

# Grandi dimensioni

Luca Cabibbo  
marzo 2010

## Grandi dimensioni

Vengono ora studiate alcuni aspetti caratteristici tipici della gestione di dimensioni grandi (ovvero, molto numerose)

- ad esempio, questa discussione potrebbe riguardare dimensioni come prodotto e cliente
- studieremo anche le modalità di gestione delle “variazioni” nelle dimensioni e nei loro membri

# La dimensione prodotto

La dimensione prodotto descrive il catalogo completo dei prodotti venduti dall'organizzazione

- i membri della dimensione prodotto possono essere decine di migliaia
  - spesso, sono varianti produttive di tipologie di prodotti

La dimensione prodotto è spesso derivata dal file principale dei prodotti gestito dal sistema di produzione

- l'esistenza di una tale sorgente informativa semplifica la gestione della tabella dimensione per i prodotti

# La dimensione prodotto

## Product Dimension

```
product_key  
SKU_description  
SKU_number  
package_size  
brand  
subcategory  
category  
department  
package_type  
diet_type  
color  
weight  
weight_unit_of_measure  
units_per_retail_case  
units_per_shipping_case  
cases_per_pallet  
shelf_width  
shelf_height  
shelf_depth  
...
```

## Trasformazioni

In generale, i dati del file principale dei prodotti devono essere sottoposti a diverse trasformazioni, tra cui

- trasformazione del codice usato come chiave primaria del prodotto nel sistema di produzione
  - ad esempio, perché la vita del data warehouse è più lunga di quella dei prodotti, e nel sistema di produzione è ritenuto accettabile riassegnare un codice di un prodotto fuori produzione
- trasformazione del codice (trasformato) usato come chiave primaria del prodotto nel sistema di produzione in un codice compatto
  - per questioni di efficienza

## Trasformazioni

- generalizzazione del codice del prodotto per tenere traccia della modifica della descrizione o formulazione del prodotto nel tempo
  - alcune dimensioni sono soggette a modifiche nel corso del tempo
  - questo aspetto sarà trattato più avanti
- generalizzazione del codice del prodotto per descrivere “prodotti aggregati”
  - ad esempio, per assegnare un codice alle marche e alle categorie di prodotto
  - questo aspetto sarà trattato più avanti nel corso

## Trasformazioni

- introduzione di descrizioni testuali per sostituire descrizioni o codifiche criptiche
  - gli attributi nelle dimensioni sono alla base dei criteri di selezione e raggruppamento dei dati, e quindi devono essere facilmente comprensibili dagli utenti del data warehouse
    - sia nell'intestazione che nel valore
- verifica di qualità delle descrizioni testuali
  - le descrizioni testuali corrette possono poi essere utilizzate per effettuare la pulizia del file principale dei prodotti

## Trasformazioni

Le sorgenti informative del data warehouse devono essere trasformate e migliorate su una base continua

- in generale, le chiavi delle dimensioni nel data warehouse sono diverse e più generali di quelle adottate nei sistemi operazionali
- gli attributi descrittivi devono essere sottoposti a un processo di miglioramento della qualità

## Gerarchie

La dimensione prodotto contiene solitamente degli attributi che descrivono una gerarchia di prodotti

- ad esempio
  - **SKU**: Green 3-pack Brawny Paper Towels, UPC ...
  - **package\_size**: 3-pack
  - **brand**: Brawny
  - **subcategory**: paper towels
  - **category**: paper
  - **department**: grocery
- questa gerarchia è sostanzialmente una collezione di attributi che descrivono un prodotto e delle relazioni uno-a-molti tra alcune proprietà del prodotto

## Il significato del 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
  - qualsiasi attributo potrebbe essere utile
  - non è solo navigazione di gerarchie

(product) brand	(product) package_size	(sum of) dollars_sold
--------------------	---------------------------	--------------------------



drill down

(product) brand	(product) package_size	(product) color	(sum of) dollars_sold
--------------------	---------------------------	--------------------	--------------------------

Il drill down non deve essere limitato alla navigazione di gerarchie

## Gerarchie multiple

Ci sono altre motivazioni per non limitare le aggregazioni alle “gerarchie”

- una dimensione può presentare diverse gerarchie
  - più o meno indipendenti
- ad esempio, il processo delle vendite è interessato alla gerarchia delle merci
  - ma un altro processo potrebbe essere interessato a una gerarchia relativa all’organizzazione degli spazi di immagazzinamento
    - ad esempio, tipo di magazzino (normale o refrigerato) con le relative varianti di classificazione (ad esempio, frigorifero o congelatore per refrigerato)
  - raggruppamenti rispetto ad entrambe le gerarchie potrebbero essere interessanti

## Resistere allo snowflaking

Lo snowflaking di una dimensione è una rappresentazione “più normalizzata” di una tabella dimensione

- in particolare, rispetto alle gerarchie

Lo snowflaking non presenta solitamente vantaggi

- in particolare, la riduzione nell’occupazione di memoria è tipicamente trascurabile rispetto all’occupazione di memoria delle tabelle fatti
- il degrado delle prestazioni può essere invece significativo
  - in genere è sconsigliato normalizzare le dimensioni (anche grandi)
- ci sono ovviamente eccezioni a questa regola
  - le minidimensioni, studiate più avanti

## La dimensione cliente

Alcuni processi devono essere analizzati rispetto alla dimensione cliente

- ad esempio, la sede destinazione nel caso delle spedizioni
- in alcuni casi, è una dimensione veramente molto grande
  - ad esempio, quando i clienti sono una porzione significativi degli abitanti di una nazione
  - casi tipici sono i clienti di compagnie telefoniche e i “clienti” del Ministero delle Finanze
- è una dimensione caratterizzata da molti attributi (nell’ordine delle centinaia) e sicuramente da diverse gerarchie

## La dimensione cliente

Una prima versione della dimensione cliente

### Customer Dimension

```
customer_key
first_name
last_name
street_address
zip
city
county
state
age
income
sex
marital_status
education_level
total_children
children_at_home
purchase_behavior
...
```

## Attributi nella dimensione cliente

Alcune attributi della dimensione cliente sono molto utili per specificare criteri di selezione e aggregazione

- gli attributi della gerarchia geografica
  - ad esempio, zip e città
- alcuni attributi demografici
  - ad esempio, sesso e stato civile

Altri attributi non sono mai usati come criteri

- nome e cognome, e probabilmente nemmeno l'indirizzo

Infine, altri attributi sarebbe più utili di quelli mostrati se opportunamente raggruppati – nel senso di categorizzati, discretizzati

- ad esempio, fascia di età anziché età, fascia di reddito anziché reddito

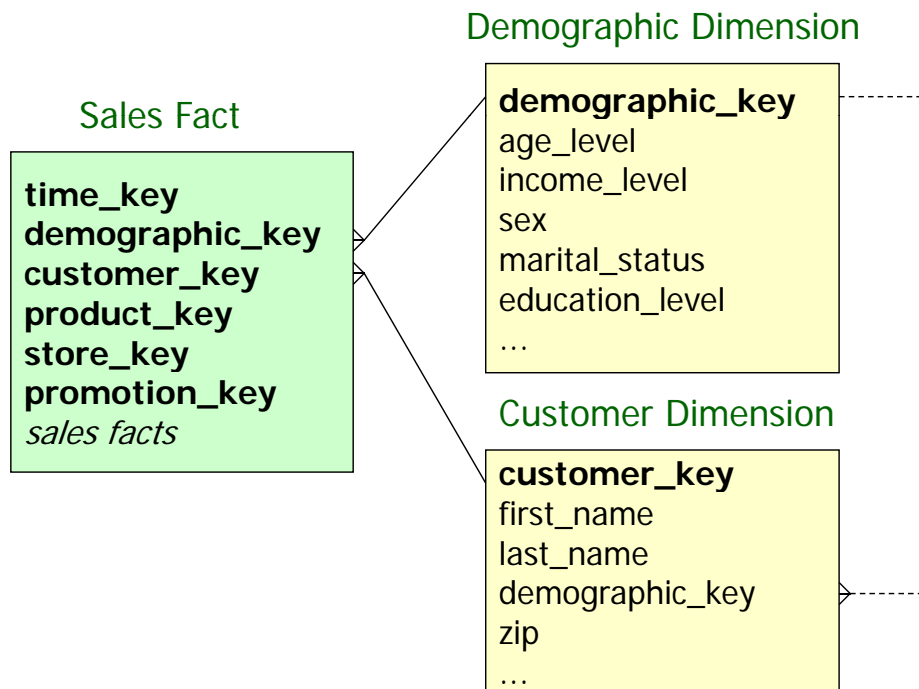
## Minidimensioni

Una tecnica utile per dimensioni con numerosi attributi è basata sulla separazione di un gruppo correlato di attributi – ad es., gli attributi demografici del cliente – in una nuova dimensione

- una tale nuova dimensione è chiamata una **minidimensione**
  - una minidimensione demografica, nell'esempio
- la minidimensione demografica descrive le possibili combinazioni significative degli attributi demografici
  - gli attributi continui sono raggruppati in fasce
    - per limitare l'esplosione nel numero di combinazioni, ad esempio, tali da definire al più 100.000 combinazioni significative distinte
  - la classificazione in fasce predefinite
    - limita (parzialmente) le possibilità di analisi
    - migliora notevolmente le prestazioni

## La dimensione cliente

La minidimensione viene solitamente referenziata sia dalla tabella fatti che dalla tabella dimensione



17

Grandi dimensioni

Luca Cabibbo

## Dimensioni che cambiano lentamente

Una delle caratteristiche delle dimensioni in uno schema dimensionale è la loro mutua indipendenza

- ogni dimensione (primaria) dovrebbe essere logicamente indipendente da tutte le altre dimensioni
  - ovvero, le dimensioni dovrebbero descrivere punti di vista sostanzialmente differenti sui fatti

In realtà, molte dimensioni dipendono dal tempo (che è una dimensione costantemente presente)

- non solo perché l'insieme dei membri della dimensione cambia nel tempo
- ma anche perché possono cambiare le descrizioni dei membri
  - come gestire questi cambiamenti?
  - ad es., se cambia la descrizione di un prodotto o un attributo demografico di un cliente?

18

Grandi dimensioni

Luca Cabibbo

## Dimensioni che cambiano lentamente

Come gestire i cambiamenti nelle dimensioni?

- un'idea potrebbe essere quella di rappresentare gli aspetti mutevoli come fatti e non come dimensioni
- tuttavia questa scelta porta comunemente a schemi poco comprensibili – nonché ad un degrado nelle prestazioni

Un altro punto di vista è il seguente

- molte dimensioni soggette a cambiamenti sono in realtà “quasi costanti” nel tempo
  - possono essere considerate sostanzialmente indipendenti dalla dimensione tempo
- oltre allo stato “corrente” della dimensione, si tiene traccia di dati che descrivono i cambiamenti nel tempo

Le dimensioni “quasi costanti” sono chiamate **dimensioni che cambiano lentamente (slowly changing dimensions)**

## Cambiamenti nelle dimensioni

Si consideri il seguente esempio

- la cliente Mary Jones non è sposata fino al 15 gennaio 2007
  - questa informazione è descritta dall'attributo **marital\_status**
- Mary Jones si sposa il 15 gennaio 2007

Come può essere gestito questo cambiamento nella dimensione cliente?

## Tipologie di cambiamenti lenti

Sono possibili tre scelte per la gestione delle dimensioni che cambiano lentamente

- *sovrascrivere il valore precedente*
  - perdendo la possibilità di tenere traccia dei cambiamenti
- *creare una nuova riga nella tabella dimensione* con i nuovi valori per gli attributi
  - segmentando accuratamente la storia delle descrizioni
  - la grana dei membri della dimensione è per versione di membri della dimensione individuata originariamente
- *definire ulteriori campi nella riga* sia per i valori correnti degli attributi che per i valori (immediatamente) precedenti
  - rappresentando un numero fissato di versioni

## Tipologie di cambiamenti lenti

Le tre scelte di gestione proposte sono rispettivamente chiamate – con poca fantasia – dimensioni che cambiano lentamente di tipo 1, 2 e 3

- **tipo 1** – *sovrascrivere il valore precedente*
  - viene modificato il campo **marital\_status** del record relativo a Mary Jones
- **tipo 2** – *creare una nuova riga*
  - viene creato una nuova riga
  - ogni transazione relativa a Mary Jones successiva al 15 gennaio 2007 verrà associata a questa nuova riga
- **tipo 3** – *definire più campi nella riga*
  - vengono usati (e aggiornati opportunamente) i campi **current\_marital\_status** e **old\_marital\_status**

## Dimensioni che cambiano lentamente di tipo 1

Il tipo 1 – *sovrascrivere il valore precedente* – è la modalità di gestione dei cambiamenti nel tempo più semplice ma, talvolta, meno efficace

- non tiene affatto traccia della storia passata dei membri della dimensione
  - infatti, dopo il 15 gennaio 2007, risulterà che Mary Jones è sposata “da sempre”
  - non è possibile partizionare la storia

Questa modalità di gestione è comunque utile nella correzione degli errori

- ad esempio, se il 15 gennaio 2007 si scopre che Mary Jones è, in effetti, sposata

## Dimensioni che cambiano lentamente di tipo 2

La modalità di gestione di tipo 2 – *creare una nuova riga* – partiziona automaticamente la storia

- per considerare le variazioni di stato dei membri non è necessario imporre vincoli sulla data dei cambiamenti nelle interrogazioni
  - le interrogazioni sono solitamente corrette ignorando l'esistenza delle versioni
  - ad esempio, se si vuole aggregare per stato civile, le transazioni di Mary Jones saranno partizionate automaticamente nei gruppi “single” e “sposata” in modo corretto

## Dimensioni che cambiano lentamente di tipo 2

Un aspetto caratteristico della modalità di gestione di tipo 2 è che gestisce “versioni di oggetti”

- ovvero, la tabella dimensione non contiene più una riga per ciascun membro della dimensione
  - piuttosto, contiene una riga per ciascuna “versione di membro” della dimensione
- in alcune dimensioni, può essere talvolta utile introdurre forzatamente e periodicamente nuove versioni dei suoi membri
  - ad es., versioni “annuali” degli impiegati di un’azienda – con attributi che descrivono lo stato di avanzamento della carriera in quel momento – consente ad esempio di analizzare lo stato della popolazione degli impiegati in precisi istanti di tempo

## Dimensioni che cambiano lentamente di tipo 2

Un aspetto caratteristico della modalità di gestione di tipo 2 è che gestisce “versioni di oggetti”

- può essere utile avere una chiave “generalizzata” della dimensione – univoca per versione di oggetto
  - bisogna tenere traccia delle chiavi di produzione e generalizzate degli oggetti soggetti a versione
    - ad esempio, gestendo la chiave di produzione “Mary Jones” e le chiavi generalizzate “Mary Jones 00” e “Mary Jones 01”
  - informazioni sulle chiavi generalizzate sono solitamente gestite nell’area di preparazione dei dati, mediante dei metadati
- nel caso dei prodotti, nuove versioni di prodotti sono associate a nuovi UPC, e considerate nuove SKU
  - in questo caso, gli UPC sono già chiavi generalizzate

## Dimensioni che cambiano lentamente di tipo 2

La modalità di gestione di tipo 2 impedisce di analizzare congiuntamente versioni diverse di uno stesso oggetto

- spesso è quello che si vuole
- in alcuni casi è necessario correlare queste diverse versioni
  - ad esempio, quando ci sono cambiamenti geopolitici, come l'istituzione di nuove province, la modifica delle estensioni di un gruppo di comuni, l'unione e la decomposizione di nazioni

In questi casi, sarebbe necessaria una modalità di gestione di tipo 3

- spesso, l'interesse nell'effettuare correlazioni è limitato nel tempo
- può allora essere sufficiente gestire questi cambiamenti come di tipo 2

## Dimensioni che cambiano lentamente di tipo 3

La modalità di gestione di tipo 3 – *definire più campi nella riga* – è la modalità di gestione più complessa da realizzare

- sono possibili diverse varianti
  - il campo “precedente” può avere il significato di valore immediatamente precedente (**old\_marital\_status**) oppure di valore originale (**original\_marital\_status**)
    - oppure possono esistere entrambi i campi
  - può avere senso un campo **current\_marital\_status\_effective\_date**
    - è necessario se si vuole partizionare la storia
- la modalità di gestione di tipo 3 viene solitamente evitata, preferendo la modalità di gestione di tipo 2

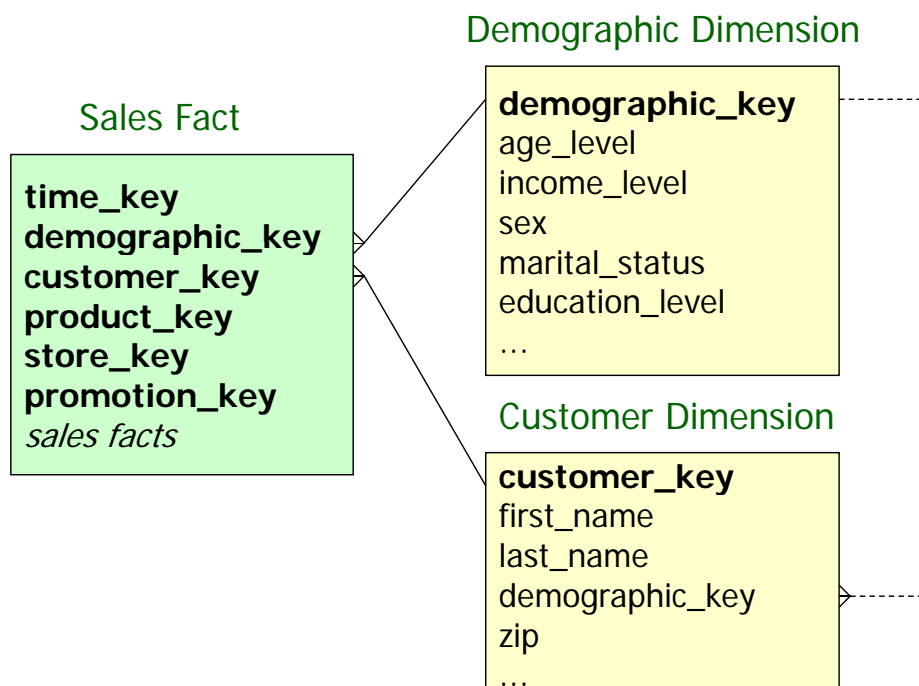
## Minidimensioni che cambiano lentamente

L'introduzione di minidimensioni (come la dimensione demografica) può avere un effetto positivo nella gestione dei cambiamenti di una dimensione principale (grande)

- ad esempio, per i clienti, i cambiamenti di cui si vuole tenere traccia avvengono solitamente nella minidimensione demografica
- in questo caso, la dimensione cliente può essere utilmente gestita con la modalità di tipo 1
  - ovvero, semplicemente cambiando il valore di **demographic\_key** nel record del cliente

## La dimensione cliente

La minidimensione è solitamente referenziata sia dalla tabella fatti che dalla tabella dimensione



## Minidimensioni che cambiano lentamente

Adottando per la minidimensione lo schema mostrato in precedenza (riferimento alla minidimensione sia nella tabella fatti che nella tabella dimensione)

- si noti che dal cliente è possibile accedere alle informazioni demografiche correnti
- da una riga della tabella fatti è possibile accedere
  - sia (direttamente) alle informazioni demografiche del cliente *al momento della transazione*
  - che (indirettamente) alle informazioni demografiche *correnti* del cliente