  - hanno natura simile a saldi e bilanci economici, e a misure di intensità come la temperatura
- Caratteristiche dei livelli di inventario
  - non sono additivi rispetto al tempo
    - ma sono additivi rispetto ad altre dimensioni
  - guardando solo ai livelli di inventario in due istanti di tempo non è possibile determinare l'effettivo flusso tra i due istanti

## Livelli di inventario e semi additività

- I livelli di inventario possono essere aggregati (nel tempo) rispetto ad alcune operazioni diverse dalla somma
  - ad esempio, media e massimo
  - le medie però devono essere effettuate rispetto ai periodi di tempo
    - la funzione aggregativa AVG di SQL potrebbe aggregare in modo non corretto
    - ad esempio, il livello totale giornaliero medio di un prodotto in un'area geografica che contiene 4 magazzini in una settimana (ovvero, la media giornaliera del livello totale del prodotto nei magazzini) può essere calcolata
      - prima sommando i  $4 \cdot 7 = 28$  dati
      - e poi dividendo per 7 (e non per 28)

# Modelli di inventario

- Esistono tre diversi modelli di inventario
  - modello ad **istantanee (inventory snapshot)**
    - i livelli di inventario sono misurati periodicamente (ad esempio, giornalmente)
      - un record per prodotto, magazzino, unità di tempo
  - modello per **stato delle consegne (delivery status)**
    - viene gestito lo stato di ciascuna consegna di prodotto derivante da un ordine
      - un record per prodotto, magazzino, ordine
      - il record viene aggiornato a fronte di consegne in ingresso e uscita del prodotto dal magazzino
    - possibile solo se prodotti identici relativi a ordini diversi sono distinguibili

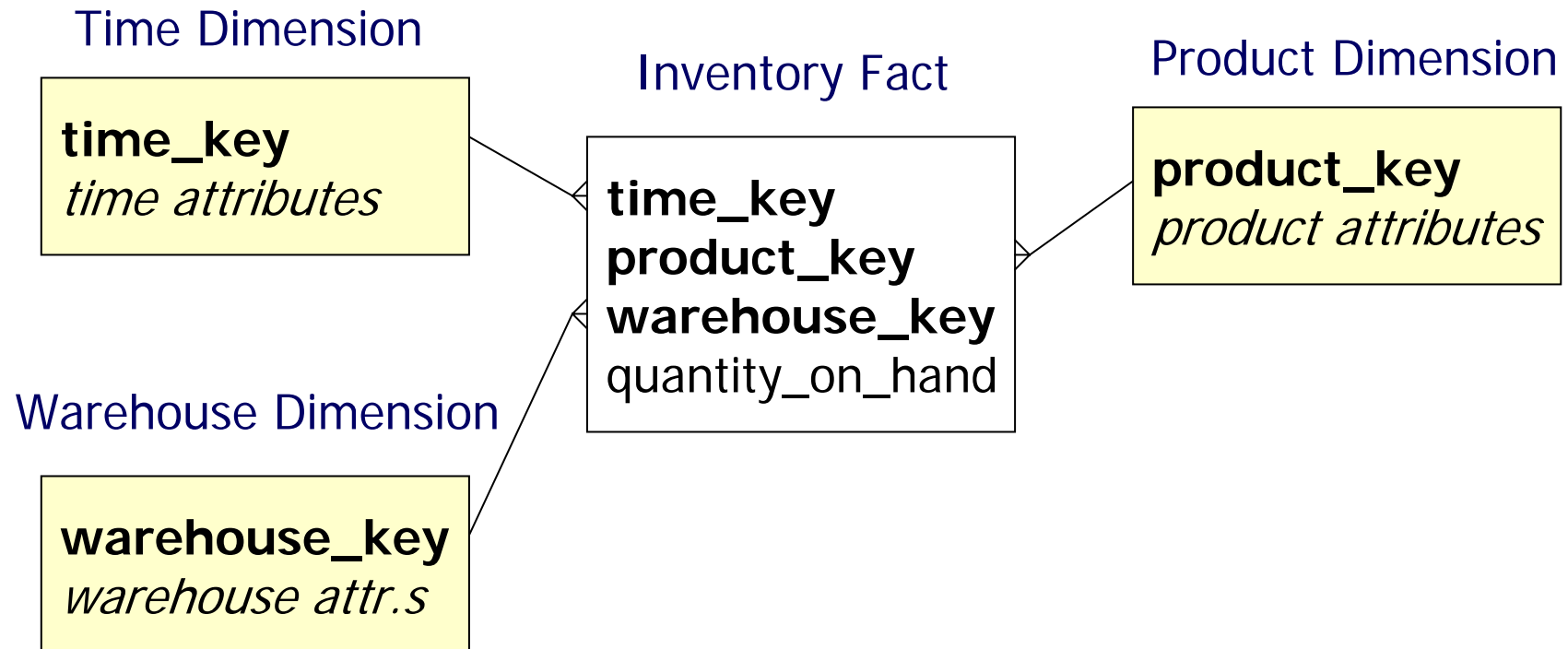
# Modelli di inventario

- modello a **transazioni (transaction)**
  - vengono misurate tutte le modifiche dei livelli
    - un record per prodotto, magazzino, transazione
- Ciascuno dei tre modelli supporta diverse modalità di analisi
  - in pratica, il data mart di un processo di inventario può utilizzare contemporaneamente anche due o tutti e tre i modelli di inventario, mediante uno schema dimensionale per ciascun modello utilizzato

# Il modello inventory snapshot

- Il modello di inventario ad istantanee prevede solitamente tre dimensioni primarie
  - tempo, prodotto e magazzino
    - in casi più generali, i magazzini sono associati ai negozi e/o ai clienti, e bisogna rappresentare anche tali dimensioni
    - gli inventari non sono correlati con le promozioni (né è solitamente possibile farlo)
    - una possibile ulteriore dimensione è il fornitore
      - se è possibile distinguere tra prodotti identici consegnati da fornitori diversi
- Prevede un singolo fatto misurabile
  - la quantità disponibile (quantity\_on\_hand)

# Schema dimensionale per inventory snapshot





## Un possibile uso di inventory snapshot

- Se ci sono molti prodotti e magazzini e i livelli di inventario sono misurati frequentemente (giornalmente) è possibile applicare questo modello a dei casi concreti interessanti
  - ad esempio, i magazzini possono essere negozi, e il livello di inventario può essere il livello del prodotto sullo scaffale (in un certo istante, ad esempio alla chiusura)

# Caratteristiche di inventory snapshot

- Rispetto al data mart delle vendite, il data mart dell'inventario ad istantanee è **denso** (non è sparso)
  - per ogni giorno, magazzino e prodotto c'è un fatto (record) da misurare
  - nel caso di una catena di negozi con 100.000 prodotti in 2.000 negozi ci sono  $100.000 \cdot 2.000 \cdot 365 = 73.000.000.000$  record per anno
    - con un record di 14 byte richiede oltre 1TB di spazio primario (per ciascun anno accumulato)
  - spesso è necessario un compromesso
    - es.: dati giornalieri per l'ultimo mese, settimanali per gli undici mesi precedenti, mensili per anni precedenti
      - per avere tre anni di dati servono così circa 100 snapshot anziché 1.100

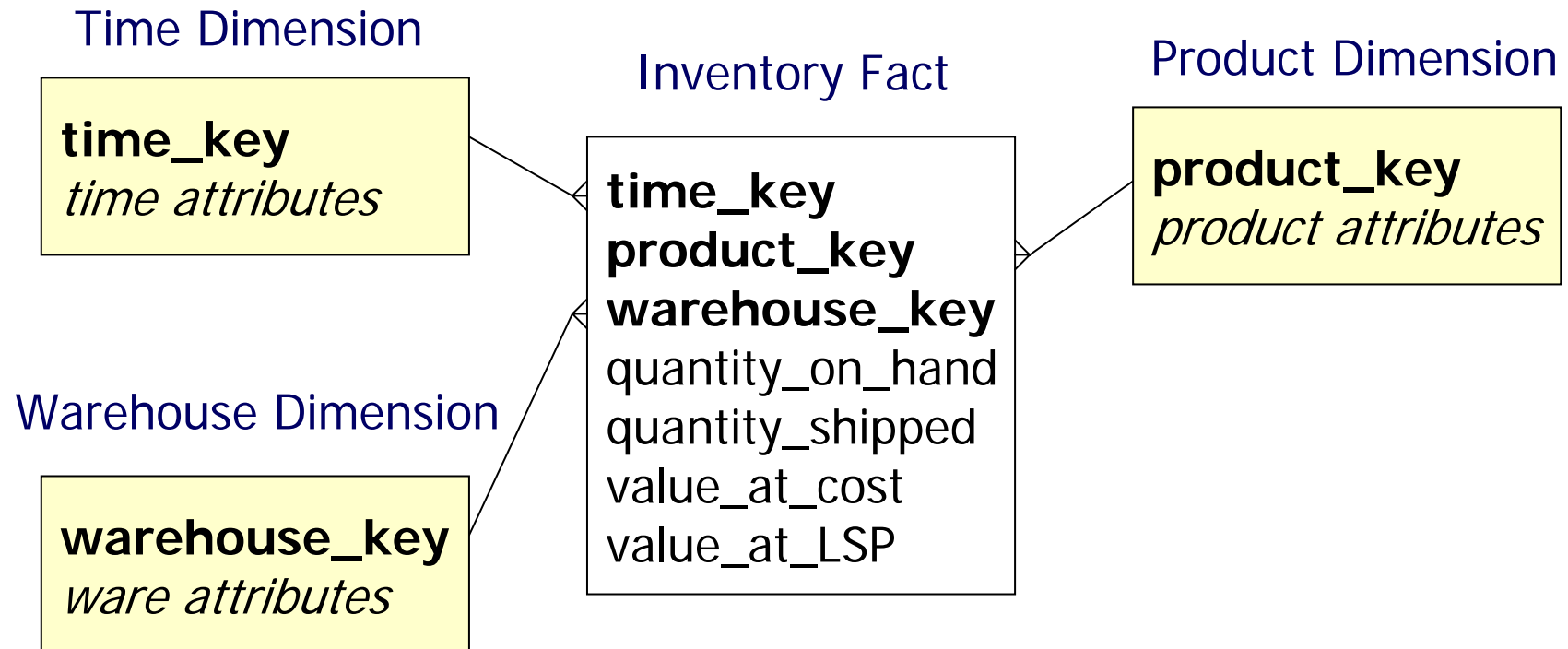
## Limiti di inventory snapshot

- Il modello di inventario ad istantanee è basato sulla sequenza temporale di livelli di inventario dei prodotti individuali
  - non permette di calcolare alcune metriche di processo interessanti, come
    - velocità di rotazione
    - giorni di approvvigionamento
    - margine lordo di ritorno sull'inventario (**GMROI**)
  - GMROI (Gross Margin Return on Inventory Investment) è una metrica standard adottata dagli analisti degli inventari per giudicare la qualità dell'investimento in giacenze
    - Intuitivamente, è pari al profitto (cioè incasso meno costo) nell'unità di tempo diviso per l'immobilizzo

# Advanced inventory snapshot

- Una variante dell'inventario ad istantanee (che permette di calcolare le misure descritte precedentemente) è il modello di **inventario ad istantanee avanzato (advanced inventory snapshot)**
  - che si ottiene rappresentando anche i seguenti fatti
    - quantità spedita (**quantity\_shipped**) o consumata o venduta
    - valore del costo (**value\_at\_cost**) (per unità di prodotto)
    - valore della vendita (**value\_at\_LSP**) (all'ultimo prezzo di vendita)

# Advanced inventory snapshot



# Analisi con advanced inventory snapshot

- numero di rotazioni giornaliere
  - **quantity\_shipped / quantity\_on\_hand**
- numero medio di rotazioni giornaliere
  - somma di **quantity\_shipped** / media giornaliera di **quantity\_on\_hand**
- numero medio di giorni di approvvigionamento
  - valore finale di **quantity\_on\_hand** / media di **quantity\_shipped**
- profitto lordo (per unità di prodotto)
  - **value\_at\_LSP – value\_at\_cost**
- GMROI
  - (somma di **quantity\_shipped**) \* (**value\_at\_LSP – value\_at\_cost**) diviso (media giornaliera di **quantity\_on\_hand**) \* (**value\_at\_cost**)

# Additività

- Malgrado la quantità disponibile sia semi additiva, la quantità spedita, il valore del costo e il valore della vendita sono fatti additivi
  - il GMROI non è additivo
    - il GMROI è una misura che deve essere calcolata, mentre non è utile memorizzarla (materializzarla) nella tabella fatti
  - il profitto lordo può essere definito come attributo calcolato mediante la definizione di una vista sulla tabella fatti
    - i calcoli intra-record possono essere calcolati in modo molto efficiente

## Il modello delivery status

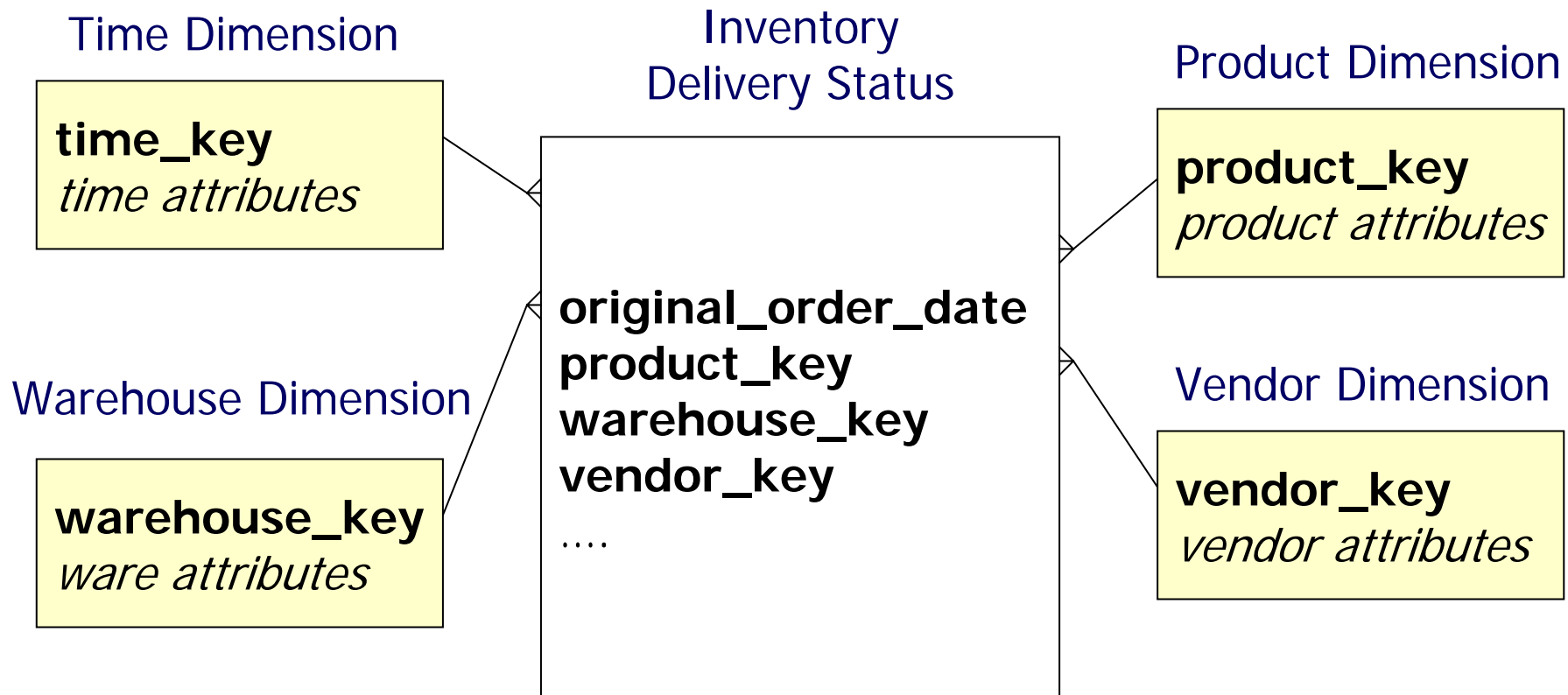
- Nel modello di inventario per stato delle consegne viene memorizzato un record per ciascun ordine di acquisto di prodotto di un magazzino
  - ogni record nella tabella fatti corrisponde a una voce su un ordine di acquisto
  - è utile quando ciascuna consegna è relativa a una quantità grande di un prodotto, che viene via via consumata dal magazzino
    - in questo caso, ha senso mantenere traccia di una serie eventi di ben definiti, dalla consegna all'esaurimento della merce
    - non è appropriato se i prodotti arrivano con un flusso continuo e in diverse consegne (prima di esaurirsi)



## Attività nel magazzino

- In un magazzino rappresentato con il modello per stato delle consegne, le merci attraversano tipicamente le seguenti fasi
  - sequenza normale di fasi (in ordine)
    - ricezione, ispezione, collocamento nel magazzino, autorizzazione alla vendita, ritiro dal magazzino, imballaggio, consegna
  - fasi eccezionali
    - ispezione fallita, restituzione al fornitore, danneggiamento, perdita, restituzione dal cliente, restituzione al magazzino, cancellazione, rimborso
- La tabella fatti deve memorizzare informazioni aggiornate sullo stato dei prodotti
  - potrebbe essere anche aggiornata più volte nello stesso giorno

# Schema dimensionale per delivery status



## Due caratteristiche “nuove”

- Dimensioni degeneri
- Date ausiliarie

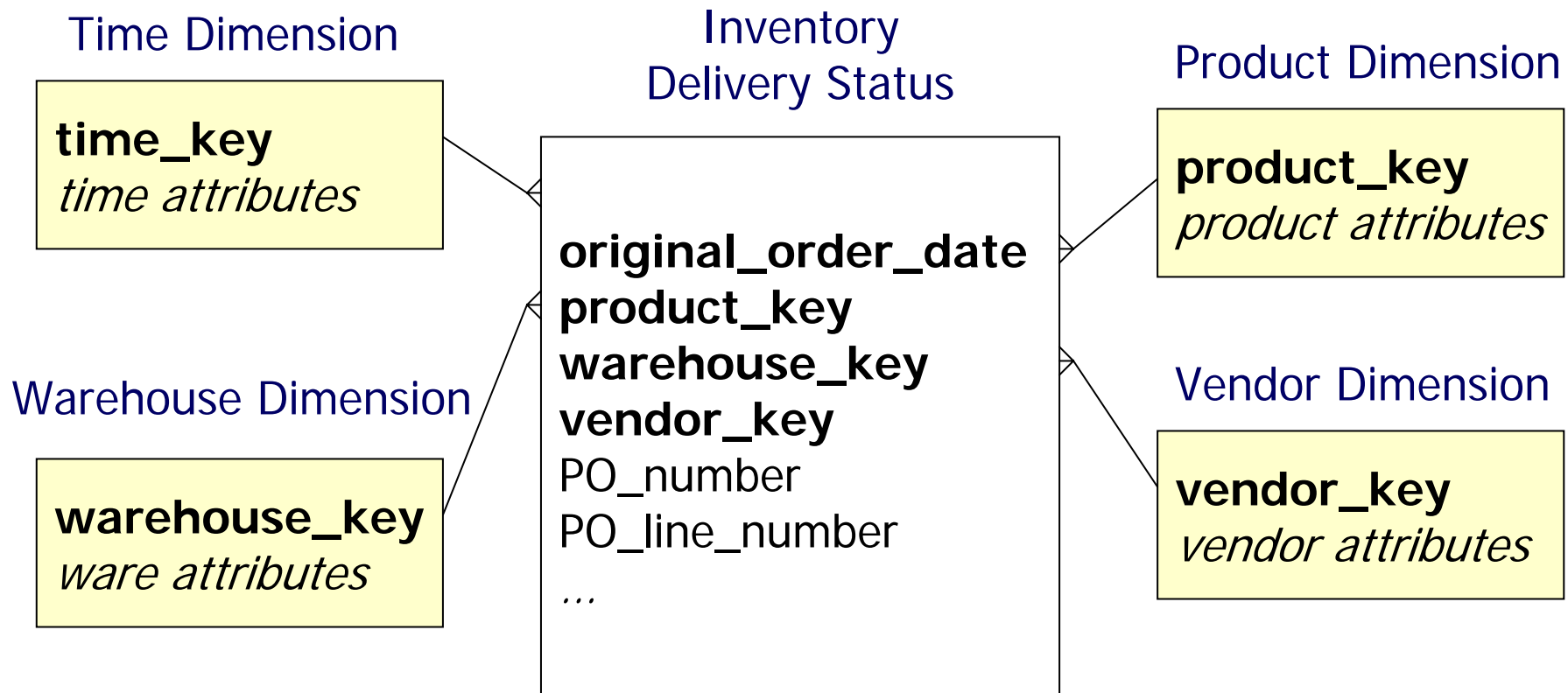
## Una dimensione particolare

- Ci interessa rappresentare
  - ordine e linea d'ordine ?
- Sì
  - per identificare i singoli fatti
  - anche per raggruppare i fatti relativi ad un ordine
- Si tratta quindi di una dimensione, ma con quali proprietà?

# Dimensioni degeneri

- Il numero dell'ordine di acquisto sarebbe la chiave in una tabella degli ordini
  - che memorizza informazioni complessive circa i singoli ordini, come il venditore, la data dell'ordine e il magazzino di destinazione
  - queste informazioni sono già rappresentate nello schema dimensionale da altre dimensioni e/o fatti
- Possiamo introdurre una “dimensione senza tabella dimensionale”
  - **dimensione degenera**
- È utile rappresentare dati come dimensioni degeneri
  - quando la loro grana corrisponde a quella della tabella fatti
  - e la loro utilità si limita al poter raggruppare direttamente i fatti
    - ad esempio, per ordine d'acquisto

# Schema dimensionale per delivery status



## Dimensioni degeneri

- I campi numero dell'ordine di acquisto (**PO\_number**) e numero di linea nell'ordine di acquisto (**PO\_line\_number**) formano, insieme, una dimensione
  - tuttavia, lo schema dimensionale non contiene nessuna “dimensione voce di ordine”
    - una dimensione la cui chiave è presente nella tabella fatti ma non è rappresentata da una tabella dimensione è chiamata una **dimensione degenera**

## Campi data ausiliari

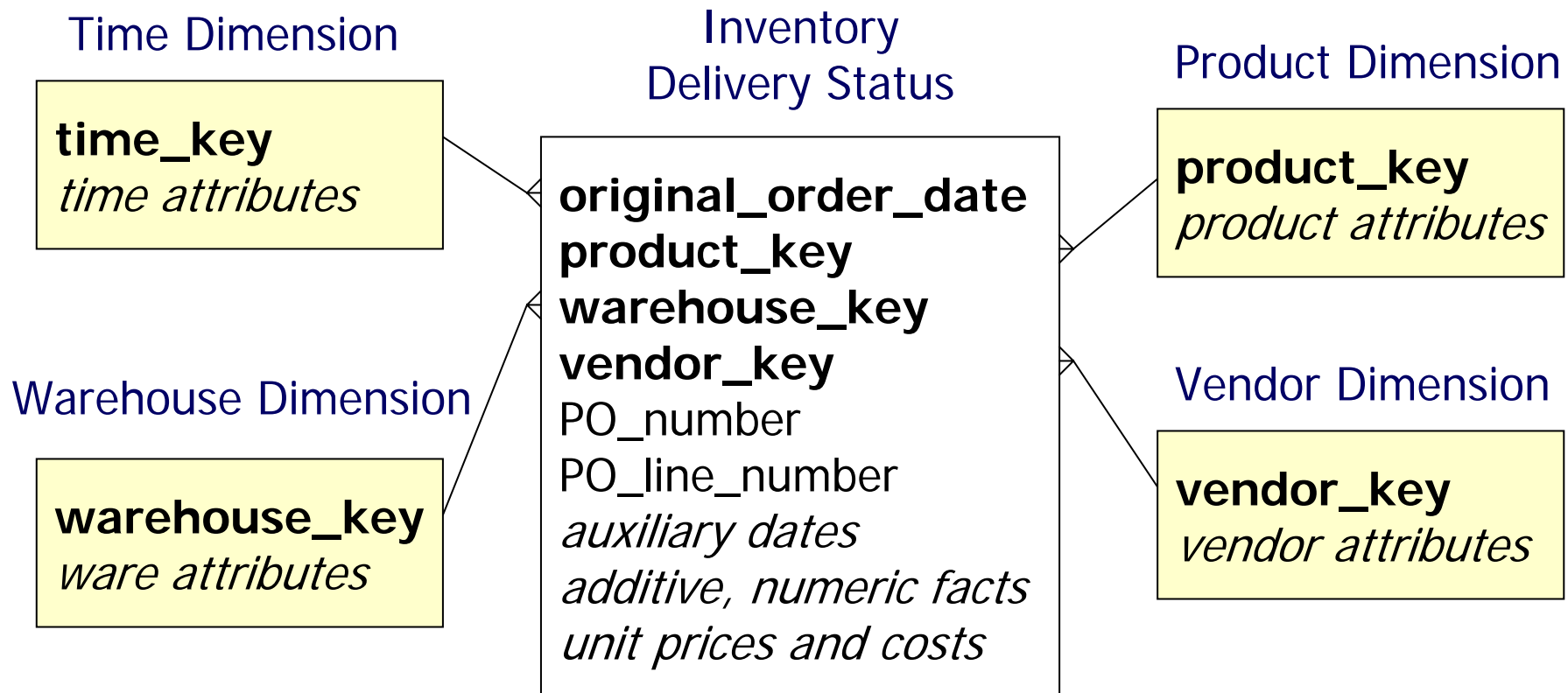
- Se una tabella fatti viene usata per rappresentare la traccia dello stato di una entità, allora possono essere utili diversi valori di data
  - per gli istanti in cui avvengono eventi interessanti
  - c'è solitamente una data primaria
    - la data dell'ordine (**original\_order\_date**)
  - altre date (ad esempio, della prima e dell'ultima consegna) sono usate per differenza per misurare la durata delle attività nel magazzino
    - dalla differenza di due date si ottiene un numero (di giorni) di cui può essere calcolata la media rispetto a tutte le dimensioni
    - queste differenze sono misure calcolate che possono essere definite da una vista



## Campi data ausiliari

- Campi data ausiliari (oltre a **original\_order\_date**)
  - **first\_received\_date**
  - **last\_received\_date**
  - **first\_inspect\_date**
  - **first\_auth\_to\_sell\_date**
  - **first\_shipment\_date**
  - **last\_shipment\_date**
  - **last\_return\_date**

# Schema dimensionale per delivery status



# Fatti

- Fatti (numerici e additivi) nella tabella fatti
  - **qty\_received**
  - **qty\_inspected**
  - **qty\_returned\_to\_vend**
  - **qty\_placed\_in\_inv**
  - **qty\_auth\_to\_sell**
  - **qty\_picked**
  - **qty\_boxed**
  - **qty\_shipped**
  - **qty\_returned\_by\_cust**
  - **qty\_returned\_to\_inv**
  - **qty\_damaged**
  - **qty\_lost**
  - **qty\_written\_off**

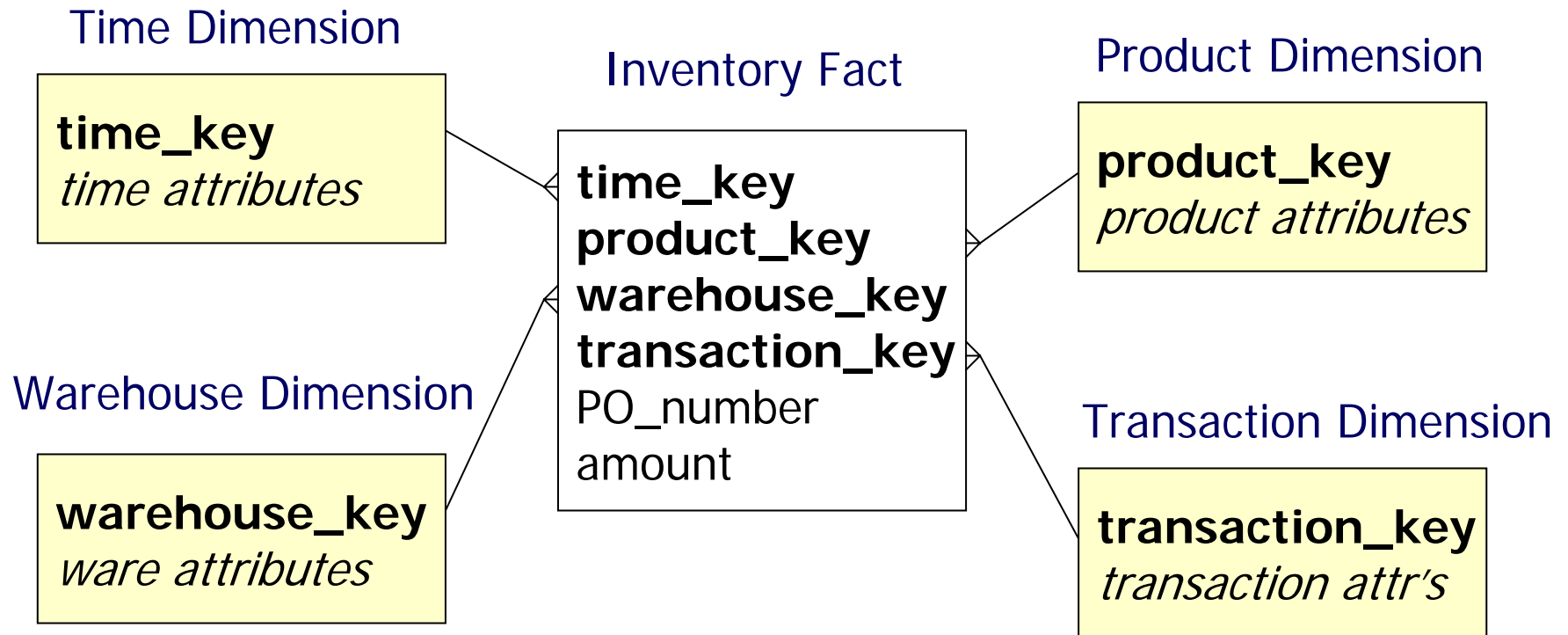
# Prezzi e costi unitari

- Altri quattro fatti sono relativi a prezzi e costi per unità di prodotto
  - **unit\_cost**, **orig\_selling\_price**, **last\_selling\_price**, **avg\_selling\_price**
- Prezzi e costi unitari non sono additivi
  - tuttavia, è conveniente memorizzare questi 4 dati come dati unitari, perché è possibile combinarli con i 13 fatti numerici in ogni modo
    - sarebbero altrimenti necessari 52 campi
  - i totali di interesse possono essere definiti come fatti calcolati mediante una vista
- In generale, fatti non additivi possono essere preferiti a fatti additivi se questi fatti additivi possono poi essere calcolati mediante calcoli intra-record

# Il modello transaction

- Nel modello di inventario per transazioni viene memorizzato un record per ciascuna transazione che modifica lo stato dell'inventario
  - ogni record nella tabella fatti corrisponde a una transazione in un magazzino
  - le transazioni possibili comprendono
    - ricezione (della voce di un linea d'ordine), collocamento per l'ispezione, rilascio da ispezione, collocamento nel magazzino, autorizzazione alla vendita, imballaggio, consegna, ...
    - ispezione fallita (con ragione), restituzione al fornitore (con ragione), danneggiamento, perdita, restituzione dal cliente (con ragione), ...
  - ciascuna transazione è relativa a una certa quantità di prodotto

# Schema dimensionale per transaction



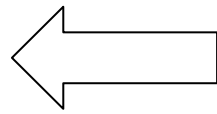
# Schema dimensionale per transaction

- La dimensione transazione ha un record per ciascun possibile tipo di transazione (con ogni possibile “ragione”, per le transazioni con ragione)
  - il numero di tipologie di transazioni possibili (con ragioni) è comunque piccolo (centinaia)
- Il singolo fatto quantità (**amount**) è tipico delle tabelle fatti con grana delle transazioni individuali
  - lo scopo di ogni transazione è tipicamente quello di muovere una quantità di qualcosa
  - in generale, le tabelle fatti con questa grana
    - hanno un singolo fatto, quantità
    - il contesto della transazione è rappresentato da dimensioni (e non da fatti)
    - possono essere presenti dimensioni degeneri

## Uso del modello transaction

- Nel modello di inventario per transazioni viene rappresentato il massimo livello di dettaglio possibile per un inventario
  - è però difficile da usare direttamente per fini di analisi
    - ad esempio, per conoscere i livelli di inventario in una certa data è necessario conoscere i livelli di inventario in una data iniziale ed elaborare tutti i record relativi a transazioni dalla data iniziale alla data di interesse
  - per questo motivo, il modello per transazioni è spesso accompagnato da una rappresentazione dell'inventario basata su qualche modello ad istantanee
    - il data mart dell'inventario è in questo caso composto da più schemi dimensionali





# Studi di caso

- Vendite
- Inventario
- **Catena del valore**

# Catena del valore

- Sono stati finora studiati separatamente alcuni processi che possono essere inquadrati in un contesto più ampio
  - il processo che segue il progresso del prodotto
    - dalla produzione alla vendita
    - attraverso delle fasi intermedie che formano la **catena del valore** del prodotto
  - è importante comprendere come le diverse fasi intermedie contribuiscano individualmente al valore complessivo del prodotto
- Due tipologie di catene del valore
  - dal lato della domanda
    - dal magazzino alla vendita
  - dal lato della produzione
    - dal materiale grezzo al magazzino

# Catena del valore della domanda

- Uno scenario tipico legato alla domanda dei prodotti è rappresentato da sei schemi dimensionali, ordinati dal punto in cui il prodotto ha origine al punto in cui viene venduto all'utente finale
  - magazzino dei prodotti finiti (inventario)
  - spedizione al centro di distribuzione (flusso)
  - magazzino del centro di distribuzione (inventario)
  - spedizione ai negozi di vendita (flusso)
  - magazzino dei negozi di vendita (inventario)
  - vendita al dettaglio (flusso)
- Il prodotto si muove sequenzialmente attraverso la sua catena del valore, attraverso fasi che sono alternativamente di inventario e di flusso

# Dimensionalità delle fasi nella domanda

- Magazzino dei prodotti finiti (inventario)
  - tempo, prodotto (SKU), magazzino
- Spedizione al centro di distribuzione (flusso)
  - tempo, prodotto (SKU), magazzino, centro di distribuzione, contratto (o accordo commerciale o promozione), modalità di consegna (compreso il vettore)
- Magazzino del centro di distribuzione (inventario)
  - tempo, prodotto (SKU), centro di distribuzione

# Dimensionalità delle fasi nella domanda

- Spedizione ai negozi di vendita (flusso)
  - tempo, prodotto (SKU), centro di distribuzione, negozio, contratto (o accordo commerciale o promozione), modalità di consegna (compreso il vettore)
- Magazzino dei negozi di vendita (inventario)
  - tempo, prodotto (SKU), negozio
- Vendita al dettaglio (flusso)
  - tempo, prodotto (SKU), negozio, promozione, cliente (se disponibile)

# Drill across

- Gli schemi dimensionali condividono alcune dimensioni
  - ad esempio, la dimensione tempo è presente in tutti gli schemi, la dimensione negozio è presente negli ultimi tre schemi
- Se dimensioni che hanno lo stesso nome in più schemi dimensionali hanno anche lo stesso significato (intensionale ed estensionale) allora ha senso effettuare interrogazioni trasversali (**drill across**) tra i diversi schemi
  - che hanno lo scopo di comprendere il valore aggiunto da ciascuna fase nella catena del valore
    - ad esempio, per confrontare le vendite medie giornaliere dei prodotti con i livelli medi in inventario
  - i diversi schemi dimensionali formano un data mart

# Dimensioni conformi

- Una **dimensione conforme** (o conformata, **conformed dimension**) è una dimensione che ha esattamente lo stesso significato in più schemi dimensionali
  - rispetto a ogni possibile tabella fatti con cui può essere correlata mediante una operazione di join
  - ad esempio, una dimensione è conforme se può essere rappresentata da tabelle identiche in schemi dimensionali diversi
- Un insieme di schemi dimensionali forma un data mart (e un insieme di data mart forma un data warehouse) se è stato costruito attorno a un insieme coerente e coordinato di dimensioni conformi
  - in questo caso, infatti, i dati dei diversi schemi dimensionali e data mart possono essere correlati in modo utile



## Dimensioni con dettaglio ridotto

- Si consideri la seguente situazione
  - il magazzino dei prodotti finiti conosce alcune informazioni circa i prodotti (ad esempio, il lotto di produzione) che non sono note (o non possono essere raccolte) nella vendita nei negozi
  - in questo caso, la dimensione prodotto nel magazzino dei prodotti finiti può essere diversa da quella della vendita nei negozi
    - in particolare, può essere basata su una grana più fine
  - possono esistere diverse versioni della stessa dimensione
    - purché opportunamente costruite mediante operazioni di aggregazione (e, quindi, conformate)

## Dimensioni con dettaglio ridotto

- Se sono presenti diverse versioni di una dimensione, a diversi livelli di dettaglio
  - sono possibili operazioni di drill across basate solo su attributi che esistono in tutte le versioni della dimensione (e hanno lo stesso significato)
    - “basate” significa che gli attributi usati per le selezioni e i raggruppamenti sono presenti in tutte le versioni interessate della dimensione
  - evidentemente, infatti, non è possibile fare analisi per il processo di vendita al livello di dettaglio del lotto di produzione
    - una interrogazione di questo tipo potrebbe essere “compilabile” ma fornire un risultato non corretto

# Dimensioni derivate

- La possibilità di costruire tabelle fatti derivate mediante aggregazioni è fondamentale nelle applicazioni di data warehousing
  - nel caso di dimensioni con dettaglio ridotto, può essere utile costruire dimensioni conformate come tabelle derivate mediante aggregazioni
    - le dimensioni derivate devono contenere solo gli attributi significativi alla grana usata per l'aggregazione

# Catena del valore della produzione

- Il processo di produzione riguarda l'acquisizione di parti (e materie grezze) e il loro montaggio in prodotti finiti
  - la catena del valore della produzione è profondamente diversa da quella della domanda
    - sia per quanto riguarda le analisi di interesse
    - sia per quanto riguarda i dati
  - ad esempio, la nozione di prodotto non esiste in tutte le fasi
  - inoltre, esiste una relazione multi-a-molti tra parti e prodotti, che non è solitamente possibile rappresentare in modo diretto

# Schemi dimensionali nella produzione

- Ordinazione materiali
  - tempo, ingrediente (o parte), fornitore, accordo commerciale (o contratto)
- Consegna materiali
  - tempo, ingrediente (o parte), fornitore, stabilimento, modalità di consegna (compreso il vettore), accordo commerciale (o contratto)
- Magazzino materiali
  - tempo, ingrediente (o parte), stabilimento
- Monitoraggio dei processi produttivi
  - tempo, ingrediente (o parte), processo, stabilimento

# Schemi dimensionali nella produzione

- Montaggio (bill of materials)
  - tempo, ingrediente (o parte), prodotto (SKU)
- Inventario prodotti finiti
  - tempo, prodotto (SKU), magazzino
- Programmazione della produzione
  - tempo, prodotto (SKU)