Vendite al dettaglio (studio di caso)

Luca Cabibbo

aprile 2012

Studi di caso

Le tecniche di modellazione dimensionale saranno illustrate mediante un certo numero di studi di caso di business

- ciascuno studio di caso è relativo a un contesto di business concreto – ma non necessariamente propone una soluzione completa o definitiva per quel particolare contesto di business
- ciascuno studio di caso ha lo scopo di introdurre (o rafforzare o specificare meglio) una o più tecniche (pattern) di modellazione dimensionale
- ciascuna tecnica più chiaramente essere usata in più contesti di business
  - anche diversi da quelli dello studio di caso in cui è stata presentata

Progettazione di uno schema dimens.

Preliminarmente, introduciamo un approccio per la progettazione di uno schema dimensionale

- uno schema dimensionale è composto da una singola tabella fatti e da un insieme di tabelle dimensione
- la progettazione di uno schema dimensionale può essere guidata da quattro decisioni principali

Bisogna però tenere anche presente che

- un data mart può essere composto da più schemi dimensionali
- un data warehouse è composto da più data mart
- una metodologia per la progettazione di un data warehouse dimensionale completo, secondo l'architettura a bus, verrà proposta più avanti

Progettazione di uno schema dimensionale

La progettazione di uno schema dimensionale può essere basata sullo svolgimento di quattro passi

- scelta del **processo di business** da modellare
- dichiarazione della **grana** del processo di business
- scelta delle **dimensioni** da cui dipende ciascuna riga della tabella fatti
- identificazione dei **fatti** misurabili che popoleranno ogni riga della tabella fatti

Queste scelte devono essere guidate

- soprattutto dai requisiti degli utenti di business
- anche dalle sorgenti informative disponibili e dalla loro qualità – bisogna però resistere alla tentazione di considerare solo i dati (apparentemente) disponibili

Progettazione di uno schema dimensionale

Scelta del **processo di business** da modellare

- *quale processo di business stiamo modellando?*
- per processo di business si intende un'attività di business eseguita dall'organizzazione – di solito è supportato da uno o più sistemi operazionali, i cui dati possono essere utilizzati per popolare lo schema dimensionale
- gli indici di prestazioni a cui sono interessati gli utenti di business del sistema DW/BI derivano proprio dalla misurazione di processi di business
- alcuni esempi di processi di business
  - vendite, ordini, fatturazione, magazzino/inventario, approvvigionamento, ...

Progettazione di uno schema dimensionale

Dichiarazione della **grana** del processo di business

- *che cosa descrive, esattamente, una singola riga della tabella fatti?*
- per grana si intende appunto il livello di dettaglio atomico che deve essere rappresentato nella tabella fatti per il processo di business di interesse
- esempi di livelli per la grana
  - transazioni individuali
  - istantanea (snapshot) periodica individuale – ad es., giornaliera o mensile
  - istantanea accumulata – ad es., stato delle consegne
  - ...

Progettazione di uno schema dimensionale

Scelta delle **dimensioni** da cui dipende ciascuna riga della tabella fatti

- *in che modo sono descritte le misurazioni del processo di business per il processo che stiamo modellando?*
- una dimensione è un insieme di membri, caratterizzati da un certo numero di attributi – da usare nelle selezioni e nei raggruppamenti
- le dimensioni rappresentano il contesto in cui il processo di business viene misurato (o analizzato)
- esempi di dimensioni
  - tempo, prodotto, cliente, promozione, magazzino, tipo di transazione, stato, ...
- la scelta della grana stabilisce alcune delle dimensioni
  - è però possibile scegliere anche dimensioni aggiuntive

Progettazione di uno schema dimensionale

Identificazione dei **fatti** misurabili che popoleranno ogni riga della tabella fatti

- *che cosa stiamo misurando?*
- gli utenti di business sono interessati in analizzare proprio queste misurazioni del processo di business di interesse
- i fatti sono misure (solitamente numeriche, continue e additivi) del processo di interesse – possono essere misure dirette catturate in momenti significativi del processo, oppure dati derivati da o associati a queste misure
- esempi di fatti
  - quantità venduta, incasso della vendita, ...

Progettazione di uno schema dimensionale

La progettazione di uno schema dimensionale è guidata da due forze fondamentali

- le esigenze di analisi da parte degli utenti/analisti di business del sistema DW/BI (top down)
- le sorgenti di dati disponibili (bottom-up)

- entrambe le forze devono essere tenute in considerazione

Il processo delle vendite al dettaglio

Si consideri il seguente caso di studio, relativo al **processo delle vendite al dettaglio** in una catena di negozi alimentari (**retail case study**)

- il contesto è la direzione di una grande catena di alimentari (negli US)
- la catena comprende 100 grandi supermercati, distribuiti in 5 stati
- ogni supermercato ha numerosi reparti (department)
  - ad esempio, drogheria, surgelati, latticini, carne, frutta e verdura, pane, pasta, fiori, liquori, ...

Il processo delle vendite (2)

- la catena di negozi vende circa 60.000 tipi di prodotti individuali – chiamati **unità di vendita** (**SKU**, stock keeping unit)
  - un esempio di SKU è “lattina di Diet Coke”
  - ogni variante di confezionamento dei prodotti costituisce una diversa SKU
- circa 55.000 delle SKU provengono da fornitori esterni, e riportano un codice a barre chiamato **codice universale del prodotto** (**UPC**, universal product code)
  - la grana degli UPC è la stessa delle SKU
- le altre 5.000 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 codice SKU

Il processo delle vendite (3)

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)
  - una cassa POS è il luogo in cui i prodotti vengono venduti e “lasciano” il negozio – è un ottimo punto in cui misurare il processo delle vendite al dettaglio
- per quanti riguarda gli acquisti dai fornitori
  - 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
- inoltre, l’inventario è spesso realizzato girando tra gli scaffali e guardando quali prodotti sono assenti

Il processo delle vendite (4)

La direzione della catena di 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 (5)

Le promozioni comprendono

- riduzioni temporanee di prezzo (TPR)
- pubblicità (su diversi media)
- esposizione opportuna – sugli scaffali o alla fine dei corridoi
- buoni sconto

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 data mart e schema dimensionale che lo compone – è basata sulla comprensione del processo di business di interesse, delle relative misure delle prestazioni e dei dati effettivamente disponibili

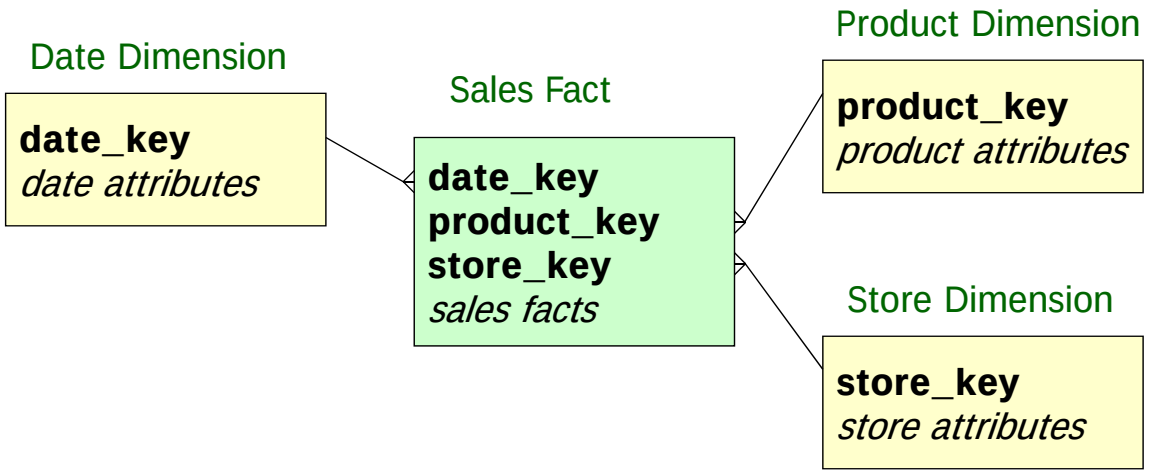
Prima decisione: scelta del **processo di business** da modellare

- il processo di business di interesse, in questo studio di caso, è il processo delle vendite al dettaglio dei prodotti nei negozi
- porterà alla definizione di un data mart delle **vendite dei prodotti**

- Scelta della grana

Seconda decisione: dichiarazione della **grana** del processo

- per il data mart per il processo delle vendite sono possibili diverse scelte per la grana delle misurazioni del processo
- ad esempio, unità di vendita (SKU) per negozio per giorno



Scelta della grana

La scelta della grana ha influenza su

- scelta delle dimensioni del data mart
- occupazione di memoria del data mart
- tipologie di analisi che possono essere effettuate – ma anche su quelle che non possono essere effettuate

Alcune possibili scelte per la grana

Alcune possibili scelte per la grana

- unità di vendita (SKU) per negozio per giorno
- unità di vendita per negozio per settimana
  - non permette di distinguere le vendite nei fine settimana da quelle degli altri giorni
- unità di vendita (SKU) per negozio per giorno per promozione
  - consente di analizzare l'impatto delle promozioni sulle vendite

Alcune possibili analisi

Se viene scelta, ad esempio, la grana “unità di vendita per negozio per promozione per giorno”, allora è possibile effettuare, tra l’altro, 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 esterni, forniti da società di analisi specializzate

La grana delle transazioni individuali

Di solito, se possibile, la grana per un processo viene scelta come la più atomica o granulare possibile

- per “dati atomici” si intendono i dati più dettagliati che vengono raccolti/misurati per il processo – ovvero, quei dati non possono essere dettagliati ulteriormente
- gli schemi dimensionali basati sui dati più granulari possibili producono il progetto più robusto – insieme alla maggior occupazione di memoria

Nel processo delle vendite al dettaglio, la grana più fine è

- per **voce di vendita** in una **transazione individuale**
  - una riga nella tabella fatti per ciascuna voce (riga) di ciascuno scontrino di vendita
  - permette di effettuare interessanti analisi di market basket – ancor di più se è nota anche l’identità del cliente

Alcune possibili analisi

Se, ad es., viene scelta la grana delle voci di vendite – ovvero, “per unità di vendita per scontrino di vendita per negozio per giorno”, allora è possibile effettuare, tra l’altro, la cosiddetta “analisi del paniere” (market basket analysis, MBA)

- quale regole di associazione esistono tra i prodotti venduti nel supermercato? ovvero, quali le coppie di prodotti che vengono venduti spesso insieme?
- ogni regola ha
  - supporto – la probabilità che i due prodotti vengano venduti insieme
  - confidenza – la probabilità che, avendo già acquistato il primo prodotto, viene acquistato il secondo prodotto
- la MBA non è possibile con le altre grane, non atomiche, descritte in precedenza – mentre le altre analisi proposte possono essere svolte con riferimento alla grana atomica

Altre considerazioni sulla grana

Nessuna delle analisi proposte è interessata direttamente al fatto che in una certa vendita sia stata venduta una certa SKU

- non è di solito di interesse presentare nel risultato dell’analisi i fatti memorizzati individualmente nel DW
- 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

Terza decisione: scelta delle **dimensioni** da cui dipende ciascuna riga della tabella fatti

- come avremo modo di vedere, ci sono due tipi principali di dimensioni
  - *dimensioni primarie* – hanno a che vedere con la grana delle misurazioni del processo, e sono indipendenti tra loro
  - *dimensioni supplementari* – non sono indipendenti dalle dimensioni primarie, ma sono utili per analizzare il processo di interesse

Dimensioni primarie

Fissati il processo (vendite al dettaglio) e la grana (ad esempio, per voce di vendita per unità di vendita per negozio per giorno) bisogna scegliere le dimensioni

- per alcune dimensioni la scelta è immediata
- si tratta delle **dimensioni primarie**, che fissano la grana della misurazione delle prestazioni del processo di interesse – ovvero, il contesto nell'ambito del quale viene effettuata ciascuna singola misurazione del processo
  - nel nostro esempio: in che transazione? che cosa? dove? quando?
  - dunque, nel data mart delle vendite, le dimensioni primarie potrebbero essere: scontrino di vendita, tempo, prodotto e negozio
- le dimensioni primarie sono sostanzialmente indipendenti tra loro

Dimensioni supplementari

La scelta di altre dimensioni è meno ovvia

- si tratta delle **dimensioni supplementari** (o **secondarie**) – dimensioni utili per descrivere o analizzare il processo di interesse
  - ad esempio, una dimensione supplementare potrebbe essere la dimensione promozione – perché si vuole analizzare l'impatto delle promozioni sulle vendite al dettaglio
- le dimensioni supplementari non sono indipendenti – ovvero, dipendono funzionalmente dalle dimensioni primarie
- per questo, l'introduzione di dimensioni supplementari non dovrebbe cambiare la grana del processo di interesse
  - se così non fosse, allora andrebbe ridefinita la grana per il processo – e in ogni caso la dimensione potrebbe essere primaria anziché supplementare

Scelta delle dimensioni

Ulteriori possibili dimensioni primarie (non scelte in questo studio di caso perché non accessibili dalle sorgenti informative a disposizione)

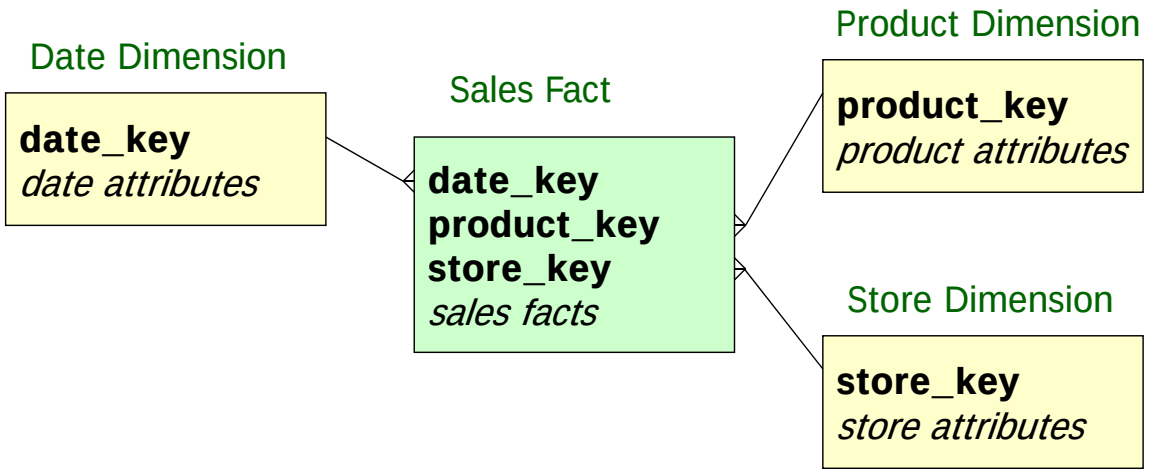
- il cliente che ha effettuato l'acquisto
- il tipo di pagamento

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
- che tempo faceva in quel giorno in quella città

Schema dimensionale

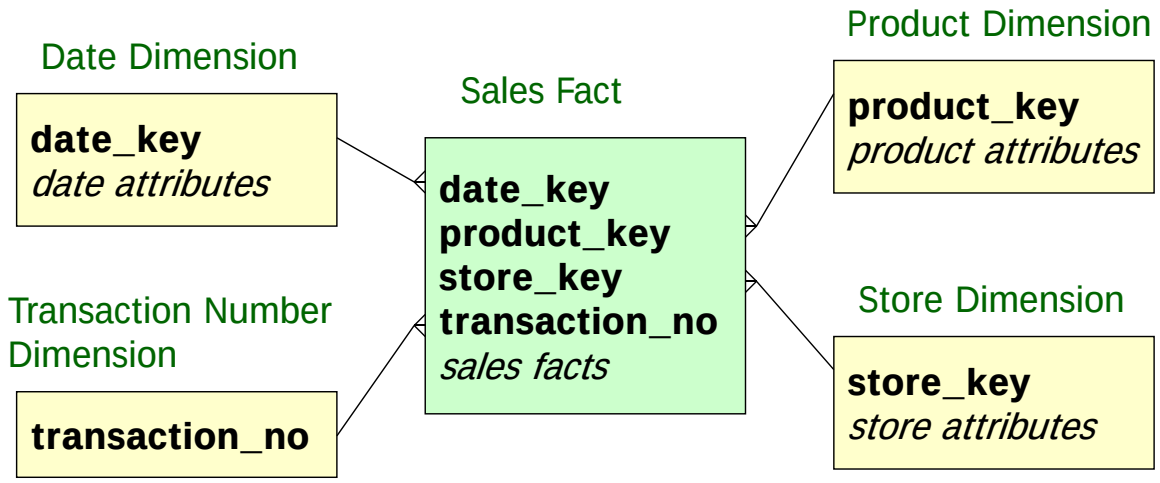
Data mart delle vendite giornaliere, per prodotto e negozio



- si tratta di una versione preliminare, la scelta degli attributi delle dimensioni verrà fatta più avanti

Schema dimensionale – varianti

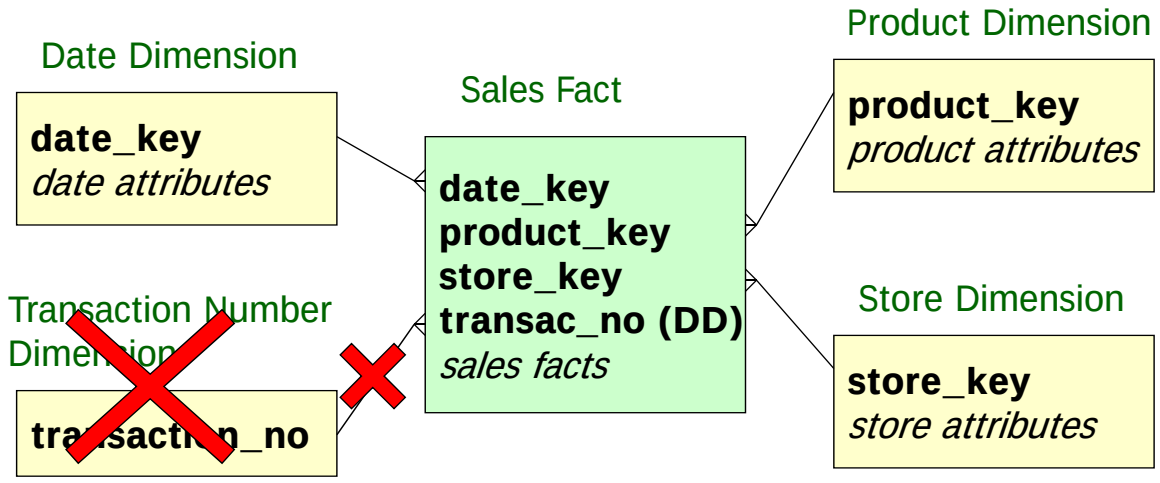
Data mart delle vendite, per riga di vendita, prodotto, negozio e giorno



Schema dimensionale – varianti

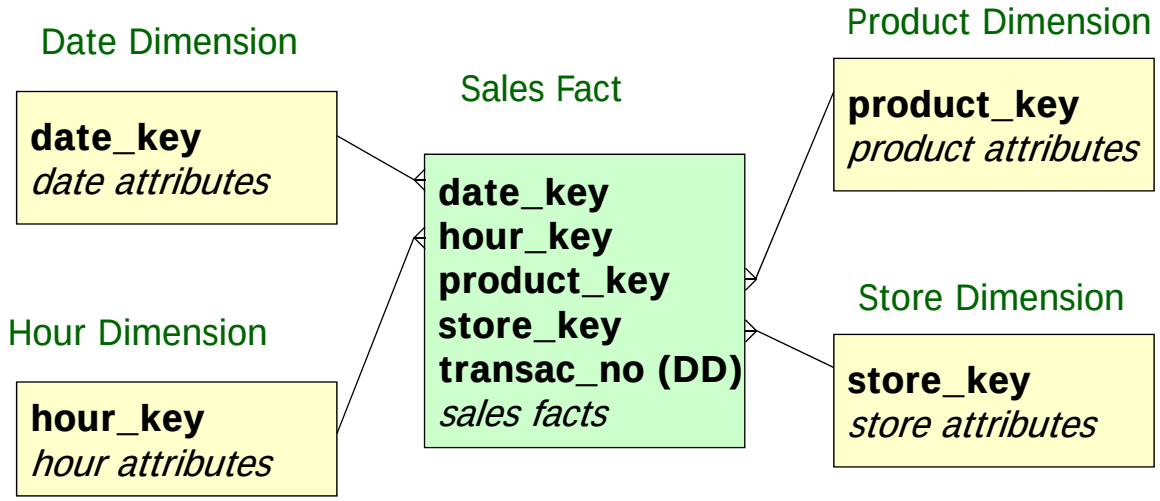
Data mart delle vendite, per riga di vendita, prodotto, negozio e giorno

- poiché la dimensione transazione (lo scontrino di vendita) non ha attributi, la relativa tabella dimensione non è strettamente necessaria – si parla in questo caso di **dimensione degenere**



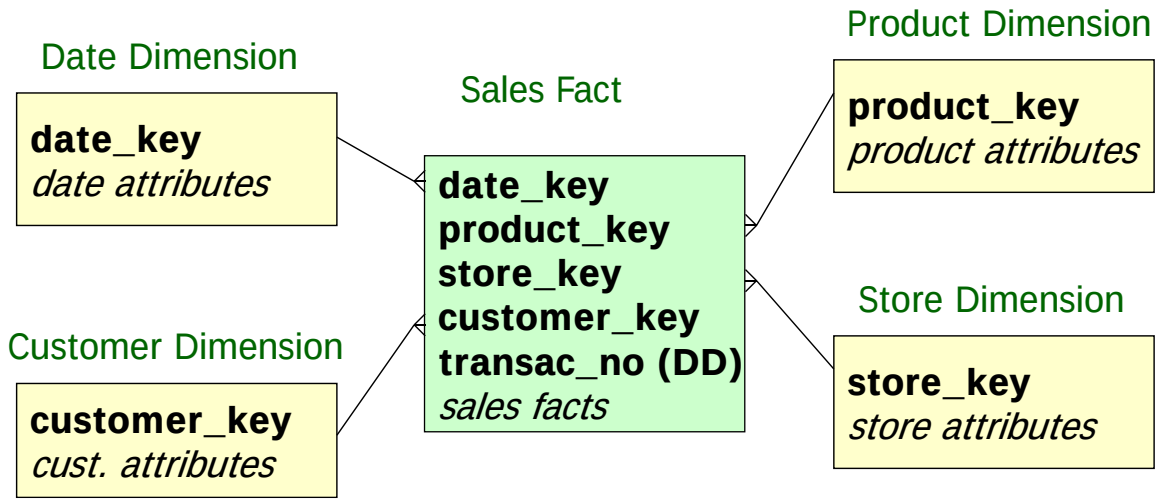
Schema dimensionale – varianti

Data mart delle vendite, per riga di vendita, prodotto, negozio, giorno e orario di vendita



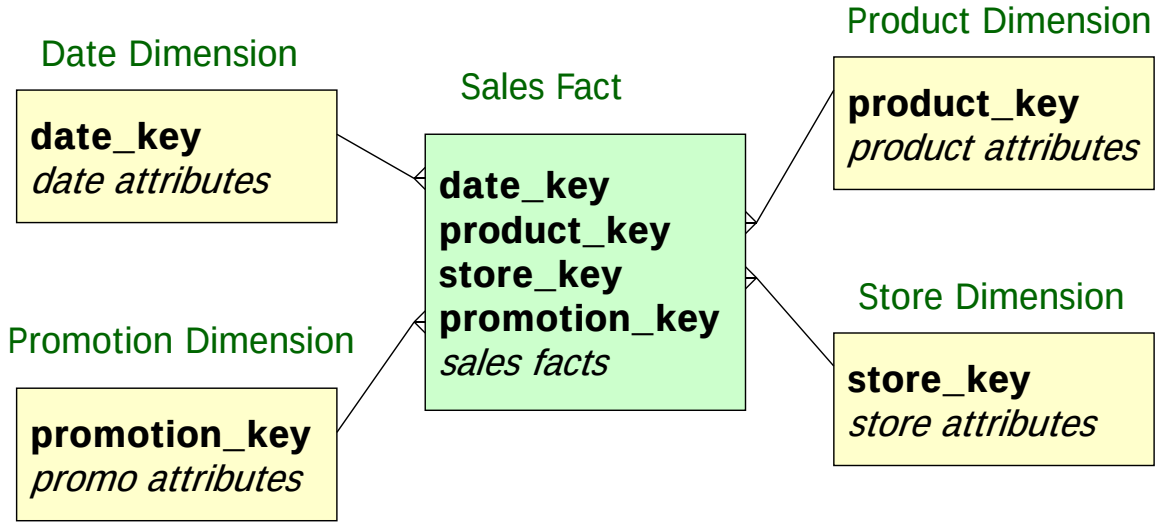
Schema dimensionale – varianti

Data mart delle vendite, per riga di vendita, prodotto, negozio, giorno e cliente



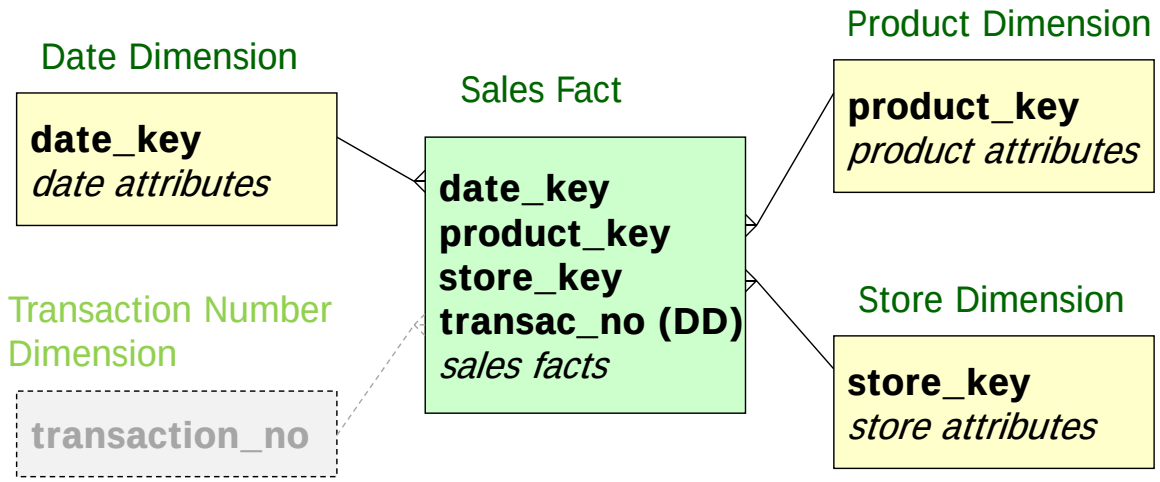
Schema dimensionale – varianti

Data mart delle vendite giornaliere, per prodotto, negozio e promozione



Schema dimensionale

D'ora in poi, faremo principalmente riferimento al data mart delle vendite, per riga di vendita, prodotto, negozio e giorno



- Identificazione dei fatti

Quarta decisione: identificazione dei **fatti** misurabili che popoleranno ogni riga della tabella fatti

- ovvero, delle misure delle prestazioni del processo di interesse
- sono misure, solitamente numeriche e additive, catturate in eventi/momenti significativi del processo, oppure altri dati associati a questi eventi

Identificazione dei fatti

Quali i fatti/misure relativi alle vendite dei prodotti (per scontrino per unità di vendita per negozio per giorno)?

- numero di pezzi venduti (**sales_quantity**)
- incasso in dollari (**sales_dollar_amount**) – **sales_quantity** x prezzo unitario
  - nota: preferibile al prezzo unitario, che non è additivo
- costo in dollari (**cost_dollar_amount**) – potrebbe essere riferito al costo standard dei prodotti venduti come stabilito dai relativi fornitori, oppure una misura più sofisticata
- profitto lordo in dollari (**gross_profit_dollar_amount**) – **sales_dollar_amount** - **cost_dollar_amount**
- perché rappresentare esplicitamente delle misure chiaramente derivate? garantisce, ad es., che tutti i report si riferiscano a quel dato in modo consistente

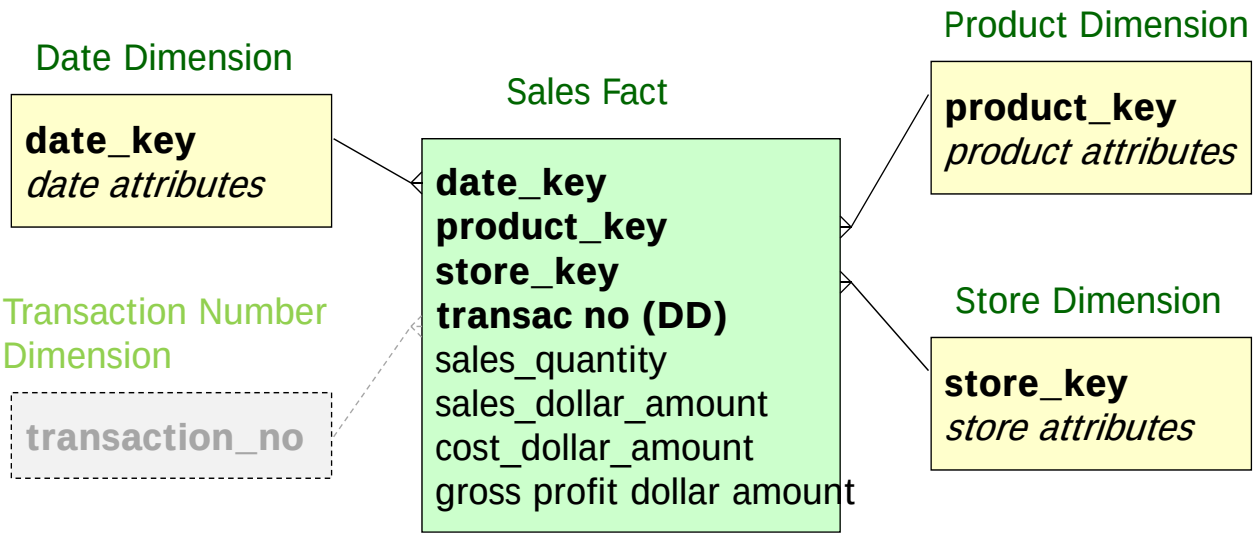
Identificazione dei fatti

E se invece avessimo scelto un data mart delle vendite giornaliere dei prodotti (per unità di vendita per negozio per giorno)?

- numero totale di pezzi venduti – per quella unità di vendita, negozio e giorno
- incasso totale in dollari
- costo totale in dollari
- profitto lordo totale in dollari
- numero di clienti (**customer_count**)
  - che hanno acquistato quel prodotto (SKU) in quel negozio e giorno
  - calcolato contando il numero di scontrini in cui è presente l'unità di vendita

Schema dimensionale

Nuova versione dello schema dimensionale



- 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 ($4·10^9$ per anno) e il costo medio della voce di vendita ($2)
  - ci sono $2·10^9$ di voci nelle transazioni individuali
- le voci nelle transazioni individuali giornaliere per negozio sono $2·10^9$ / (365·100) = 55.000 circa

Stima della taglia dei dati

Stima dell'occupazione di memoria della tabella fatti

- ipotesi
  - la chiave delle tabelle dimensione è un intero
    - di 4 byte per tempo, prodotto e numero di transazione
    - di 2 byte per negozio
  - i quattro campi chiave della tabella fatti occupano 14 byte
  - ognuno dei 4 fatti è rappresentato da un numero di 4 byte
- ogni riga della tabella fatti occupa 30 byte
- la tabella fatti contiene $2 \cdot 10^9$ righe per anno
- se vengono mantenuti dati storici relativi a tre anni, l'occupazione di memoria della tabella fatti è di circa 60GB (di spazio primario)

- Disponibilità dei fatti

Alcune misure relative al processo delle vendite al dettaglio possono essere ottenute direttamente 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
  - ad esempio, dati sugli incassi e sul numero di unità vendute

Tuttavia, per ottenere altre misure potrebbe essere necessaria l'applicazione di tecniche specifiche

- in questo caso, il costo sostenuto a fronte della vendita di alcune unità di prodotti

Parentesi – Activity Based Costing (ABC)

La **Activity Based Costing** (**ABC**) [da Wikipedia, aprile 2012] è un metodo di analisi dei costi di un'industria o impresa che fornisce dati sull'effettiva incidenza dei costi associati a ciascun prodotto e a ciascun servizio venduto dalla ditta stessa

- metodologia dell'ABC – cinque fasi per comprendere e quindi migliorare i rapporti di input-output di una ditta
  - l'input è ciò che è necessario o utilizzato per produrre l'output, ovvero il prodotto o servizio
  - l'input è rappresentato dai costi e dalle attività in gioco, che possono essere produttive e necessarie, oppure esserlo meno
  - l'analisi ABC permette di comprendere se una parte dell'input è improduttiva o non necessaria e quale parte sia

Parentesi – Activity Based Costing (ABC)

Le cinque fasi dell'ABC

- analisi delle attività – l'azienda deve comprendere esattamente quali attività svolge, intendendo con attività non tanto cosa produce, quanto le differenti cose necessarie per far sì che un dato prodotto o servizio prenda forma e sia vendibile sul mercato
- raccolta dati relativi ai costi – fissi e variabili
- riconduzione dei costi alle rispettive attività
- calcolo dell'output – *questa fase si occupa di calcolare il reale ammontare dei costi per unità produttiva*
- analisi dei costi

Progettazione delle dimensioni

Ci occupiamo ora della progettazione delle tabelle dimensione

- ovvero, della scelta degli attributi nelle tabelle dimensione
- questi attributi sono di solito descrittivi e testuali
- il loro scopo è sostenere, nelle analisi, le operazioni di *selezione* e di *raggruppamento* dei dati di interesse – e pertanto vanno scelti sulla base di necessità in questo senso

La dimensione data

La dimensione data (nel caso in esame) descrive i giorni di un intervallo temporale di interesse

- i membri della dimensione data sono i giorni dell'intervallo di interesse

La dimensione data – o in ogni caso una o più dimensioni temporali – è presente nella maggior parte degli schemi dimensionali, e praticamente in tutti i data warehouse

- la realizzazione di una tabella dimensione per le date è semplice
  - può essere facilmente pre-calcolata – ad es., usando un foglio Excel
  - i giorni per dieci anni sono poco più di 3.650

La dimensione data

È necessaria una tabella dimensione data 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 data esplicita sono relativi alla possibilità di scegliere gli attributi su cui fare selezioni, raggruppamenti, nonché quelli per realizzare dei confronti
  - ad es., 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

Attributi della dimensione data

- **date_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

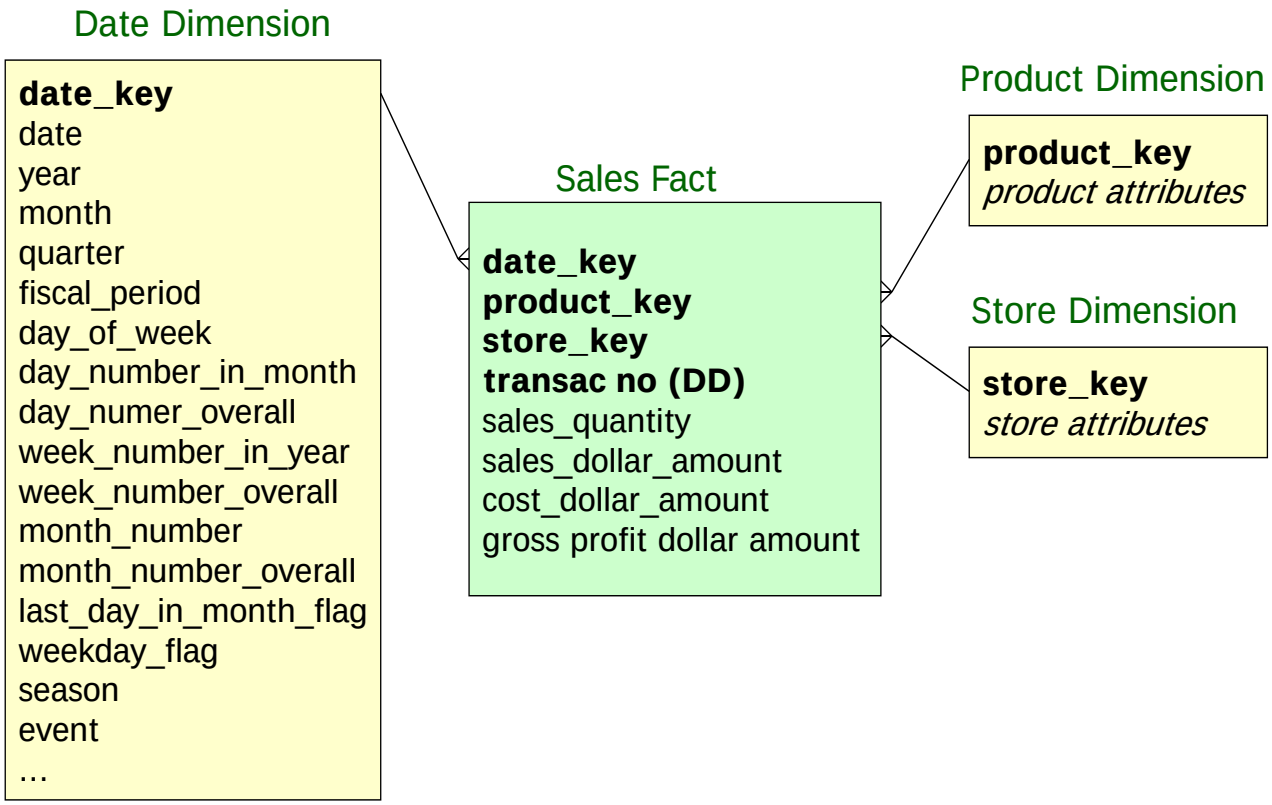
Attributi della dimensione data

- **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

Attributi della dimensione data

- **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 data



La dimensione data

Date Dimension

date_key	date	year	quarter	day_of_week	…
1	1/1/2009	2009	1Q2009	thursday	…
2	2/1/2009	2009	1Q2009	friday	…
3	3/1/2009	2009	1Q2009	saturday	…
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…
…	…	…	…	…	…
1461	31/12/2011	2011	4Q2011	saturday	…

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 master file 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 master file 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 master file 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 la dimensione geografica primaria del nostro studio di caso

- ogni negozio occupa un punto nello spazio
- i negozi possono essere organizzati e raggruppati rispetto a ogni possibile “geografia”
  - ad esempio (negli Stati Uniti) ad una geografica “politica” – per zip code, città, contea, stato
  - ma è anche possibile organizzare i negozi rispetto ad una geografia “aziendale” – ad esempio, per distretto di vendita e regione di vendita (nozioni relative alla struttura organizzativa della catena di negozi)

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
- altri attributi

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 numero di transazione

La dimensione numero di transazione rappresenta gli scontrini di vendita

- utilizzando anche questa dimensione come dimensione primaria, la grana dei dati nella tabella fatti diventa quella di una riga per ciascuna voce di scontrino di vendita
  - si tratta della grana più dettagliata possibile per il processo delle vendite al dettaglio
  - ad esempio, se anche si avesse una dimensione cliente, la grana non aumenterebbe

Dimensioni degeneri

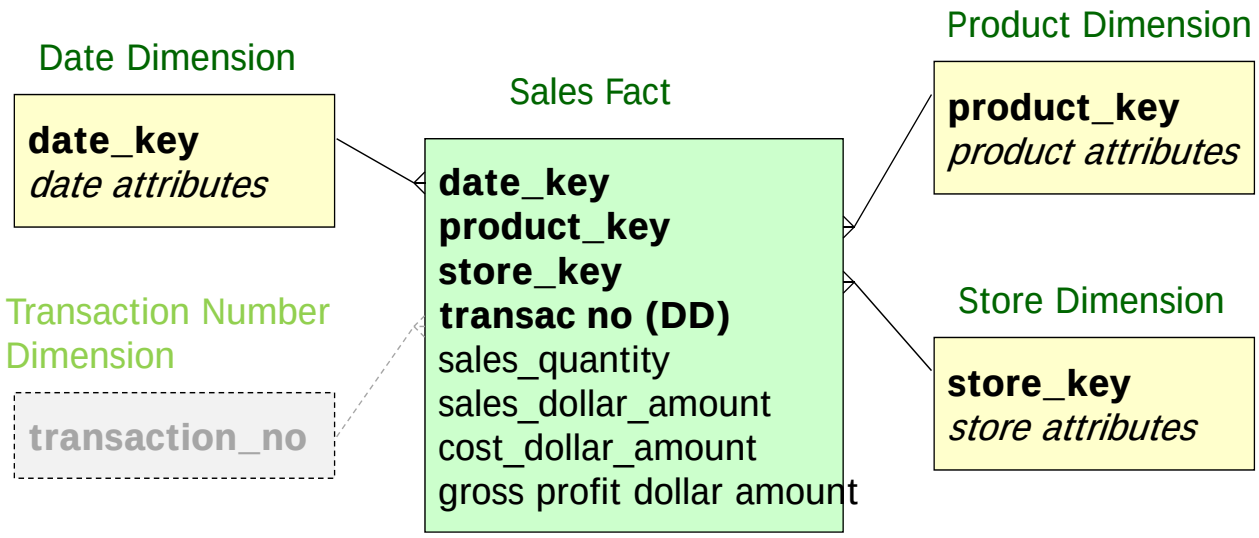
Quali attributi (informazioni) associate agli scontrini di vendita?

- ad es., data e negozio – cliente, se disponibile
- tuttavia, è preferibile rappresentare queste informazioni in dimensioni separate

Dunque, non ci sono informazioni descrittive interessanti associate ai membri di questa dimensione

- pertanto, la dimensione numero di transazione può essere rappresentata come **dimensione degenere**
  - senza tabella dimensione
  - la tabella fatti contiene il numero di transazione tra gli attributi chiave – ma non è un attributo chiave esterna
- le dimensioni degeneri sono normali, comuni e utili
  - in particolare, comuni nella progettazione di tabelle fatti orientate a “elementi” (linee o righe) di transazioni

Schema delle vendite in azione



Schema delle vendite in azione

Lo schema dimensionale della catena di negozi memorizza i seguenti fatti relativi alle vendite

- numero di pezzi venduti (**sales_quantity**)
- incasso in dollari (**sales_dollar_amount**)
- costo in dollari (**cost_dollar_amount**)
- profitto lordo in dollari (**gross_profit_dollar_amount**)
- questi fatti sono additivi rispetto a tutte le dimensioni

Che cosa possiamo calcolare?

Il *profitto lordo* (per riga di vendita, prodotto, giorno e negozio)

- è additivo rispetto a tutte le dimensioni
- aggregandolo, è possibile ad esempio calcolare il profitto lordo giornaliero, per negozio/città, e/o per categoria di prodotto

Il *margine lordo* può essere calcolato dividendo il profitto lordo per l'incasso

- per ogni possibile aggregazione, il margine lordo può essere calcolato prima sommando tutti gli incassi e i costi e poi effettuando la divisione
- alcuni fatti non additivi (calcolati da fatti additivi) possono essere aggregati – ricordandosi però di distribuire correttamente le operazioni

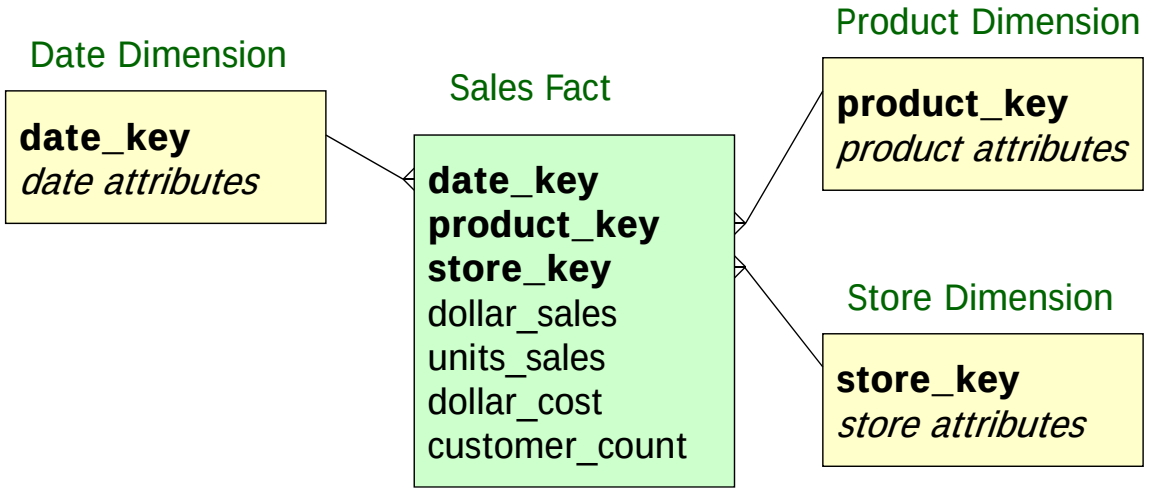
Che cosa possiamo calcolare?

Lo schema dimensionale mostrato consente inoltre – tramite strumenti di *data mining* – di effettuare le seguenti *analisi di market basket* – utili per prendere utili decisioni di marketing

- *associazioni*
  - analisi delle combinazioni di prodotti più comunemente acquistati insieme – richiede la dimensione "numero di transazione"
- *pattern sequenziali*
  - analisi delle regolarità nel comportamento dei clienti in una sequenza di acquisti – richiede transazioni non anonime, e dunque anche una dimensione "cliente"
- *classificazione*
  - dei clienti o dei prodotti – in classi predefinite
- *clustering*
  - suddivide clienti o prodotti in gruppi (classi) simili

Schema delle vendite in azione

Cosa dire invece a proposito del data mart delle vendite giornaliere, per prodotto e negozio?



Schema delle vendite in azione

In questo caso, i fatti a disposizione sono

- 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
- il numero di clienti è invece un fatto semi-additivo

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

La dimensione promozione

Uno degli scopi del data mart delle vendite al dettaglio è abilitare la comprensione dell'impatto delle varie promozioni sulle vendite

- analisi di questo tipo possono essere effettuate mediante l'introduzione di una dimensione supplementare per le promozioni
- inoltre, è necessario usare degli opportuni modelli matematici (fuori della portata di questo corso) per valutare l'andamento "base" delle vendite – in un'ipotetica assenza di promozioni – sulla base dei dati storici sulle vendite
- l'impatto di ciascun tipo di promozione può essere poi valutato in termini di crescita/decrescita delle vendite rispetto a questo andamento base
- è anche di interesse valutare se una promozione ha cambiato le abitudini dei clienti (anche solo subito prima o subito dopo la promozione) e come è cambiato il profitto complessivo delle vendite

Effetto del tempo atmosferico sulle vendite

Illustriamo l'intuizione necessaria per comprendere l'effetto delle promozioni sulle vendite con un esempio un po' diverso

Come analizzare l'impatto del tempo atmosferico sulle vendite?

- aggiungo una dimensione supplementare Tempo atmosferico
  - con membri come "bel tempo", "nuvoloso", "pioggia", ...
- utilizzo una sorgente di dati esterni per sapere quale è stato il tempo in ciascun giorno (del periodo di riferimento) nelle città in cui ci sono negozi della catena
- aggiungo a ciascuna riga della tabella fatti una chiave esterna verso la dimensione Tempo atmosferico
  - non lo faccio a mano, basta un programmino
- faccio l'analisi utilizzando degli opportuni modelli matematici
  - scopro l'eventuale impatto del tempo sulle vendite

La dimensione promozione

La dimensione Promozione – e così la dimensione Tempo atmosferico – 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

La dimensione promozione

Come introdurre una dimensione Promozione nel nostro schema dimensionale?

- da una parte, bisogna capire quali sono i suoi membri
- d'altra parte, bisogna poi capire a quale particolare membro di Promozione va collegata ciascuna riga della tabella fatti delle vendite

Promozioni

Le promozioni sui prodotti possono essere di diversi tipi

- riduzioni temporanee di prezzo (TPR)
- pubblicità (su diversi media)
- esposizione opportuna – sugli scaffali o alla fine dei corridoi
- buoni sconto

Inoltre, le diverse modalità di promozione possono essere applicate contemporaneamente

- ad esempio, riduzione temporanea del prezzo, pubblicità sui giornali e esposizione alla fine dei corridoi
- 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

Membri della dimensione promozione

Sulla scelta dei membri della dimensione Promozione

- una possibile scelta è che ogni membro di Promozione descriva una particolare combinazione delle modalità di promozione applicate in un certo periodo di tempo
  - ad es., riduzione temporanea del prezzo e pubblicità sui giornali
- queste combinazioni sono più numerose delle promozioni individuali applicate – ma non molto più numerose
  - 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)
- questa scelta potrebbe essere più vantaggiosa che non avere, ad esempio, quattro dimensioni diversi per i quattro tipi di promozioni che possono essere applicate

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

Collegare i fatti alle promozioni

Bisogna anche stabilire a quale particolare membro di Promozione va collegata ciascuna riga della tabella fatti delle vendite

- due possibili scelte
  - se una riga di vendita di uno scontrino è relativa a un prodotto in promozione, allora la riga della tabella fatti corrispondente va collegata a quella promozione
  - se una vendita avviene in un periodo di promozione – indipendentemente o meno dal fatto che il prodotto acquistato sia in promozione – allora le righe della tabella fatti corrispondenti vanno collegate a quella promozione
- in ogni caso, nella tabella Promozione va introdotto un membro speciale che rappresenta “nessuna promozione” – per evitare chiavi esterne nulle nella tabella fatti

Dimensioni “junk”

Le promozioni sono basate su quattro meccanismi causali

- riduzione di prezzo, pubblicità, esposizione, buoni sconto

La promozione è una sola dimensione – oppure 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
- l’uso di una sola dimensione porterebbe ad una “dimensione di scarti” – una **junk dimension**
  - ovvero, una raggruppamento di comodo di attributi (più o meno) casuali

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

Discussione

Alcune osservazioni sui modelli dimensionali

- gli schemi dimensionali, a fronte di alcuni cambiamenti, sono facilmente estensibili
  - ad es., aggiunta di nuovi attributi nelle dimensioni, aggiunta di nuove dimensioni supplementari
  - altri cambiamenti, invece possono avere un impatto più significativo su uno schema dimensionale
- evita schemi altamente normalizzati (snowflaking) – ci sono svantaggi nelle prestazioni, e nessun vantaggio significativo
- la presenza di troppe dimensioni è indice di una modellazione non corretta delle dimensioni – alcune possono essere combinate
- usa sempre chiavi surrogate – per le prestazioni – perché il ciclo di vita delle chiavi naturali non è di solito compatibile con la memorizzazione di dati storicizzati