Grandi dimensioni e dimensioni variabili

Luca Cabibbo

aprile 2012

Grandi dimensioni e dimensioni variabili

Questo capitolo studia alcuni ulteriori aspetti importanti e caratteristici della gestione delle dimensioni

- in particolare, delle “grandi dimensioni” – intese come dimensioni con un numero elevato di membri e con un numero elevato di attributi
  - ad esempio, questa discussione potrebbe riguardare dimensioni come prodotto e cliente
- inoltre, viene studiata anche le modalità di gestione delle “variazioni” nelle dimensioni

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 master file dei prodotti gestito da un sistema operazionale

- l'esistenza di una tale sorgente informativa semplifica la gestione della tabella dimensione per i prodotti
- tuttavia, i dati dal master file dei prodotti devono essere opportunamente trasformati prima di poter essere caricati nella dimensione prodotto

La dimensione prodotto

Product Dimension

product_key
product_description
SKU_number (natural key)
brand_description
category_description
department_description
package_type_description
package_size
fat_content
diet_type
weight
weight_unit_of_measure
storage_type
shelf_life_type
shelf_width
shelf_height
shelf_depth
...

Trasformazioni

In generale, i dati del master file 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 potrebbe essere 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 motivi 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

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 master file 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

La dimensione cliente

Alcuni processi di business devono essere analizzati rispetto alla dimensione cliente

- ad esempio, il cliente che ha effettuato una vendita – o il cliente destinatario di una spedizione
- in alcuni casi, è una dimensione veramente molto grande
  - ad esempio, quando i clienti sono una porzione significativa 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
- molti dei dati in questa dimensione potrebbero venire da un *CRM – Customer Relationship Management*

La dimensione cliente

In prima battuta, la dimensione cliente contiene una riga (membro) per ciascuno dei clienti

- inoltre, i dati in questa dimensione potrebbero essere organizzati come segue

Dimension attribute	Example values
Name	Ms. R. Jane Smith, Atty
Address-1	123 Main Rd, North West, Ste 100A
Address-2	P.O. Box 2348
City	Kensington
State	Ark.
ZIP Code	88887-2348
Phone Number	888-555-3333 x776 main, 555-4444 fax

La dimensione cliente

Tuttavia, ai fini dell'analisi (selezione e raggruppamento), è certamente più utile avere in questa dimensione attributi più specifici e valori più accurati

Dimension attribute	Example values
Salutation	Ms.
Informal Greeting Name	Jane
Formal Greeting Name	Ms. Smith
First and Middle Names	R. Jane
Last Name	Smith
Title	Attorney
Ethnicity	English
Street Number	123
Street Name	Main
Street Type	Road
Street Direction	North West
…	…

La dimensione cliente

Una prima versione della dimensione cliente

Customer Dimension

customer_key
customer_id (natural key)
customer_salutation
customer_first_name
customer_last_name
…
customer_age
customer_gender
customer_income
customer_marital_status
customer_education_level
…
customer_city
customer_county
customer_state
…

Attributi nella dimensione cliente

Molti attributi della dimensione cliente sono utili soprattutto per specificare criteri di selezione e aggregazione

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

Altri attributi non saranno mai usati come criteri

- nome e cognome, e probabilmente nemmeno l'indirizzo

Infine, altri attributi potrebbero essere più utili di quelli mostrati se opportunamente raggruppati – nel senso di categorizzati, discretizzati per "bande di valori"

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

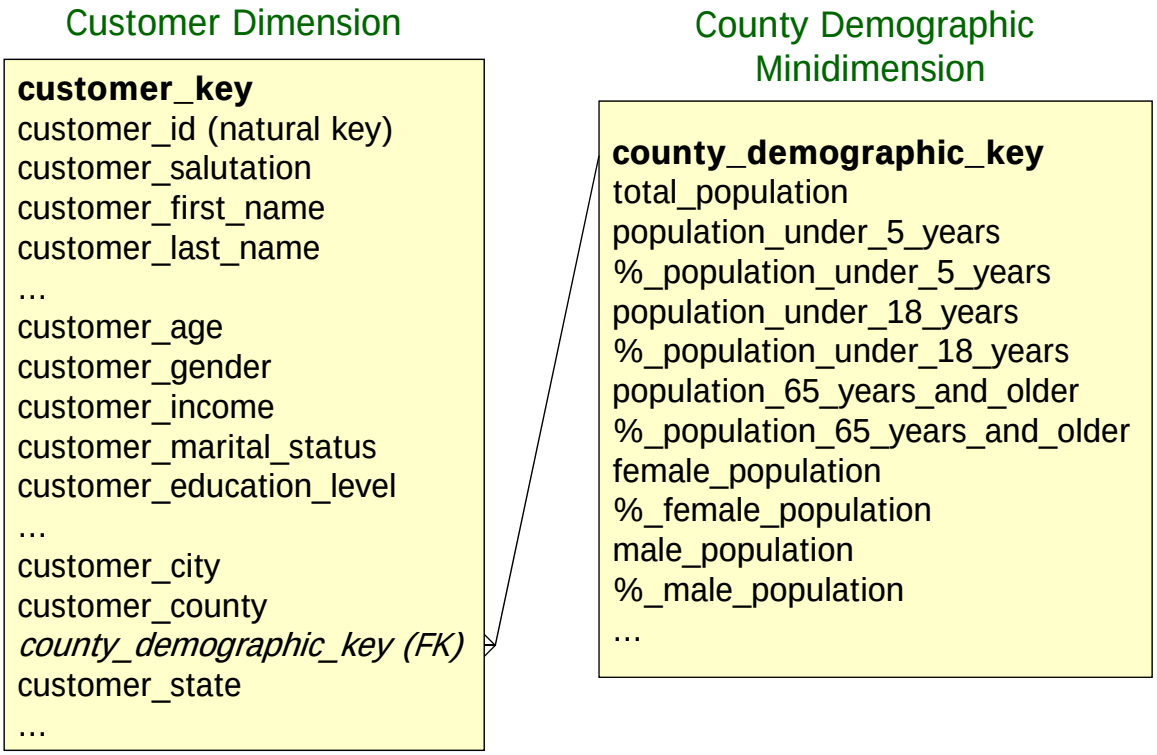
Dimensioni outrigger

Una **dimensione outrigger** – o **snowflaked** – è una dimensione derivata da un'altra dimensione – alcuni attributi di questa tabella sono separati in un'altra tabella (più) normalizzata – la tabella dimensione originale contiene una chiave esterna verso la tabella per la dimensione outrigger

- in generale lo snowflaking non è una buona idea – soprattutto in termini di prestazioni
- talvolta ha però senso usare delle dimensioni outrigger
  - ad esempio, quando la dimensione outrigger ha granularità *significativamente* diversa da quella originale
- ad esempio, è il caso degli attributi riguardanti la residenza di un cliente
  - potrebbero esserci oltre 100 attributi demografici e socioeconomici riguardanti la provincia/regione di residenza di un cliente

Dimensioni outrigger

Una seconda versione della dimensione cliente



Minidimensioni

Una **minidimensione** è simile a una dimensione outrigger

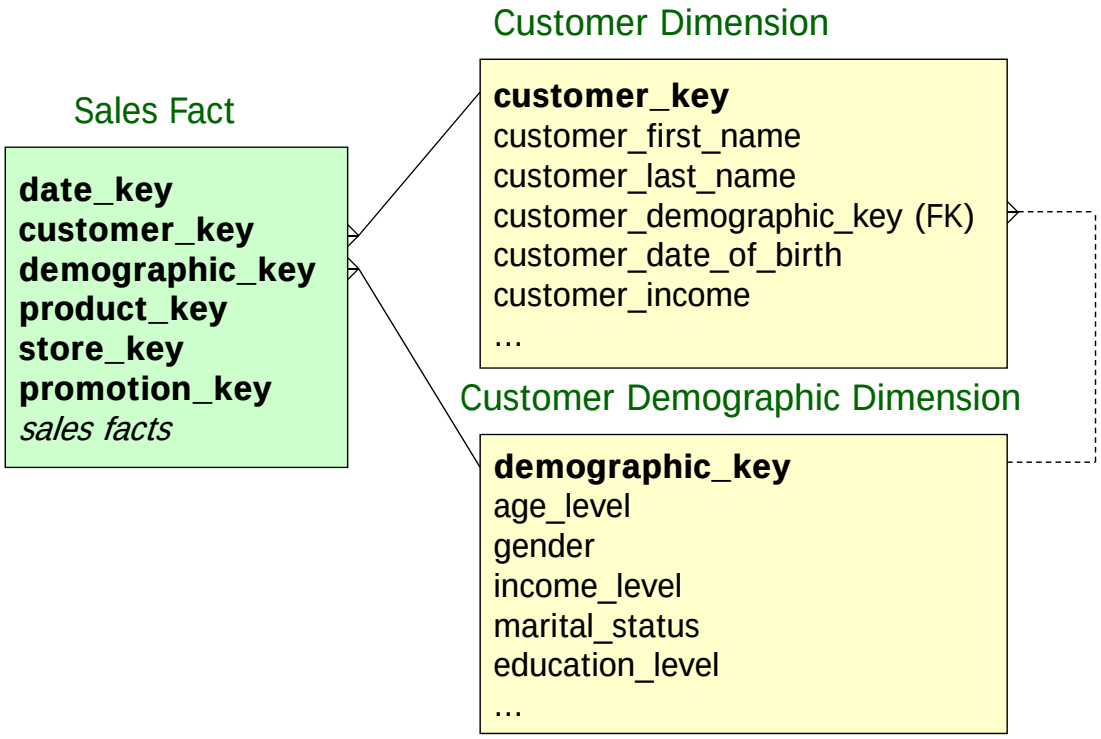
- una minidimensione è una dimensione a sé – che contiene un gruppo correlato di attributi estratti o derivati da un'altra dimensione – ad es., gli attributi demografici dei clienti
- a differenza di una semplice dimensione outrigger, una minidimensione è pensata per essere referenziata direttamente anche dalle tabelle fatti
- come vedremo, questo è utile per le alcune dimensioni che cambiano, descritte nel seguito

Ad esempio, una minidimensione demografica può descrivere le possibili combinazioni significative degli attributi demografici

- gli attributi continui sono raggruppati in fasce (predefinite)
  - potrebbe limitare (parzialmente) le possibilità di analisi
  - migliorando notevolmente le prestazioni

Minidimensione demografica

Un’altra versione della dimensione cliente



Dimensioni che cambiano lentamente

Quali cambiamenti/aggiornamenti avvengono in un data mart – ovvero, nella base di dati per un modello dimensionale?

- la tabella fatti cambia a ogni aggiornamento incrementale del data mart – ad es., giornalmente
  - normalmente vengono aggiunti nuovi fatti (righe)
  - nelle tabelle a istantanee accumulate, sono anche comuni aggiornamenti dei fatti (righe)
- le tabelle dimensione, invece, dovrebbero cambiare – e cambiano – meno frequentemente
  - ad es., vengono aggiunti nuovi prodotti, modificate le caratteristiche di prodotti esistenti, aggiunti nuovi clienti, o modificati gli attributi demografici di clienti esistenti
  - ma come gestire questi aggiornamenti delle dimensioni che cambiano (lentamente)?

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 a 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

Queste dimensioni sono chiamate **dimensioni che cambiano lentamente** – **slowly changing dimensions**, o **SCD**s

Cambiamenti nelle dimensioni

Consideriamo il seguente esempio

- un prodotto in vendita è il software *IntelliKidz 1.0* – in vendita nel reparto *Education*

Product Key	Product Description	Department	SKU Number
12345	IntelliKidz 1.0	Education	ABC922-Z

La direzione decide che, per aumentare le vendite, *IntelliKidz 1.0* debba essere spostato nel reparto *Strategy* a partire dal 15 gennaio 2010

- come può essere gestito questo cambiamento nella dimensione prodotto?

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 e della storia passata
- *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 membro della dimensione originale
- *definire ulteriori attributi nella riga* sia per i valori correnti degli attributi che per i valori (immediatamente) precedenti
  - consente di rappresentare 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 **department** della riga relativa a *IntelliKidz 1.0*
- **tipo 2** – *creare una nuova riga*
  - viene creato un'altra riga relativa a *IntelliKidz 1.0*
  - ogni transazione relativa a *IntelliKidz 1.0* successiva al 15 gennaio 2010 verrà associata a questa nuova riga
- **tipo 3** – *definire ulteriori attributi nella riga*
  - vengono usati (e aggiornati opportunamente) gli attributi **current_department** e **prior_department** della riga relativa a *IntelliKidz 1.0*

- SCD di tipo 1

SCD di tipo 1 – *sovrascrivere il valore precedente*

Product Key	Product Description	Department	SKU Number
12345	IntelliKidz 1.0	*Strategy*	ABC922-Z

La gestione dei cambiamenti adottata nelle SCD di tipo 1 è la modalità più semplice – ma, talvolta, anche la meno efficace

- non tiene affatto traccia della storia passata dei membri della dimensione
  - infatti, dopo il 15 gennaio 2010, risulterà che *IntelliKidz 1.0* è "da sempre" nel reparto *Strategy*
  - non è possibile partizionare la storia
- questa modalità di gestione è comunque utile nella correzione degli errori

- SCD di tipo 2

SCD di tipo 2 – *creare una nuova riga nella tabella dimensione*

Product Key	Product Description	Department	SKU Number
12345	IntelliKidz 1.0	Education	ABC922-Z
25984	IntelliKidz 1.0	*Strategy*	ABC922-Z

- a questa nuova riga va associato un nuovo valore per la chiave primaria
  - è un ulteriore motivo per non usare le chiavi naturali nei data mart
- ogni transazione di vendita relativa a *IntelliKidz 1.0* successiva al 15 gennaio 2010 verrà associata a questa nuova riga
  - viene partizionare la storia delle vendite

SCD di tipo 2

La modalità di gestione di tipo 2 – *creare una nuova riga* – è probabilmente la tecnica più comune nella gestione di dimensioni che cambiano lentamente

- la storia delle vendite viene partizionata automaticamente
  - tutte le vendite di *IntelliKidz 1.0* precedenti al 15 gennaio 2010 verranno associate al prodotto *12345* – e, soprattutto, al reparto *Education*
  - tutte le vendite di *IntelliKidz 1.0* successive al 15 gennaio 2010 verranno associate al prodotto *25984* – e, dunque, al reparto *Strategy*
- le interrogazioni/aggregazioni risulteranno "naturalmente" corrette
  - ad es., sia quelle per marca che quelle per reparto
  - queste interrogazioni/aggregazioni fruiranno delle variazioni senza dover far riferimento (ad es., nelle condizioni) sulle date dei cambiamenti

SCD di tipo 2

Dunque, una SCD di tipo 2 è una dimensione che rappresenta e gestisce "versioni di oggetti/membri"

- 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

SCD di tipo 2

Ulteriori osservazioni sulle SCD di tipo 2

- nella tabella dimensione, è necessario tenere traccia delle date in cui sono avvenuti cambiamenti dei suoi membri?
  - è possibile – ma non è necessario
  - è invece necessario tenere traccia dei cambiamenti e di quando sono avvenuti nell'area di preparazione dei dati

SCD di tipo 2

Ulteriori osservazioni sulle SCD di tipo 2

- che succede quando viene rilasciato il prodotto *IntelliKidz 2.0*? va gestito come un cambiamento del prodotto *IntelliKidz 1.0*?
  - no, *IntelliKidz 2.0* va gestito come un prodotto completamente diverso da *IntelliKidz 1.0*
  - ad es., le vendite di quest'ultimo proseguiranno, almeno per un po' di tempo, e non andranno certamente confuse con le vendite del nuovo prodotto

- SCD di tipo 3

SCD di tipo 3 – *definire ulteriori attributi nella riga*

Product Key	Product Description	Department	Prior Department	SKU Number
12345	IntelliKidz 1.0	*Strategy*	Education	ABC922-Z

- consente di rappresentare un numero fissato di versioni dei membri della dimensione
- consente di analizzare i fatti anche “come se” alcuni cambiamenti non fossero mai avvenuti
  - questo non è possibile né con le SCD di tipo 1 né con quelle di tipo 2

SCD di tipo 3

La modalità di gestione di tipo 3 – *definire ulteriori attributi 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 (**prior_department**) oppure di valore originale (**original_department**)
    - oppure possono esistere entrambi i campi
  - può avere senso un campo **current_department_effective_date**
    - è necessario se si vuole partizionare la storia nel tempo
- in generale, può essere applicata solo quando il numero di versioni di un oggetto – e i cambiamenti di cui si vuole tenere traccia – sono comunque fissati e limitati

SCD di tipo 3

Uno dei limiti delle SCD di tipo 2 è che non consente di associare un nuovo valore a fatti precedenti – e viceversa

- ad es., selezionando il reparto *Strategy*, non vengono considerate le vendite di *IntelliKidz 1.0* precedenti al 15 gennaio 2010 – spesso è proprio quello che si vuole

Tuttavia, in alcuni casi, si vogliono analizzare i fatti come se certi cambiamenti non fossero mai avvenuti

- ad es., quando ci sono cambiamenti geopolitici, come l'istituzione di nuove province – oppure cambiamenti nei confini delle strutture organizzative
- in questi casi, come analizzare i fatti con riferimento alla vecchia (o alla nuova) struttura geopolitica o organizzativa?
  - questo è possibile con la modalità di gestione di tipo 3

Malgrado ciò, l'uso delle SCD di tipo 3 è poco frequente

Discussione

Le SCD di tipo 1, 2 e 3 sono tre tecniche per la gestione di dimensioni che cambiano lentamente

- ciascuna tecnica ha i propri vantaggi e svantaggi
- in alcuni casi sono possibili tecniche ibride – ad es., SCD di tipo 2, in cui ciascuna versione tiene anche traccia di quale era la versione precedente
- come identificare i cambiamenti?
  - nell'area di preparazione
  - semplice se l'estrazione riesce a catturare solo le cose che sono cambiate
  - altrimenti, confrontando dati dalle ultime due estrazioni
  - come rendere efficiente questo confronto? non confrontando i campi uno alla volta – ma confrontando l'hash/checksum delle righe estratte
- come gestire dimensioni che cambiano "velocemente"?

Grandi dimensioni che cambiano

La gestione dei cambiamenti nelle grandi dimensioni è certamente problematica

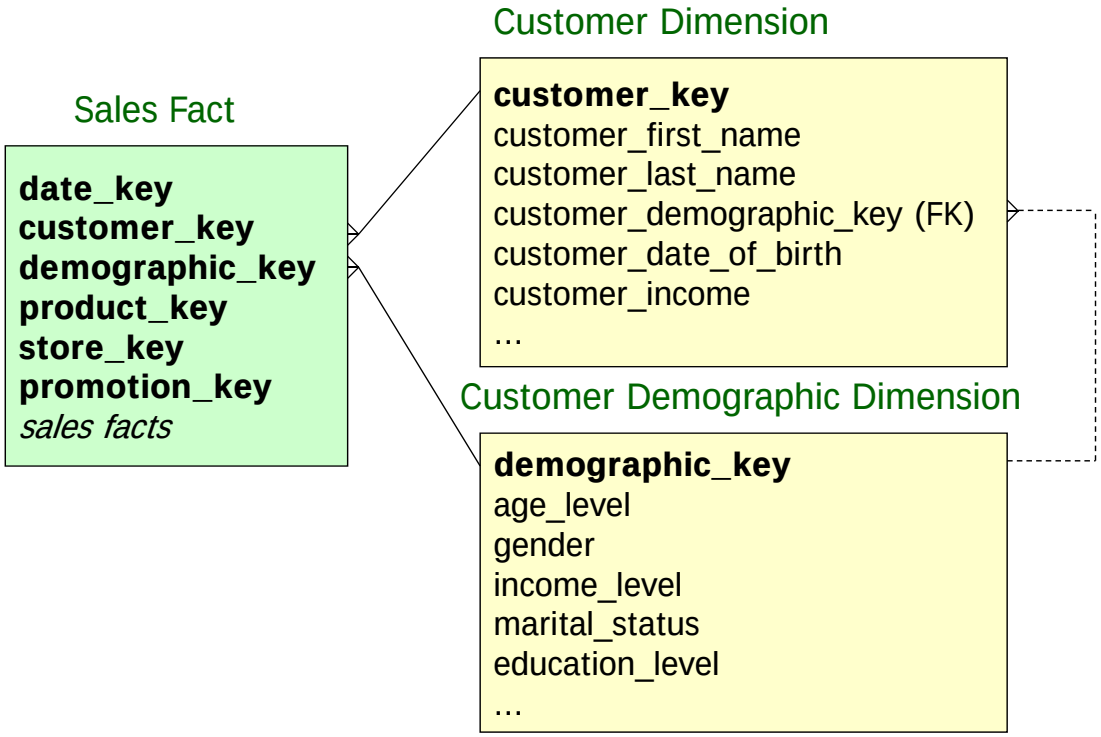
- ad es., l'uso di una modalità di gestione di tipo 2 (la più comune) renderebbe la grande dimensione *ancora più grande!*
- inoltre, le grandi dimensioni (caratterizzate spesso da molti attributi) tendono a cambiare più rapidamente delle altre dimensioni

Come gestire i cambiamenti (non lenti) di dimensioni grandi?

- un possibile approccio consiste nello spezzare gruppi di attributi che cambiano indipendentemente dagli altri gruppi di attributi in dimensioni (minidimensioni) separate
- in questo modo, inoltre, i cambiamenti nell'ambito di ciascuna minidimensione sono più lenti

Minidimensione demografica

Consideriamo nuovamente questo schema dimensionale



Minidimensione demografica

La minidimensione demografica potrebbe contenere, ad esempio, i seguenti membri

Demographic Key	Age	Gender	Income Level
1	20-24	Male	<$20,000
2	20-24	Male	$20,000-$24,999
3	20-24	Male	$25,000-$29,999
18	25-29	Male	$20,000-$24,999
19	25-29	Male	$25,000-$29,999

- ciascun membro di questa dimensione rappresenta una specifica combinazione degli attributi demografici
- notare l'uso di bande di valori

Clienti e minidimensione demografica

Ogni membro della dimensione cliente referenzia il membro della minidimensione demografica che rappresenta il suo stato demografico attuale

- ad es., *John Smith*, *maschio*, di *24* anni e reddito *$24,000* – il suo stato demografico attuale corrisponde alla chiave demografica *2*

Cust. key	Cust. name	Gender	Age	Income	Demog. key
98765	John Smith	Male	24	$24,000	2

- dunque, la tabella dimensione memorizza, di solito, lo stato demografico *corrente* del cliente

Vendite e minidimensione demografica

Una minidimensione non è semplicemente una dimensione outrigger (snowflaked)

- piuttosto, la tabella fatti referenzia direttamente le minidimensioni coinvolte
- dunque, per ogni acquisto di *John Smith*, ogni riga della tabella fatti per le vendite referenzierà sia *John Smith* (*98765*) nella dimensione cliente che il suo "stato demografico" (*2*) nella minidimensione demografica

Date key	Product key	...	Cust. key	Demog. key	...
15436	12345	...	98765	2	...

- dunque, la tabella fatti memorizza lo stato demografico del cliente *al momento della transazione*

Cambiamenti dei clienti

Come può essere gestito un cambiamento in un membro della dimensione cliente?

- ad esempio, quando *John Smith* compie *25* anni?
- nella dimensione cliente, può essere gestito come un cambiamento di tipo 1 – anche il suo stato demografico viene aggiornato, e diventa *18*

Cust. key	Cust. name	Gender	Age	Income	Demog. key
98765	John Smith	Male	*25*	$24,000	*18*

Vendite e cambiamenti

Come gestire i successivi acquisti da parte di *John Smith*?

- gli acquisti saranno ancora associati a *John Smith* – la cui chiave primaria non è cambiata
- tuttavia, i nuovi fatti nella tabella delle vendite faranno riferimento al suo nuovo stato demografico – mentre i fatti precedentemente registrati continueranno a far riferimento al suo stato demografico precedente

Date key	Product key	...	Cust. key	Demog. key	...
15436	12345	...	98765	2	...
...	...	...	...	...	...
43654	65487	...	98765	18	...

Discussione

Questo approccio presenta numerosi vantaggi

- semplice da gestire – come una SCD di tipo 1
  - la grande dimensione non diventa ancora più grande
- è possibile aggregare i dati partizionando la storia – come nelle SCD di tipo 2
  - aggregando con riferimento agli attributi della minidimensione demografica – raggiunta direttamente dalla tabella fatti
- è anche possibile aggregare i dati con riferimento allo stato corrente dei clienti – come nelle SCD di tipo 3
  - aggregando con riferimento agli attributi della minidimensione demografica – raggiunta attraverso la dimensione cliente e non direttamente dalla tabella fatti
- tuttavia, attenzione a non definire troppe dimensioni!