# Il modello dimensionale

Luca Cabibbo

aprile 2012



## Il modello dimensionale

L'organizzazione dei dati del data warehouse costituisce la pietra angolare dell'intero sistema DW/BI

- le applicazioni BI, di supporto alle decisioni, accedono i dati direttamente dal DW
  - l'organizzazione dei dati nel DW deve sostenere sia la comprensibilità dei dati di business che buone prestazioni
- il sottosistema ETL è dedicato a trasformare i dati estratti dalle sorgenti informative in una forma adeguata al caricamento nel DW
- la modellazione dimensionale si è dimostrata la tecnica più valida per l'organizzazione dei dati nel DW

# Modellazione dimensionale, in breve

Organizzazione dei dati in un DW

- i dati di interesse per ciascun processo di business dell'organizzazione sono rappresentati mediante un *modello dimensionale* – anche chiamato *data mart*
  - in un data mart, i dati di interesse sono *misure* (di solito numeriche) del processo (*fatti*) – ciascun fatto è caratterizzato da un *contesto* (prevalentemente testuale) vero nel momento in cui il fatto è stato catturato (*dimensioni*)



# Modellazione dimensionale, in breve

Organizzazione dei dati in un DW

- inoltre, è necessario che i dati nei diversi data mart/modelli dimensionali siano organizzati in modo conforme (coerente) – per questo viene adottata un'*architettura a bus del data warehouse*
  - nell'ambito dei diversi processi di business, i dati sono organizzati attorno a un insieme di *dimensioni conformi* e di *fatti conformi*
  - questa conformità serve ad assicurare che i dati nei diversi data mart, relativi a processi diversi, possano essere correlati e integrati nell'ambito dell'intero data warehouse e nel corso del tempo

## Processi di business

Un **processo di business** (o **processo aziendale**, **business process**) è un evento o attività operativa principale per l'organizzazione

- ad esempio, il processo di gestione degli ordini
- di solito supportato da un sistema sorgente – da cui è possibile collezionare le misure di prestazioni associate
- pertanto, è possibile pensare a ciascun processo di business come a una sorgente di dati principale per il DW

I processi di business sono importanti nella modellazione dimensionale

- i processi di business possono essere usati come criterio per il raggruppamento coerente delle risorse informative dell'organizzazione e dei dati del data warehouse
- un data warehouse viene realizzato mediante uno o più data mart per ciascun processo di business di interesse



## Modelli dimensionali/data mart

Un **modello dimensionale per un processo di business** – anche chiamato semplicemente **modello dimensionale** o **data mart** – rappresenta nel DW i dati di interesse relativi ad uno specifico processo di business

- detto in altri modi
  - un data mart è un sottoinsieme logico dell'intero data warehouse
  - un data mart è la restrizione del data warehouse a un singolo processo di business
  - il data warehouse è l'unione dei data mart che lo costituiscono

## Data warehouse dimensionali

Un **data warehouse dimensionale** è un data warehouse realizzato come unione di un insieme di modelli dimensionali (data mart) che hanno le seguenti caratteristiche

- ogni data mart, relativo a un processo di business, è un insieme di schemi dimensionali
  - ogni *schema dimensionale* è organizzato in termini di misure (*fatti*) e contesto (*dimensioni*)
- viene adottata un'**architettura a bus del data warehouse** – **data warehouse bus architecture** – ovvero, i vari data mart sono costruiti usando
  - dimensioni conformi (o conformate)
    - ciascuna dimensione ha lo stesso significato in ciascuno schema dimensionale e data mart
  - fatti conformi
    - anche i fatti hanno un'interpretazione uniforme



## Dati multidimensionali

L'organizzazione logica dei dati in un modello dimensionale è descritta in termini di fatti e dimensioni

- secondo la prospettiva degli utenti di business del sistema DW/BI – e non secondo i modelli (ad es., quello relazionale) adottati nei sistemi che gestiscono le sorgenti informative
- un **fatto** rappresenta una misurazione delle prestazioni di un processo di business
  - una misura, di solito numerica e additiva, del processo da analizzare
- una **dimensione** rappresenta una prospettiva rispetto alla quale effettuare l'analisi
  - una dimensione rappresenta un'informazione del contesto in cui è stata catturata una misurazione delle prestazioni del processo di business di interesse
  - le dimensioni descrivono domini discreti, solitamente organizzati in livelli di aggregazione

# Dati multidimensionali – esempi

Modello dimensionale per il processo delle vendite

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

Modello dimensionale per il processo telefonate

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

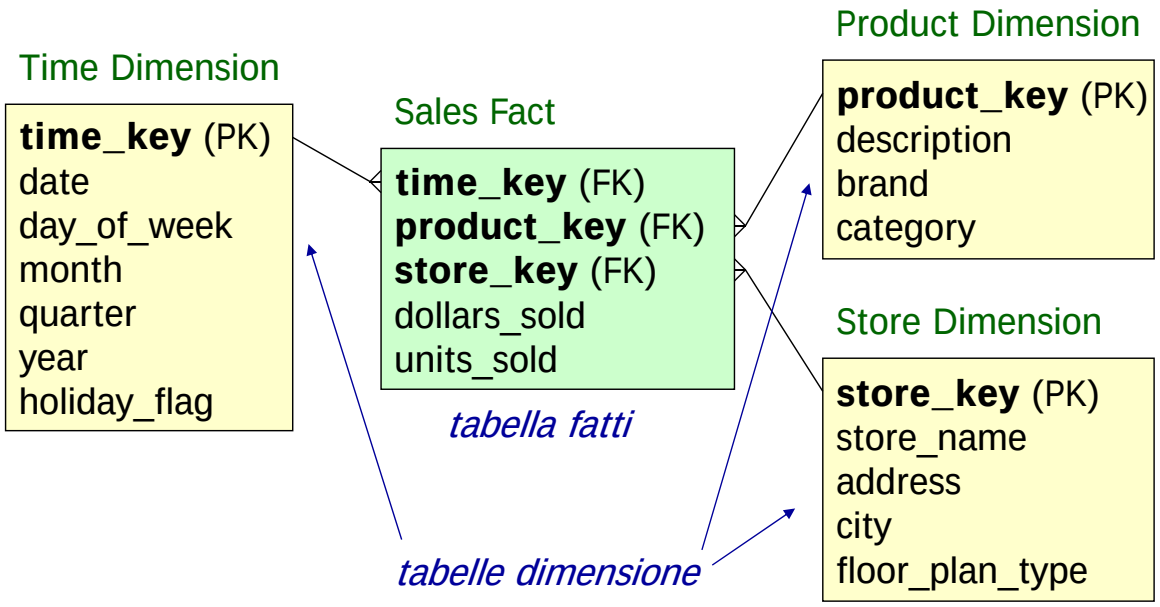


# Schemi dimensionali

Uno **schema dimensionale** – chiamato anche **star schema** o **schema a stella** – è una base di dati relazionale, usata per rappresentare dati multidimensionali – uno schema dimensionale è di solito composto da

- una tabella principale, chiamata **tabella fatti**
  - la tabella fatti memorizza i fatti misurabili di un processo
    - i fatti più comuni sono numerici, continui e additivi
- alcune tabelle ausiliarie, chiamate **tabelle dimensione**
  - ciascuna tabella dimensione fornisce un contesto ai fatti della tabella fatti – inoltre, rappresenta una dimensione rispetto alla quale è interessante analizzare i fatti
    - memorizza i membri della dimensione, che caratterizzano la granularità dei fatti, nonché gli attributi (solitamente testuali, discreti e descrittivi)

# Esempio di schema dimensionale




- questo schema modella 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



# Scopo di uno schema dimensionale

In uno schema dimensionale

- una *tabella dimensione* serve a rappresentare un insieme di elementi (un insieme in senso matematico), chiamati *membri della dimensione*
- una *tabella fatti* serve a memorizzare un insieme di funzioni numeriche (funzioni in senso matematico)

Nell'esempio, lo schema rappresenta

- una dimensione **Product** di tipi di prodotti in vendita
- una dimensione **Time** di giorni in un intervallo di interesse
- una dimensione **Store** dei negozi di una catena di negozi
- una funzione **dollars_sold**: **Product ×Time × Store → R**
- una funzione **units_sold**: **Product × Time × Store → N**

# Tabelle dimensione

Una **tabella dimensione** (**dimension table**) memorizza una dimensione rispetto a cui è di interesse analizzare un processo

- una **dimensione** è un dominio (insieme) di elementi, chiamati membri
  - ad es., un insieme di prodotti o un insieme di negozi
- ciascun **membro** della dimensione è rappresentato da una riga della tabella dimensione
  - ad es., ciascuna riga della tabella **Product Dimension** descrive uno dei prodotti in vendita nella catena di negozi
- la chiave è semplice e artificiale – di solito è un intero
- gli altri campi (non chiave) della tabella dimensione memorizzano gli **attributi** dei membri
  - gli attributi sono le proprietà dei membri – che sono solitamente testuali, discrete e descrittive
  - sono usati dalle interrogazioni per vincolare e raggruppare i fatti



# Tabelle dimensione

Time Dimension

| **time_key** | date | ... |
|---|---|---|
| 1 | 1/1/2009 | ... |
| 2 | 2/1/2009 | ... |
| 3 | 3/1/2009 | ... |
| ... | ... | ... |
| ... | ... | ... |
| ... | ... | ... |
| ... | ... | ... |
| ... | ... | ... |
| ... | ... | ... |
| 1461 | 31/12/2011 | ... |

Product Dimension

| **product_key** | description | ... |
|---|---|---|
| 1 | Lattina Coca Cola | ... |
| 2 | Lattina Coca Cola Diet | ... |
| 3 | Tubo Pringles Original | ... |
| ... | ... | ... |
| ... | ... | ... |
| ... | ... | ... |
| ... | ... | ... |
| ... | ... | ... |
| ... | ... | ... |
| 9827 | Spaghetti De Cecco | ... |

| **store_key** | store_name | ... |
|---|---|---|
| 1 | MegaStore Marconi (RM) | ... |
| 2 | MegaStore EUR (RM) | ... |
| ... | ... | ... |
| ... | ... | ... |
| ... | ... | ... |
| ... | ... | ... |
| 49 | HyperStore Duomo (MI) | ... |

Store Dimension

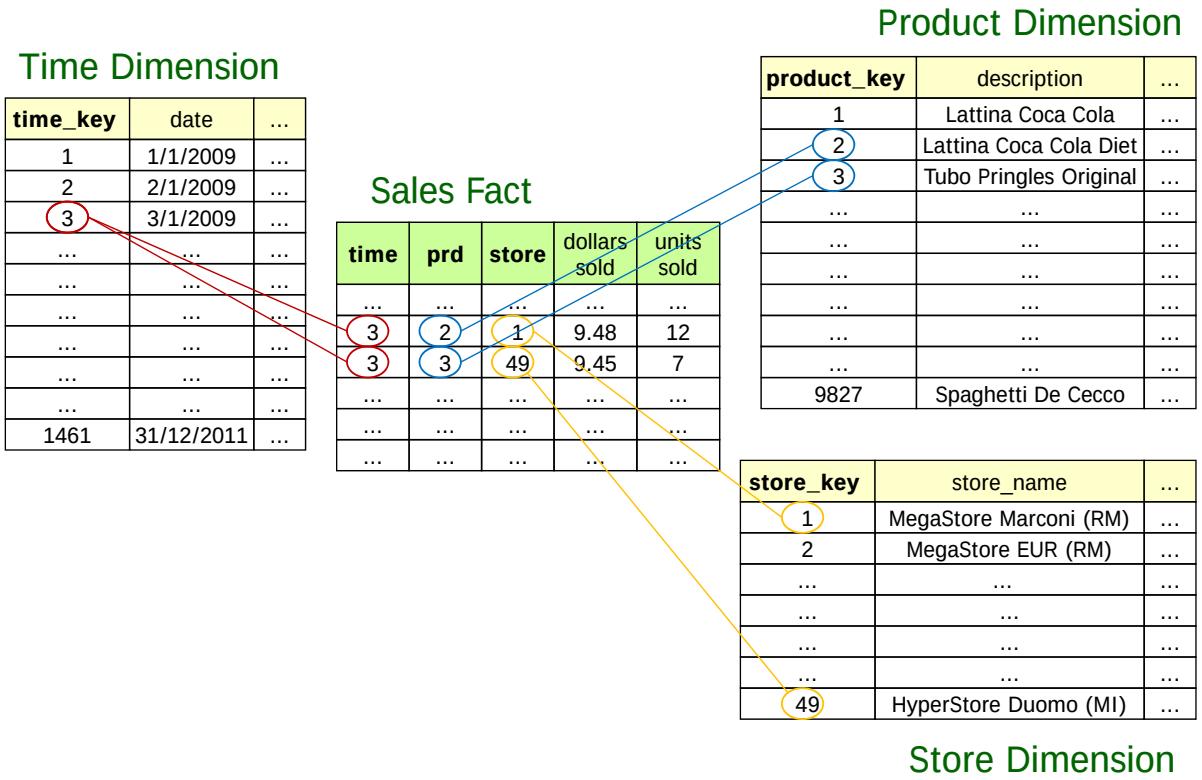
# Tabella fatti

Una **tabella fatti** (**fact table**) memorizza le misure numeriche (fatti) di un processo di business

- per **fatto** si intende una specifica misura delle prestazioni di un processo di business
- ogni fatto viene misurato durante un momento significativo dell'erogazione di un processo – con un contesto ben preciso, descritto da un insieme di membri, uno per ciascuna delle dimensioni significative per il fatto
  - per definire questo contesto, la chiave della tabella fatti è composta da chiavi esterne verso le varie tabelle dimensione, che referenziano i membri coinvolti
- gli altri campi rappresentano i fatti
  - questi fatti sono solitamente numerici, continui e additivi
- ciascuna riga della tabella fatti memorizza un insieme di misure (fatti) associati ad una particolare combinazione di membri, presa all'intersezione di tutte le dimensioni



# Tabella fatti



# Tabella fatti

Una tabella fatti serve

- a memorizzare un insieme di funzioni numeriche (in senso matematico)
  - una funzione per ciascuno dei fatti
  - il cui dominio è dato dall'insieme delle dimensioni
  - ciascuna di queste funzioni (parziale) associa un valore a ciascuna possibile combinazione dei membri delle dimensioni
- in un modo adeguato per l'analisi dimensionale

Nell'esempio

- una funzione **dollars_sold**: **Product ×Time × Store → R**
- una funzione **units_sold**: **Product × Time × Store → N**



# Schemi dimensioni, processi e grana

Ciascuno schema dimensionale serve a rappresentare i fatti di interesse per un certo processo di business, ad una certa granularità

- nell'esempio, il processo è la vendita di prodotti nei negozi di una catena di negozi
- i fatti sono
  - l'incasso in dollari (**dollars_sold**)
  - la quantità venduta (**units_sold**)
- in questo caso, la granularità a cui sono rappresentati di dati sono *il totale giornaliero per prodotto e negozio*

# Additività dei fatti

Un fatto è **additivo** se ha senso sommarlo rispetto a ogni possibile combinazione delle dimensioni da cui dipende

- nell'esempio, l'incasso in dollari è additivo – infatti 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 di BI devono spesso combinare (aggregare) i fatti descritti da molte righe 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 – ad esempio, min, max, avg



# 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 operazioni diverse dalla somma
  - ad esempio, min, max

# Sulla scelta delle chiavi

Alcuni commenti circa l'uso delle chiavi nei vari tipi di tabelle in uno schema dimensionale

- la chiave primaria di ciascuna tabella dimensione deve essere semplice e artificiale ("surrogata")
  - tipicamente, un semplice numero intero
  - non deve avere nessun significato "naturale"
  - le chiavi originali "di produzione" non vanno assolutamente usate
  - perché? riuso di chiavi naturali, cambiamento delle chiavi naturali, dimensioni che cambiano lentamente, ...
  - corrispondenza con tra chiavi naturali e chiavi artificiali gestita nell'area di preparazione dei dati
- le chiavi delle tabelle fatti sono composte da chiavi esterne delle tabelle dimensione coinvolte



# Attributi e interrogazioni

Gli attributi delle tabelle dimensione sono il principale strumento per l'interrogazione del data warehouse

- gli attributi delle dimensioni vengono usati principalmente con due finalità
  - per selezionare un sottoinsieme dei dati di interesse
    - vincolando il valore di uno o più attributi
    - ad esempio, le vendite nel corso dell'anno 2011
  - per 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

Un esempio di interrogazione

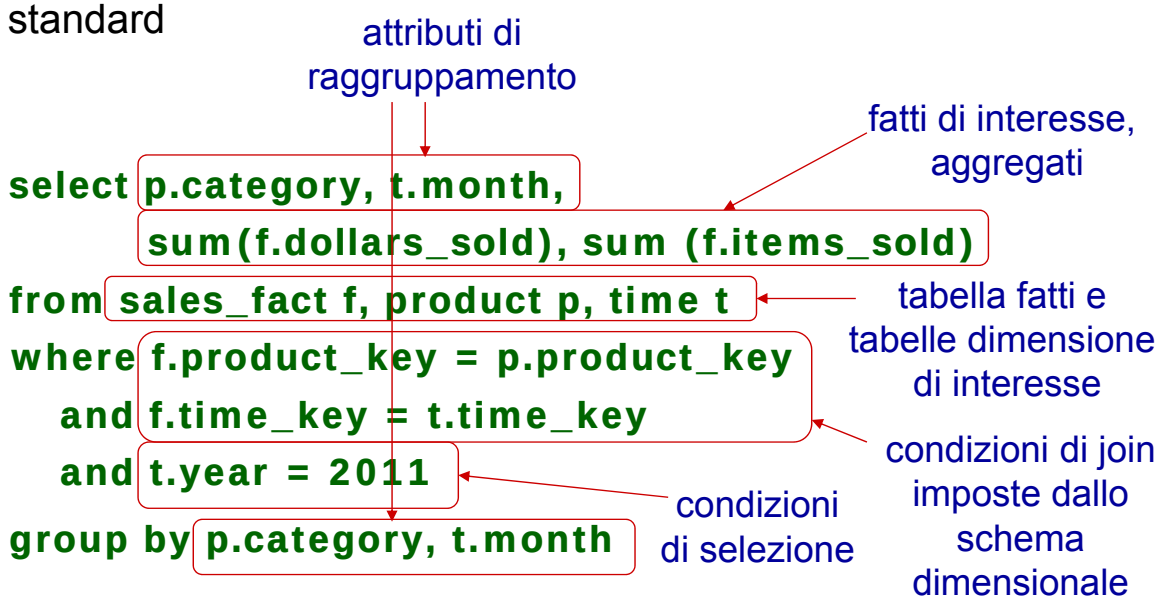
- somma degli incassi in dollari e delle quantità vendute
- per ciascuna categoria di prodotto in ciascun mese
- nel corso dell'anno 2011

| (product) category | (time) month | (sum of) dollars_sold | (sum of) units_sold |
|---|---|---|---|
| Drinks | gennaio 2011 | 21.509,05 | 23.293 |
| Drinks | febbraio 2011 | 19.486,93 | 22.216 |
| Drinks | marzo 2011 | 21.986,43 | 23.532 |
| | | | |
| Food | gennaio 2011 | 86.937,77 | 55.135 |
| | | | |
| Supplies | gennaio 2011 | 21.554,17 | 13.541 |
| | | | |



# Formato delle interrogazioni

Le interrogazione assumono solitamente la seguente forma standard




- sono comuni anche interrogazioni che effettuano confronti e/o rapporti (tassi)

# 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 un'interrogazione
- diminuisce la grana dell'aggregazione – aumenta il dettaglio dei dati

| (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 |
|---|---|---|---|---|



# Drill up

L'operazione di drill up riduce il dettaglio dei dati restituiti da una interrogazione

- il drill up avviene rimuovendo un attributo dall'intestazione di un'interrogazione
- aumenta la grana dell'aggregazione – diminuisce il dettaglio dei dati

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

drill up

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

# 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
  - consente 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



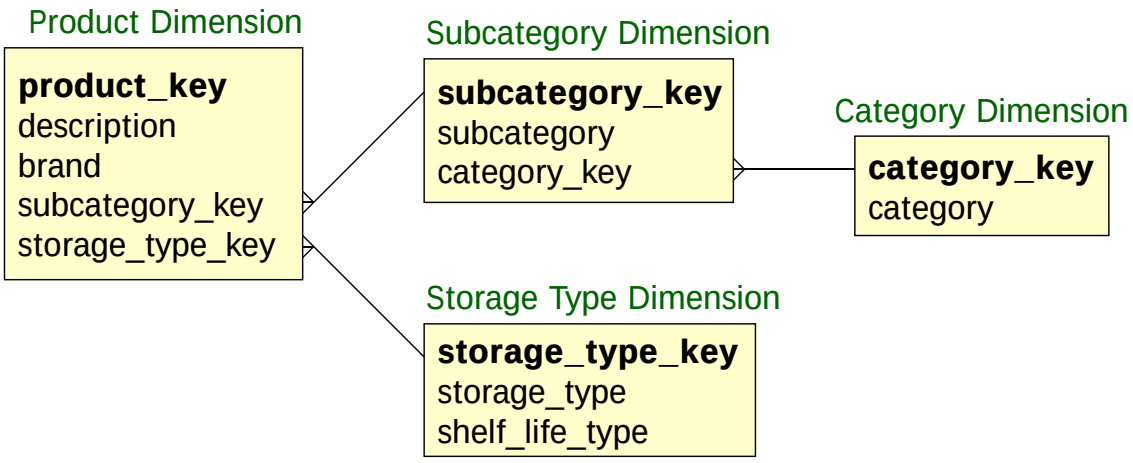
# Vantaggi della modellazione dimensionale

Ulteriori vantaggi, che vedremo successivamente

- 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
  - ad esempio, dimensioni lentamente variabili, prodotti eterogenei, tabelle fatti senza fatti, ....
  - alcune di queste tecniche saranno presentate nel seguito di questo corso

# Snowflaking

Per snowflaking di una dimensione si intende una rappresentazione “più normalizzata” di una tabella dimensione, che evidenzia delle “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 negozi
- prodotti
  - 30.000 prodotti
- 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 x 300 x 3000 = 630.000.000 record

# Resistere allo snowflaking

Lo snowflaking è solitamente svantaggioso

- inutile per l'occupazione di memoria
  - ad es., una dimensione prodotto con 30.000 record, di circa 2.000 byte ciascuno -> 60MB di memoria primaria
  - tabella fatti con invece 630.000.000 record, di circa 10 byte ciascuno -> 6.3GB di memoria primaria
  - 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 le prestazioni
- tuttavia, ci sono delle situazioni in cui è utile definire delle "sottodimensioni" – con l'apparenza di uno snowflake
  - si veda la tecnica delle mini-dimensioni