Sistemi Informativi Avanzati
Anno Accademico 2015/2016
Prof. Domenico Beneventano

Modellazione concettuale

Dal Capitolo 5 del libro
Data Warehouse - teoria e pratica della Progettazione
Autori: Matteo Golfarelli, Stefano Rizzi;
Editore: McGraw-Hill

Quale formalismo?

- Mentre è universalmente riconosciuto che un DW si appoggia sul modello multidimensionale, non c'è accordo sul formalismo di modellazione concettuale e quindi sulla metodologia di progettazione concettuale.

- Il modello Entity/Relationship è molto diffuso nelle imprese come formalismo per la documentazione dei sistemi informativi relazionali, ma *non può essere usato per modellare il DW*.

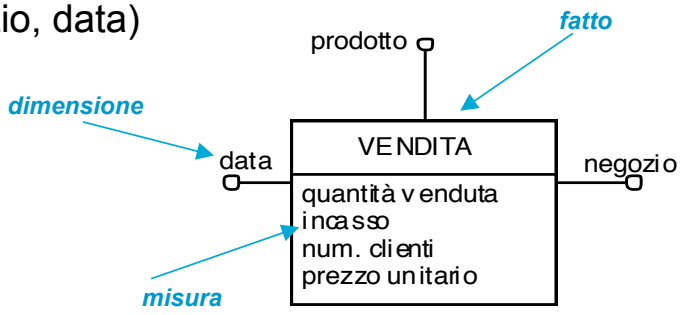
Il Dimensional Fact Model (DFM)

- E' un modello concettuale grafico per data mart, pensato per:
  - ✓ supportare efficacemente il progetto concettuale;
  - ✓ creare un ambiente su cui formulare in modo intuitivo le interrogazioni dell'utente;
  - ✓ permettere il dialogo tra progettista e utente finale per raffinare le specifiche dei requisiti;
  - ✓ creare una piattaforma stabile da cui partire per il progetto logico (*indipendentemente dal modello logico target*);
  - ✓ restituire una documentazione a posteriori espressiva e non ambigua.
- La rappresentazione concettuale generata dal DFM consiste in un insieme di schemi di fatto. Gli elementi di base modellati dagli schemi di fatto sono i fatti, le misure, le dimensioni e le gerarchie

Il DFM: costrutti di base

- Un *fatto* è un concetto di interesse per il processo decisionale; tipicamente modella un insieme di eventi che accadono nell'impresa (ad esempio: vendite, spedizioni, ...). È essenziale che un fatto abbia aspetti dinamici, ovvero evolva nel tempo
- Una *misura* è una proprietà numerica di un fatto e ne descrive un aspetto quantitativo di interesse per l'analisi (ad esempio, ogni vendita è misurata dal suo incasso)
- Una *dimensione* è una proprietà con dominio finito di un fatto e ne descrive una coordinata di analisi (dimensioni tipiche per il fatto vendite sono prodotto, negozio, data)

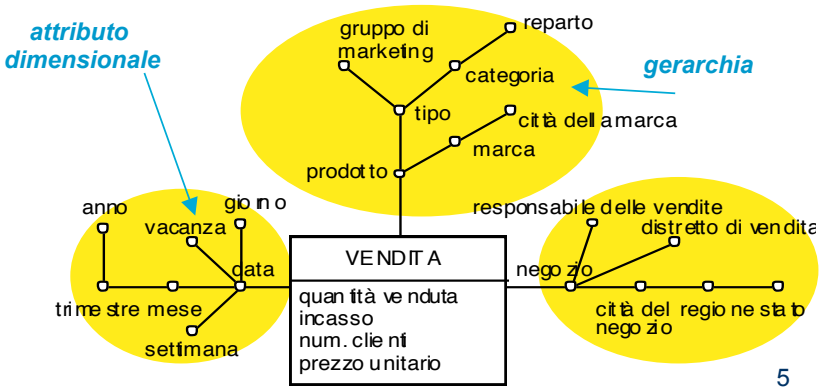
Un fatto esprime una associazione molti-a-molti tra le dimensioni



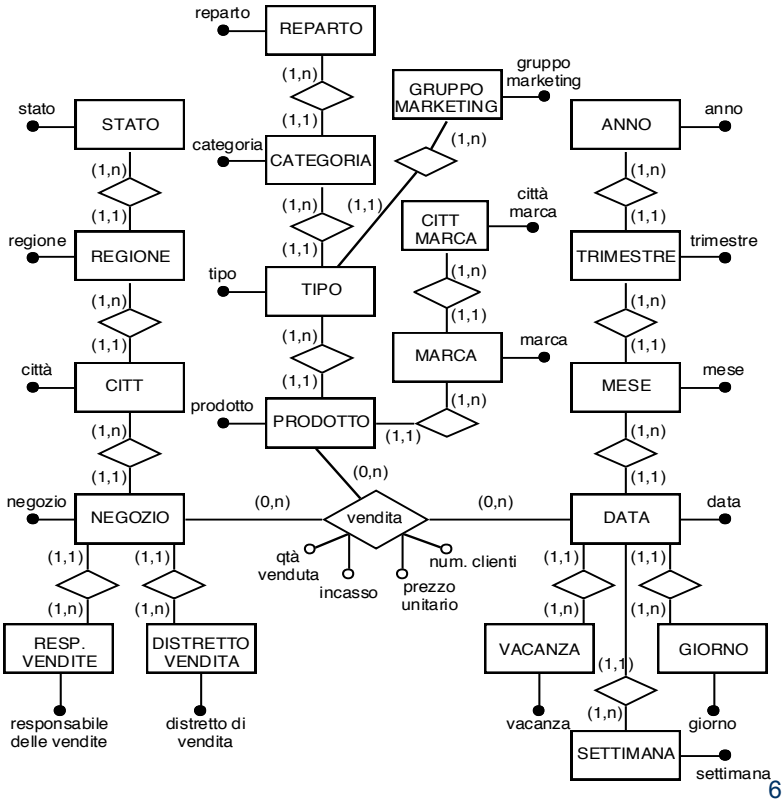
Il DFM: costrutti di base

- Con *attributo dimensionale* si intendono le dimensioni e gli altri attributi, che le descrivono (per esempio, un prodotto è descritto dal suo tipo, dalla categoria cui appartiene, dalla sua marca, dal reparto in cui è venduto)
- Una *gerarchia* è un albero direzionato i cui nodi sono attributi dimensionali e i cui archi rappresentano associazioni molti-a-uno tra coppie di attributi dimensionali:
  l'arco da X a Y rappresenta la dipendenza funzionale X→Y

- La gerarchia racchiude una dimensione, posta alla radice dell'albero, e tutti gli attributi dimensionali che la descrivono



Il DFM: corrispondenza con l'E/R



"Naming conventions"

- Tutti gli attributi dimensionali in ciascuno schema di fatto dovrebbero avere nomi diversi
- Eventuali nomi uguali dovrebbero essere differenziati qualificandoli con il nome di un attributo dimensionale che li precede nella gerarchia
  - ✓ Ad esempio, *warehouse city* è la città in cui si trova un magazzino, mentre *store city* è la città in cui si trova un negozio
- I nomi degli attributi non dovrebbero riferirsi esplicitamente al fatto a cui appartengono
  - ✓ Ad esempio, si evitino *shipped product* e *shipment date*

Eventi primari e dimensioni

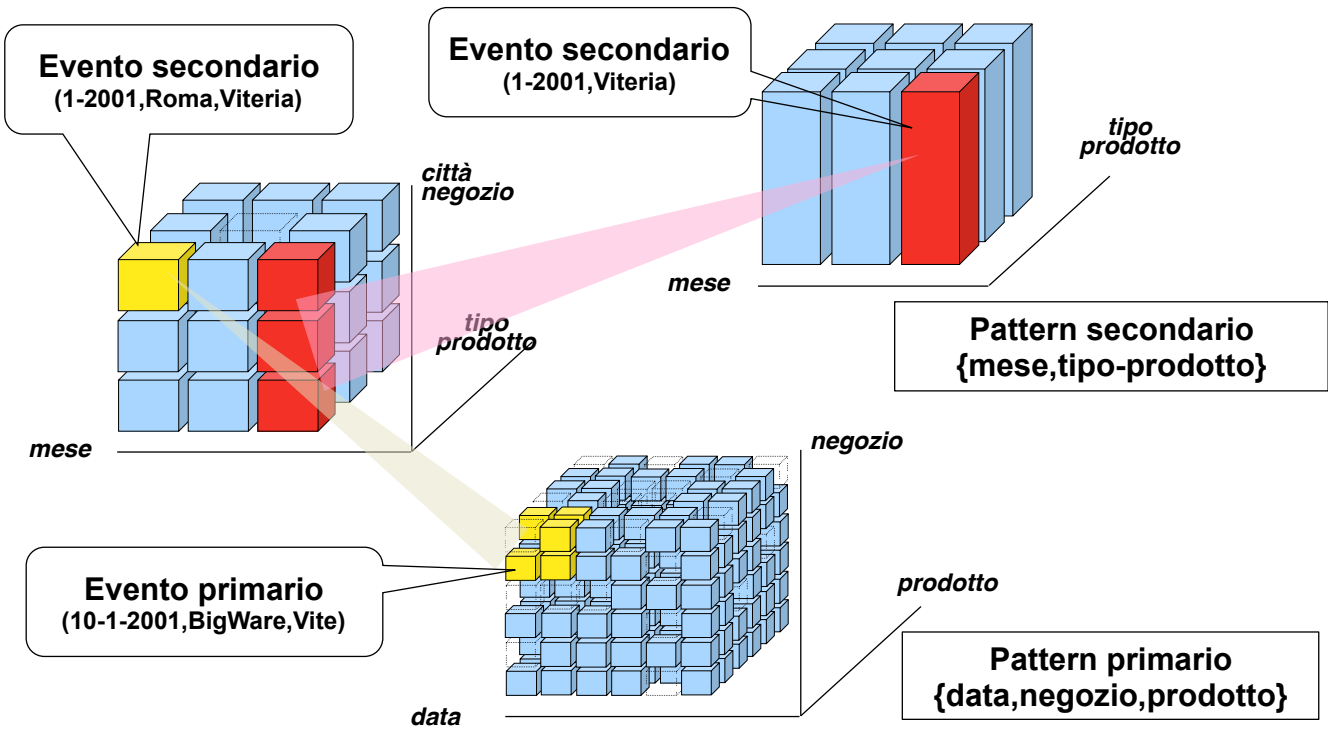
- Un *evento primario* è una particolare occorrenza di un fatto, individuata da una ennupla costituita da un valore per ciascuna dimensione. A ciascun evento primario è associato un valore per ciascuna misura
  - ✓ Nelle vendite, un possibile evento primario registra per esempio che, il 10/10/2001, nel negozio NonSoloPappa sono state vendute 10 confezioni di detersivo Brillo per un incasso complessivo di 25 euro
  - ✓ Un fatto F con *n* dimensioni Dim_1, …, Dim_n e k misure Mis_1, …, Mis_k si può considerare come una relazione
    $$F(\underline{Dim_1, \ldots, Dim_n}, Mis_1, \ldots, Mis_k)$$
    che ha come chiave $D = \{ Dim_1, \ldots, Dim_n \}$
    quindi ciascuna misura dipende funzionalmente da D
  - ✓ Questo parallelo con il modello relazionale ci consente di parlare di dipendenze funzionali tra le dimensioni, tra le dimensioni e le misure …

Eventi secondari e pattern

- Dato un insieme di attributi dimensionali (*pattern*), ciascuna ennupla di loro valori individua un *evento secondario* che aggrega tutti gli eventi primari corrispondenti. A ciascun evento secondario è associato un valore per ciascuna misura, che riassume in sé tutti i valori della stessa misura negli eventi primari corrispondenti
  - ✓ Pertanto, le gerarchie definiscono il modo in cui gli eventi primari possono essere aggregati e selezionati significativamente per il processo decisionale; mentre la dimensione in cui una gerarchia ha radice ne definisce la granularità più fine di aggregazione, agli altri attributi dimensionali corrispondono granularità via via crescenti

- Pattern Primario e Pattern Secondari
  - ✓ **Pattern primario**: è il pattern formato dall'insieme delle dimensioni
  - ✓ **Pattern secondario**: è un qualsiasi altro pattern diverso dal primario, ovvero contenente almeno un attributo dimensionale che non è una dimensione

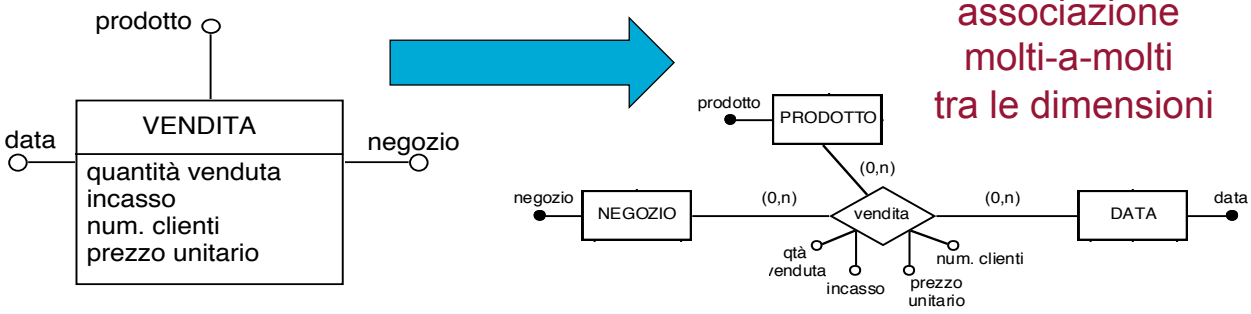
Eventi e aggregazione



Corrispondenza tra DFM ed E/R

- Utile per spiegare la semantica dei costrutti del modello DFM a partire da quella del modello E/R (e quindi del modello relazionale)

Un fatto esprime una associazione molti-a-molti tra le dimensioni

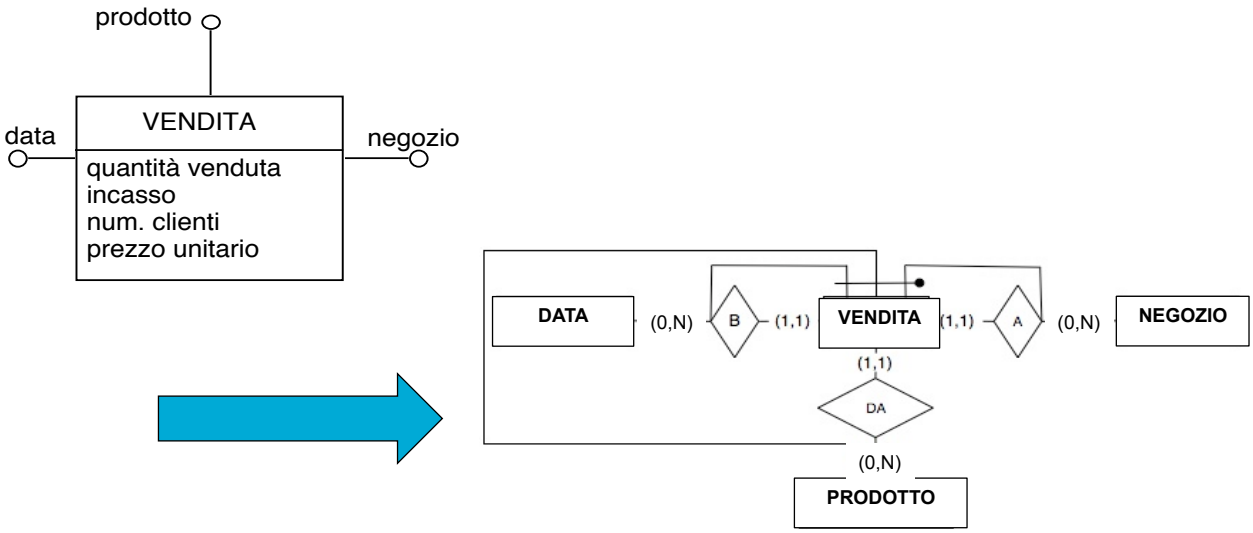


Le istanze sono gli **eventi primari** (si riporta solo Incasso)

Prodotto	Negozio	Data	Incasso
P1	A	18/4	13
P2	B	18/4	12

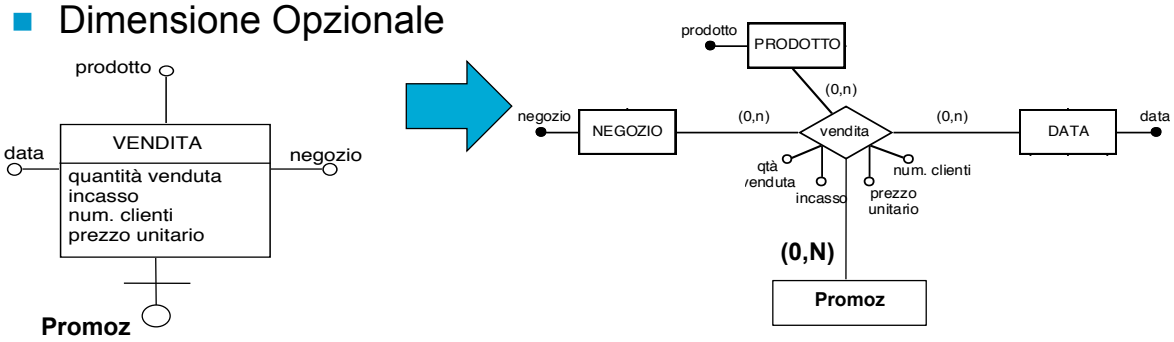
Corrispondenza tra DFM ed E/R

- L'associazione molti-a-molti VENDITA si può anche esprimere, in modo del tutto equivalente, in ***forma reificata***



Corrispondenza tra DFM ed E/R

- Dimensione Opzionale



- Corrispondenza *non del tutto esatta*: nello schema di fatto **Promoz** è opzionale (ci sono vendite senza Promoz) mentre in E/R Vendita è un'associazione che necessita anche di Promoz! Per una corrispondenza esatta si dovrebbe reificare VENDITA e specializzare in VENDITA_IN_PROMO con Promoz …
  - … si semplifica considerando un particolare valore di Promoz, cioè l'opzionalità viene **codificata** con un opportuno valore a livello di eventi primari

In un DW i **valori nulli** derivanti dalle opzionalità **sono codificati**.

PROMOZ	Prodotto	Negozi	Data	Incasso
Estate	P1	A	18/4	13
NO_PROMO	P2	B	18/4	12
Estate	P2	B	18/4	12

Modellazione e progettazione concettuale

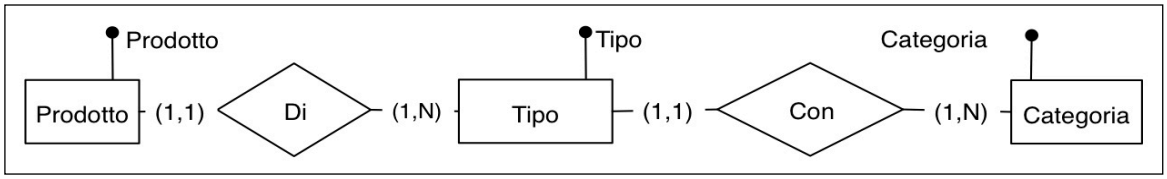
- **Modellazione:** sintassi e semantica del modello DFM
  - ✓ la semantica dei costrutti del modello DFM viene spiegata a partire dalla semantica del modello E/R

- **Progettazione:** metodi per progettare uno schema secondo il modello DFM
  - ✓ Progettazione da schemi E/R: dato uno schema E/R ed **i requisiti** del Data Warehouse, progettare lo schema di fatto
  - ✓ Le scelte fondamentali che deve fare il progettista sono
    1. **Dimensioni** (granularità)
    2. **Misure** e relativi **operatori di aggregazione**
    3. **Gerarchia** associata a ciascuna dimensione

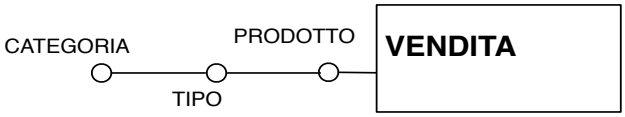
Esempio di progettazione concettuale

- Consideriamo la gerarchia di PRODOTTO in VENDITA

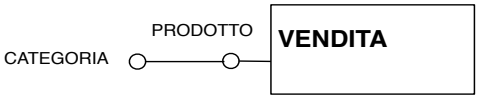
DBO



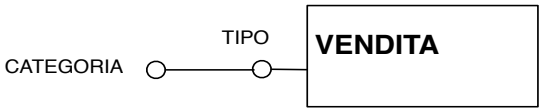
- Lo schema di fatto può riportare tutta la gerarchia …



- … oppure si decide che TIPO non è utile ai fini dell'analisi:



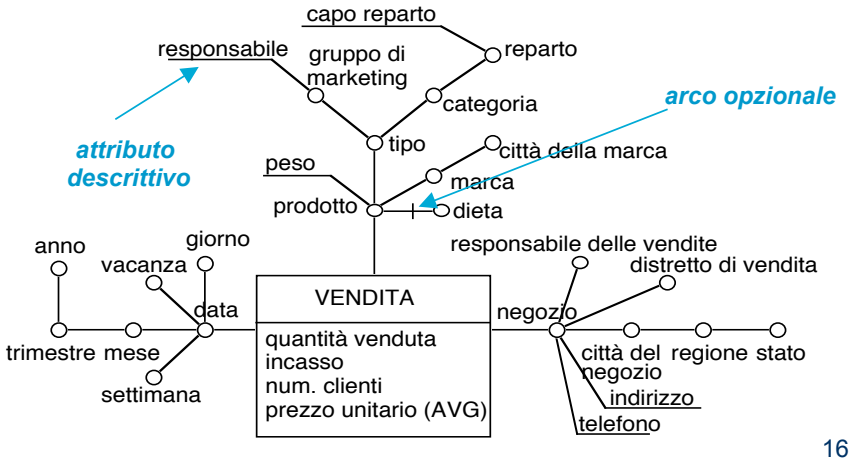
- … oppure si sceglie una granularità meno fine, non considerando PRODOTTO (schema di fatto *temporale*):



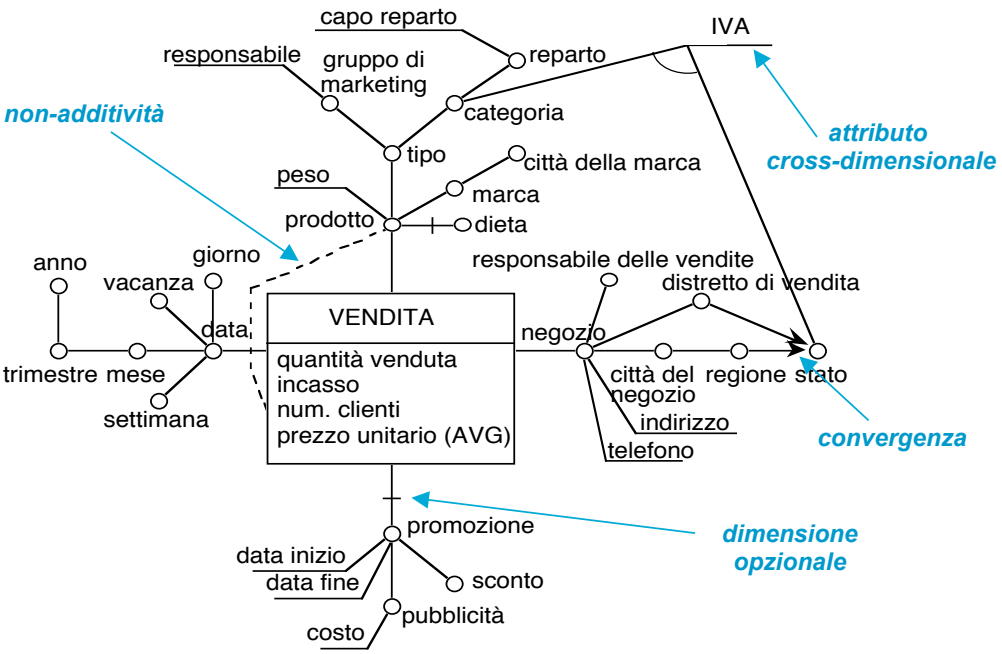
Il DFM: costrutti avanzati

- Un *attributo descrittivo* contiene informazioni aggiuntive su un attributo dimensionale, a cui è connesso da una associazione uno-a-uno. Non è usato per l'aggregazione poiché ha valori continui e/o poiché deriva da un'associazione uno-a-uno

- Alcuni archi dello schema di fatto possono essere *opzionali*



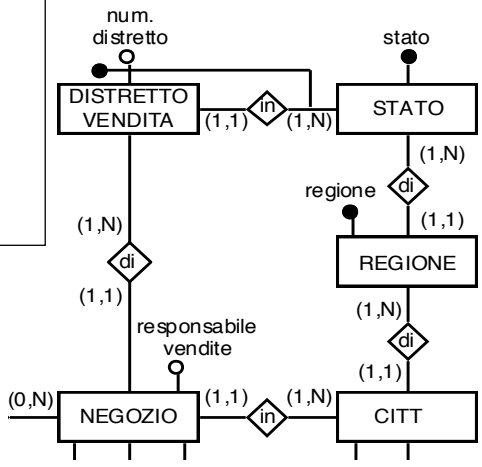
Il DFM: costrutti avanzati



Convergenza

Vincolo di integrità (non esprimibile in E/R):
"lo stato della città del negozio deve essere lo stesso di quello del distretto del negozio"

Questa informazione è esprimibile sullo **schema di fatto** indicando una **convergenza**:
la convergenza rappresenta un vincolo di integrità.

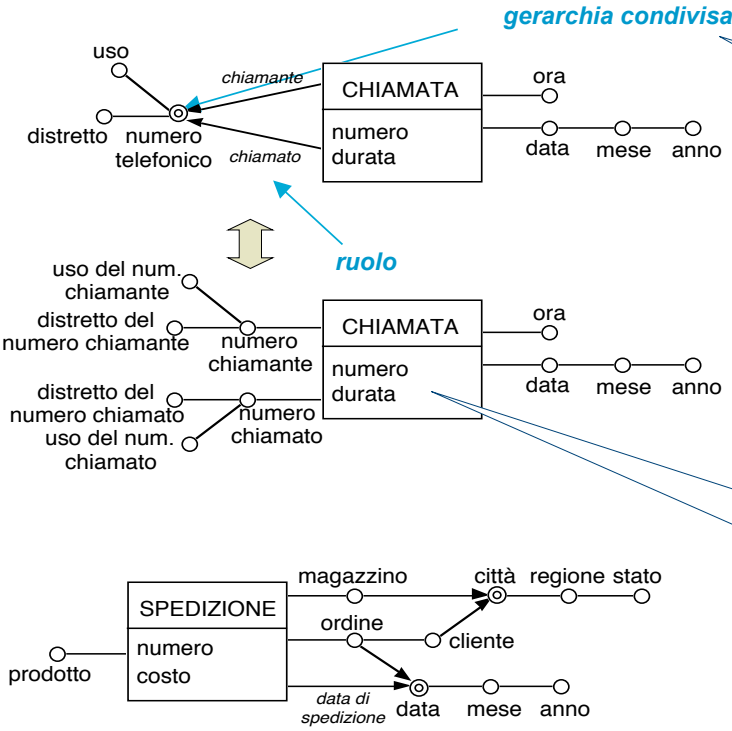


Il Pattern
{ Negozio.DistrettoVendite.Stato,
Negozio.Citta.Regione.Stato }

ammette come eventi secondari solo coppie di valori uguali:
si considera solo {Stato }

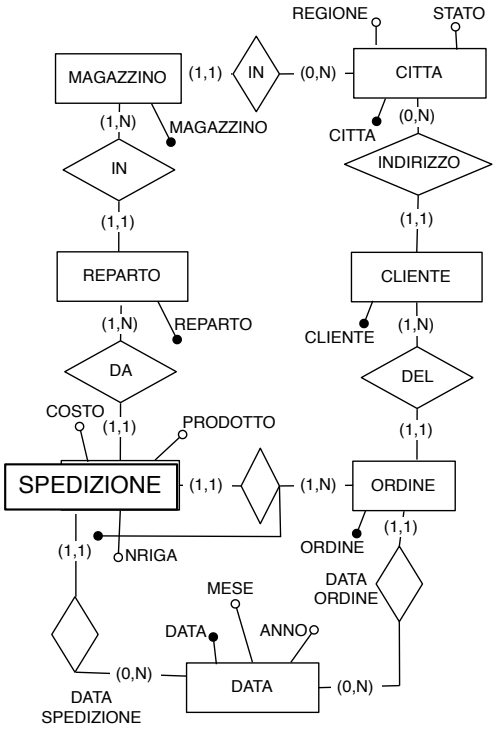
In un pattern con un attributo
Di condivisione
per **distinguere** le due occorrenza
occorre **qualificare** con
Il percorso nella gerarchia

Il DFM: costrutti avanzati



La gerarchia è sicuramente condivisa:
il numero del chiamante deve essere diverso dal numero del chiamato

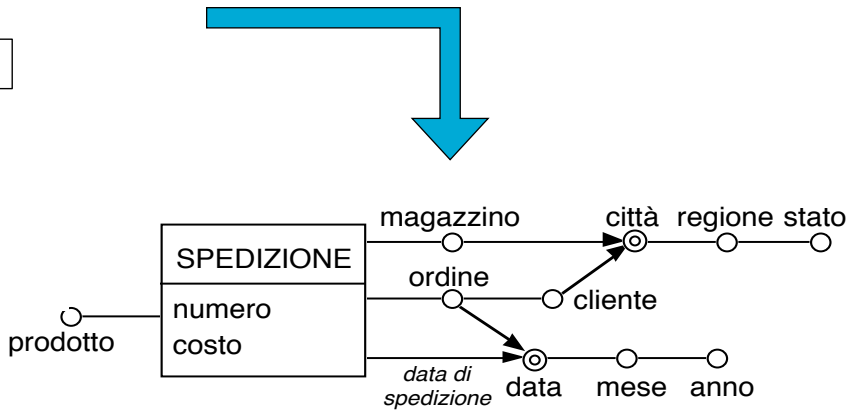
È il numero di chiamate, mentre la durata è quella complessiva.



PROGETTAZIONE CONCETTUALE

Basata sulle sorgenti

Requisiti : Fatto VENDITA con DIMENSIONI
- Magazzino
- Ordine
- Prodotto
- DataDiSpedizione

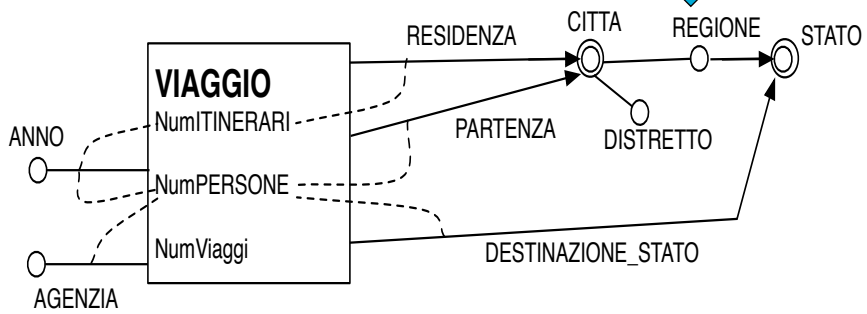


Schema di Fatto: Esempio

- Esempio di progettazione dallo schema relazionale del DB operazionale

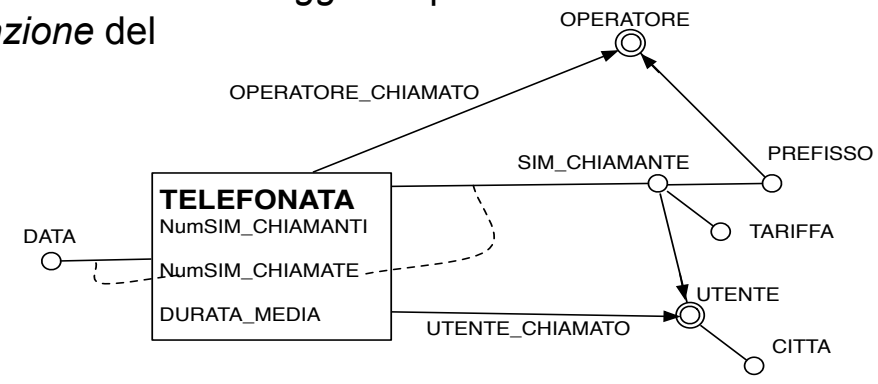
```
VIAGGIO(PERSONA:PERSONA,DATA, ITINERARIO:ITINERARIO)
ITINERARIO(ITINERARIO,PARTENZA:CITTA,DESTINAZIONE:CITTA,AGENZIA,TIPO)
        FD: PARTENZA, DESTINAZIONE → AGENZIA
PERSONA(PERSONA, RESIDENZA:CITTA)
CITTA(CITTA, REGIONE:REGIONE, DISTRETTO)
REGIONE(REGIONE, STATO)
```

Requisiti : Fatto VIAGGIO con DIMENSIONI
ANNO
AGENZIA
STATO DI DESTINAZIONE
CITTA DI RESIDENZA
CITTA DI PARTENZA



Schema di Fatto: Esempio

- Uno Schema di Fatto è facilmente *leggibile* quindi ha anche un ruolo di *documentazione* del DataWarehouse



- Dimensioni
  1. **DATA**
  2. **UTENTE_CHIAMATO**
  3. **SIM_CHIAMANTE**
  4. **OPERATORE_CHIAMATO**

- Siccome OPERATORE è condiviso, la dimensione OPERATORE_CHIAMATO viene indicata come ruolo. (stesso discorso per UTENTE_CHIAMATO)

Archi e dimensioni opzionali

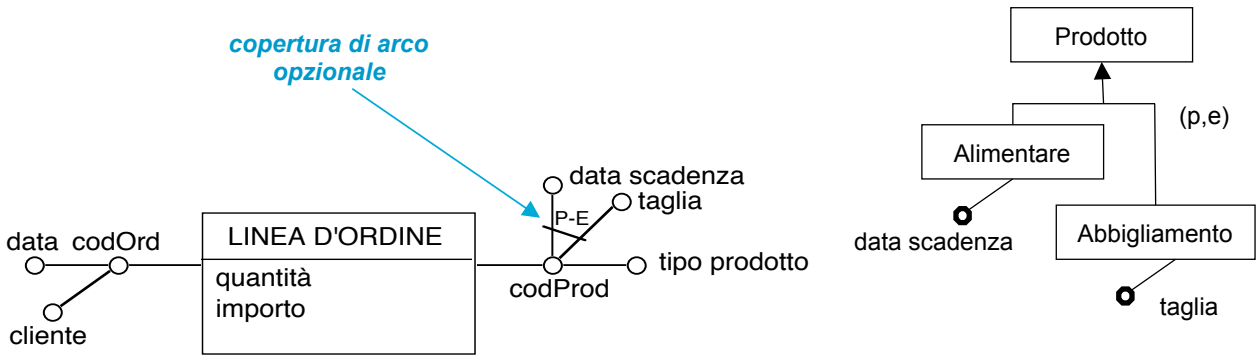
- Derivano da una cardinalità minima pari a zero nelle associazioni, ovvero da *associazioni opzionali*

- **Arco Opzionale** Prodotto →Dieta
  - ✓ Un Prodotto ha una sola Dieta; per alcuni prodotti la dieta è indefinita, ovvero assume un valore NULL
  - ✓ A livello di analisi dei dati, normalmente tale valore NULL viene “rappresentato” con un valore significativo, quale ‘NESSUNA DIETA’

- **Dimensione Opzionale** Promozione
  - ✓ Un evento primario è identificato dalle dimensioni; se una dimensione è opzionale alcuni eventi primari sono identificati solo dalle altre dimensioni: le vendite senza promozione sono identificate da prodotto-negozio-data e con un valore significativo per Promozione, es. ‘NESSUNA PROMOZIONE’

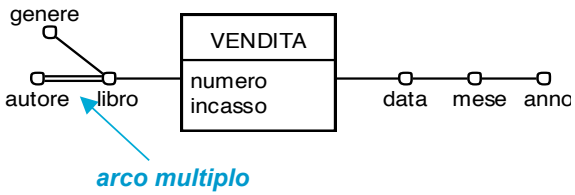
- **L’opzionalità si propaga ai discendenti nella gerarchia**
  - ✓ Gli eventi senza promozione non hanno sconto
  - ✓ A livello di analisi dei dati, al valore ‘NESSUNA PROMOZIONE’ faremo corrispondere ‘NESSUN SCONTO’ per l’attributo dimensionale sconto.

Copertura di un arco opzionale



- La proprietà di copertura influisce sul numero di eventi secondari ammissibili
  - ✓ Il Pattern {Data scadenza, Taglia} in caso di copertura esclusiva non ammette eventi secondari

Arco Multiplo



- Un arco multiplo corrisponde ad un'associazione molti-a-molti: il padre (libro) non determina funzionalmente il figlio (autore)
  - ✓ Nell'esempio si aggregano le vendite dei libri sulla base dei loro autori: un libro è *scritto* da più autori quindi non si può associare ad un unico autore

- Gli archi multipli verranno trattati a parte: in particolare si vedrà che per definire in modo consistente l'aggregazione anche per gli archi multipli sia a volte necessario definire un "peso"
  - ✓ Nell'esempio delle vendite di un libro, il "peso" stabilisce la percentuale dell'incasso di un libro che deve essere attribuita a ciascuno dei suoi autori

Dipendenze funzionali tra dimensioni

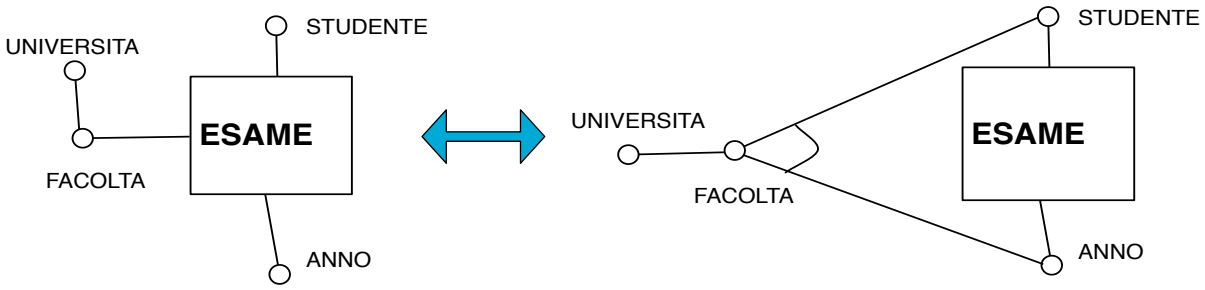
- Una **FD tra le dimensioni** si ha quando, dato l'insieme delle dimensioni D,
  esistono due sottoinsiemi X ed Y di D tali che **X → Y**.
  - ✓ ogni misura M dipende solo da X, cioè X → M

- Con FD tra dimensioni il **pattern primario** è **ridondante**

- In questo caso lo schema di fatto F è equivalente ad uno schema di fatto F' con dimensioni X e con le restanti "vecchie" dimensioni in Y sono inclusi come attributi cross-dimensionali determinate da X
  - ✓ Rappresentare gli attributi di Y come dimensioni è comunque più utile per dare maggiore risalto al loro ruolo nell'aggregazione.

Dipendenze funzionali tra dimensioni

- Schema di Fatto Esame con D = {STUDENTE,FACOLTA,ANNO}
  e con FD: {STUDENTE,ANNO} → FACOLTA



- I due schemi sono *"equivalenti"*; nello schema di destra la FD è stata esplicitata grazie al costrutto di attributo cross-dimensionale
- Le FD tra dimensioni influenzano la *Aggregabilità delle Misure*
- L'effetto delle FD tra dimensioni è *visibile* nei pattern/report:

Dipendenze funzionali tra dimensioni

- Pattern/report {STUDENTE,UNIVERSITA,ANNO}

1. **Senza** FD:{STUDENTE,ANNO}→FACOLTA

		ANNO		
		2012	2013	Grand Total
STUDENTE	UNIVERSITA	NESAMI	NESAMI	NESAMI
S2	K2		4	4
	U2	2	2	4
	W2	4		4
	Total	6	6	12
S4	K2	2	4	6
	U2	2		2
	W2		4	4
	Total	4	8	12
Grand Total		10	14	24

2. **Con** FD:{STUDENTE,ANNO}→FACOLTA

in un certo *anno*, uno *studente* ha una sola facoltà e quindi una sola *università*
molte *celle* sono vuote

		ANNO		
		2012	2013	Grand Total
STUDENTE	UNIVERSITA	NESAMI	NESAMI	NESAMI
S2	U2		2	2
	W2	4		4
	Total	4	2	6
S4	K2		4	4
	W2	2		2
	Total	2	4	6
Grand Total		6	6	12

in questo caso per realizzare il pattern {STUDENTE,UNIVERSITA,ANNO} meglio un report bidimensionale {STUDENTE,ANNO}, riportando tra () l'UNIVERSITA

	ANNO		
	2012	2013	Grand Total
STUDENTE	NE	NE	NE
S2	4 (W2)	2 (U2)	6
S4	2 (W2)	4 (K2)	6
Grand Total	6	6	12