Il ciclo di vita del Data Warehouse

Prof. Stefano Rizzi

Perché?

- Molte organizzazioni mancano della necessaria esperienza e capacità per affrontare con successo le sfide implicite nei progetti di data warehousing
- Uno dei fattori che maggiormente minaccia la riuscita dei progetti è la mancata adozione di una approccio metodologico, che minimizza i rischi di insuccesso essendo basato su un'analisi costruttiva degli errori commessi

Fattori di rischio

- ✓ Rischi legati alla gestione del progetto
- ✓ Rischi legati alle tecnologie
- ✓ Rischi legati ai dati e alla progettazione
- ✓ Rischi legati all'organizzazione

- Il rischio di ottenere un risultato insoddisfacente nei progetti di data warehousing è particolarmente alto a causa delle elevatissime aspettative degli utenti
- Nella cultura aziendale contemporanea è infatti diffusissima la credenza che attribuisce al data warehousing il ruolo di panacea
- In realtà una larga parte della responsabilità della riuscita del progetto ricade sulla qualità dei dati sorgente e sulla lungimiranza, disponibilità e dinamismo del personale dell'azienda

Approccio top-down

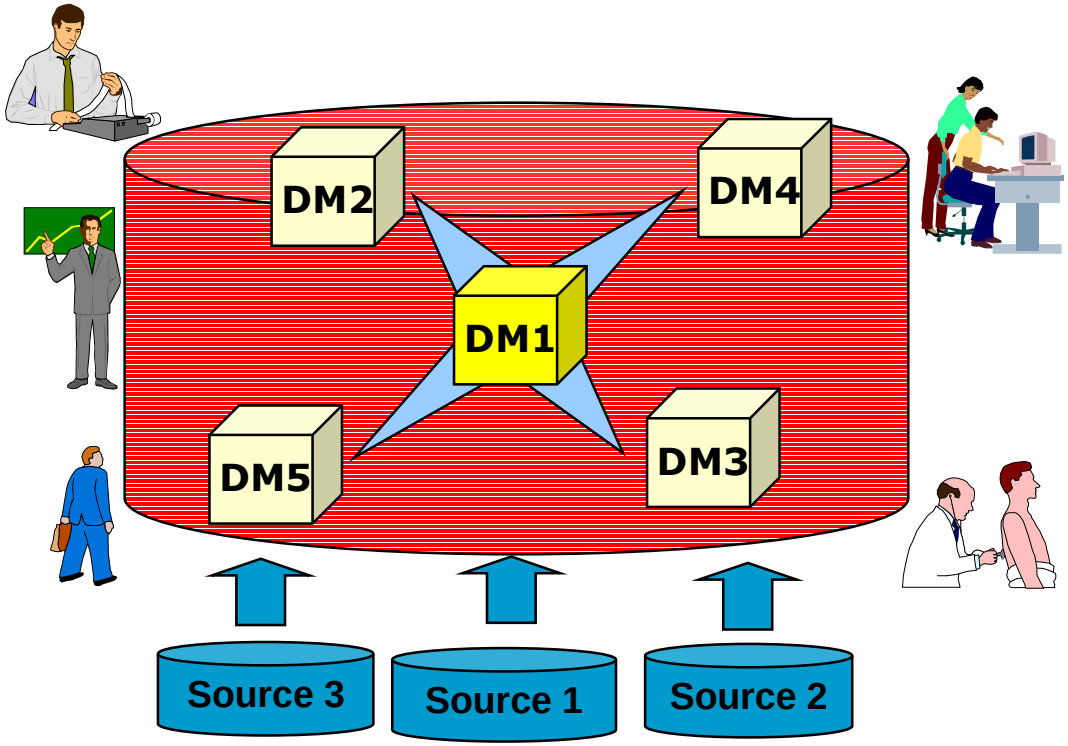
- Analizza i bisogni globali dell'intera azienda e pianifica lo sviluppo del DW per poi progettarlo e realizzarlo nella sua interezza
  - Promette ottimi risultati poiché si basa su una visione globale dell'obiettivo e garantisce in linea di principio di produrre un DW consistente e ben integrato
  - Il preventivo di costi onerosi a fronte di lunghi tempi di realizzazione scoraggia la direzione dall'intraprendere il progetto
  - Affrontare contemporaneamente l'analisi e la riconciliazione di tutte le sorgenti di interesse è estremamente complesso
  - Riuscire a prevedere a priori nel dettaglio le esigenze delle diverse aree aziendali impegnate è pressoché impossibile, e il processo di analisi rischia di subire una paralisi
  - Il fatto di non prevedere la consegna a breve termine di un prototipo non permette agli utenti di verificare l'utilità del progetto e ne fa scemare l'interesse e la fiducia

Approccio bottom-up

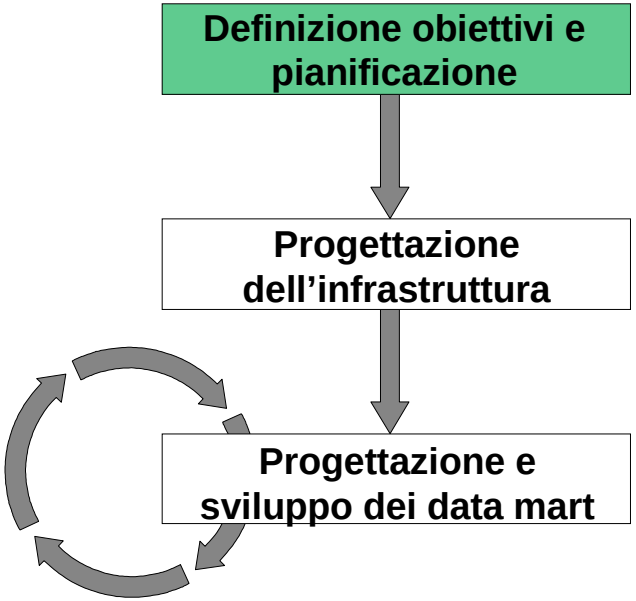
- Il DW viene costruito in modo incrementale, assemblando iterativamente più data mart, ciascuno dei quali incentrato su un insieme di fatti collegati a uno specifico settore aziendale e di interesse per una certa categoria di utenti
  - Determina risultati concreti in tempi brevi
  - Non richiede elevati investimenti finanziari
  - Permette di studiare solo le problematiche relative al data mart in oggetto
  - Fornisce alla dirigenza aziendale un riscontro immediato sull'effettiva utilità del sistema in via di realizzazione
  - Mantiene costantemente elevata l'attenzione sul progetto
  - Determina una visione parziale del dominio di interesse

Il primo data mart da prototipare...

- ✓ deve essere quello che gioca il ruolo più strategico per l'azienda
- ✓ deve ricoprire un ruolo centrale e di riferimento per l'intero DW
- ✓ si deve appoggiare su fonti dati già disponibili e consistenti

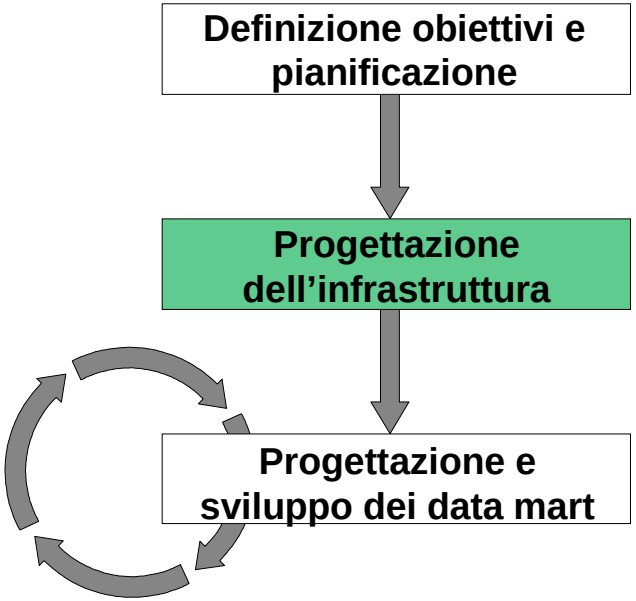


Il ciclo di sviluppo



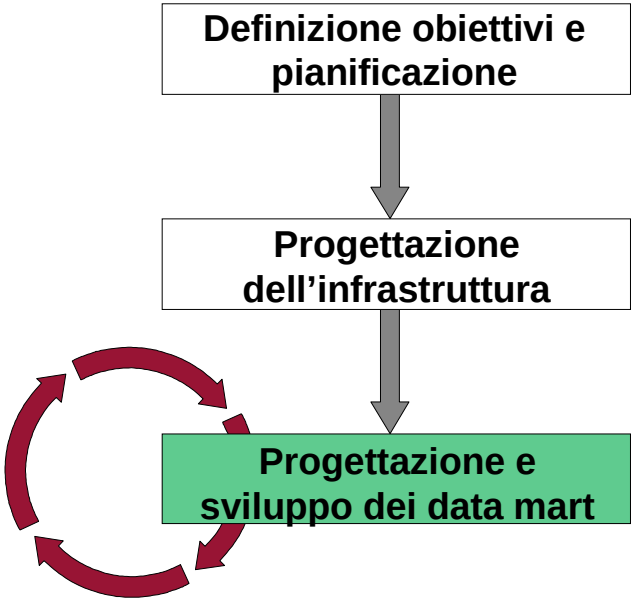
- individuazione degli obiettivi e dei confini del sistema
- stima delle dimensioni
- scelta dell'approccio per la costruzione
- valutazione dei costi e del valore aggiunto
- analisi dei rischi e delle aspettative
- studio delle competenze del gruppo di lavoro

Il ciclo di sviluppo



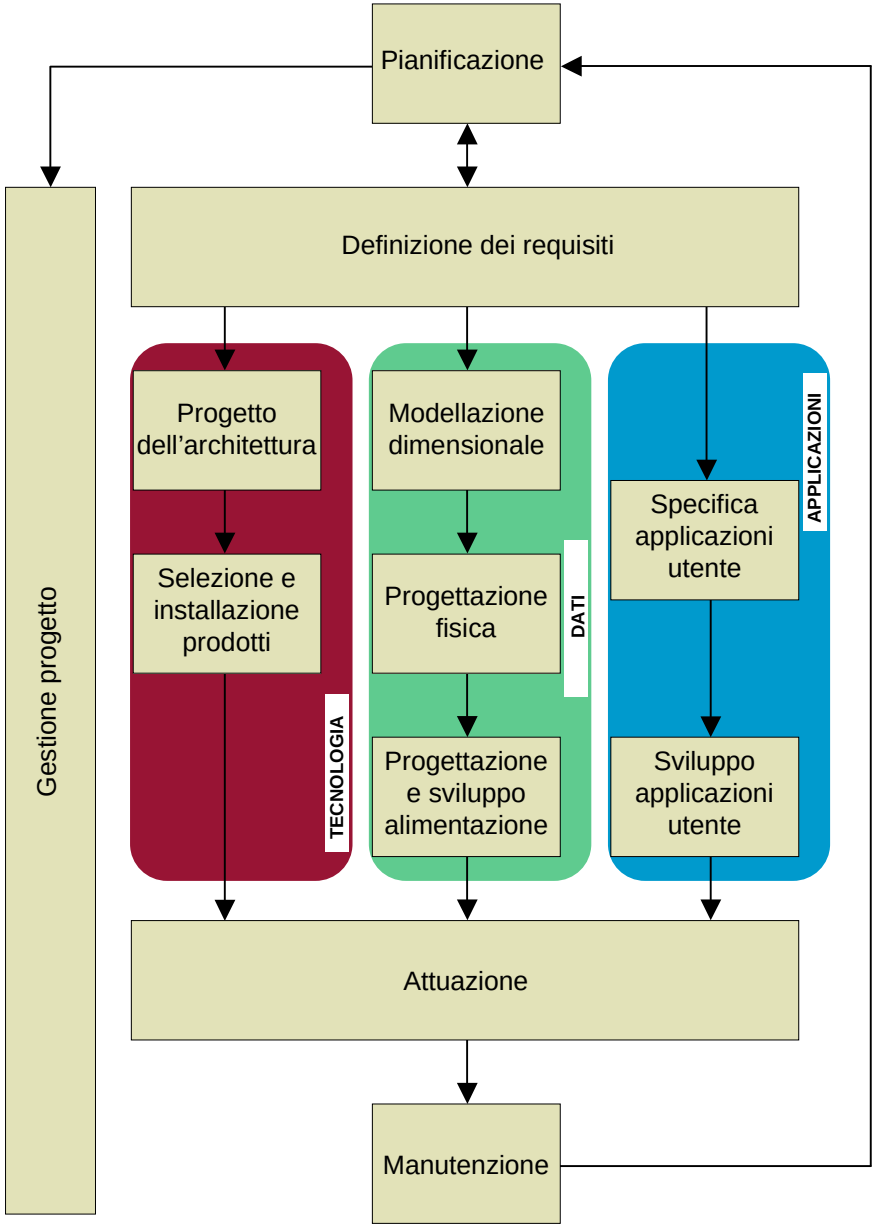
Si analizzano e si comparano le possibili soluzioni architetturali valutando le tecnologie e gli strumenti disponibili, al fine di realizzare un progetto di massima dell'intero sistema.

Il ciclo di sviluppo

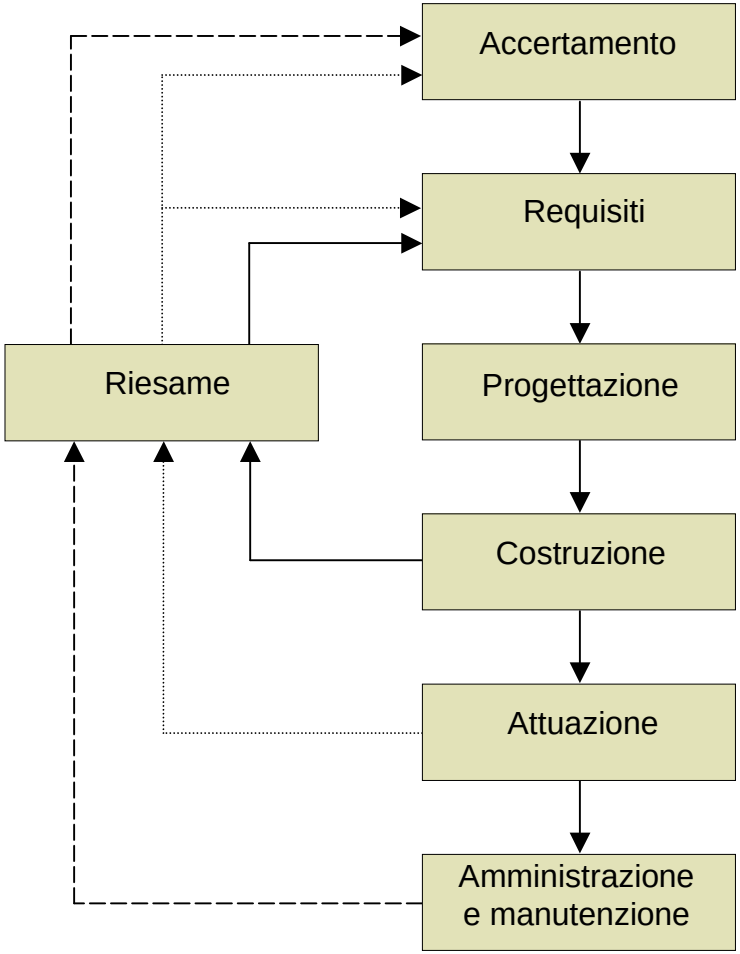


Ciascuna iterazione comporta la creazione di un nuovo data mart e di nuove applicazioni, che vengono via via integrate nel sistema di data warehousing.

Il "Business Dimensional Lifecycle" (Kimball)



La "Rapid Warehousing Methodology" (SAS)



La progettazione del data mart

Analisi e riconciliazione delle sorgenti

amministratore db

Analisi dei requisiti

Progettazione concettuale

utente finale

Raffinamento del carico di lavoro

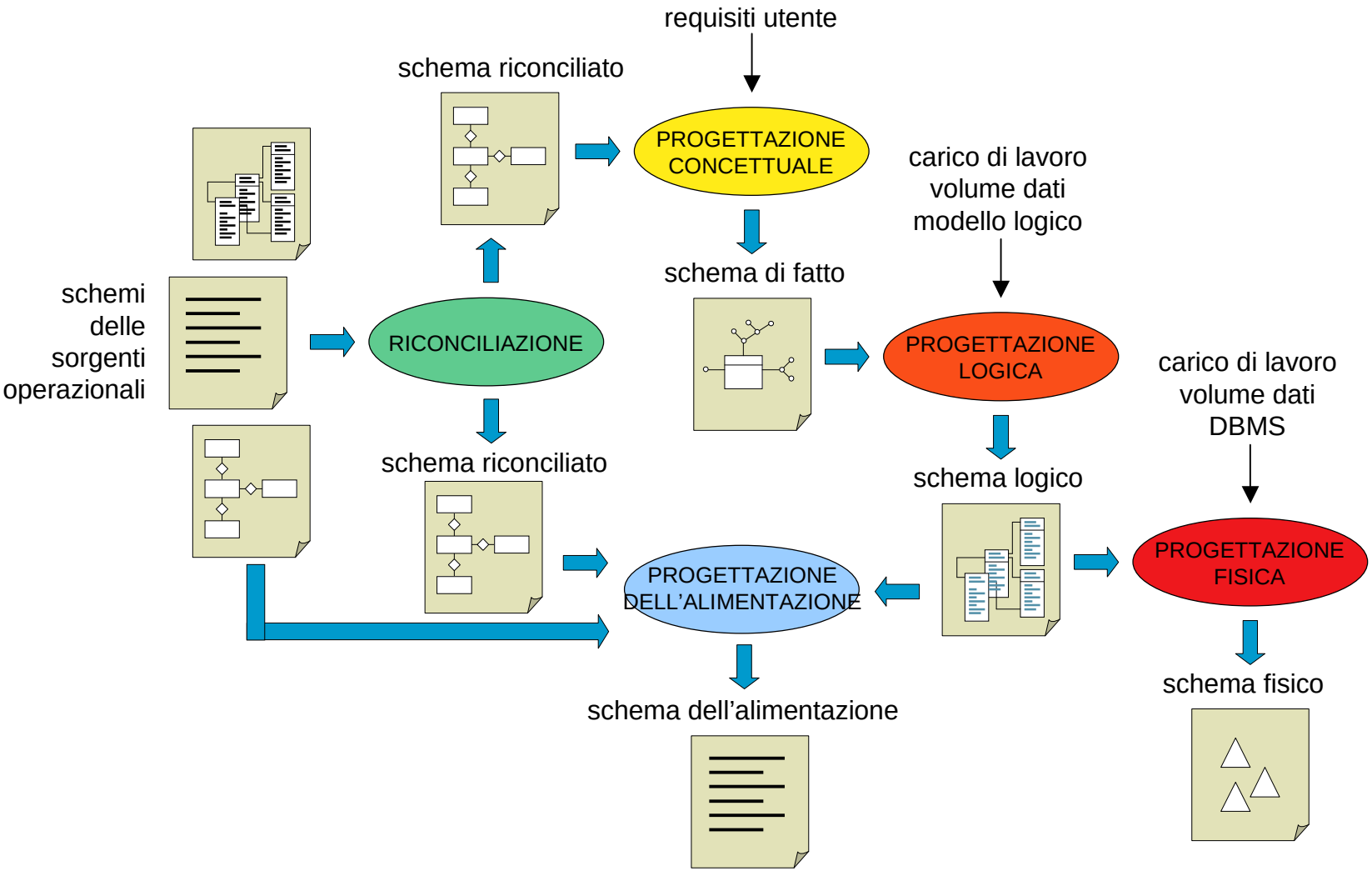
progettista

Progettazione logica

Progettazione dell'alimentazione

Progettazione fisica

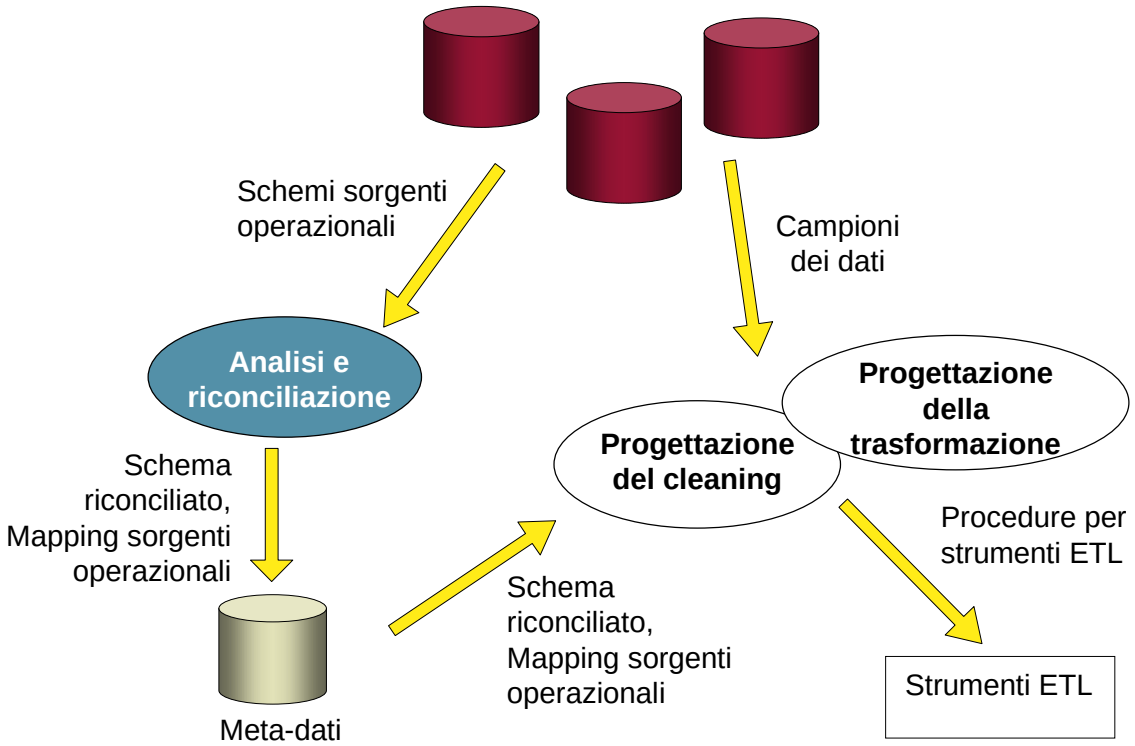
La progettazione del data mart



Analisi e riconciliazione delle sorgenti operazionali

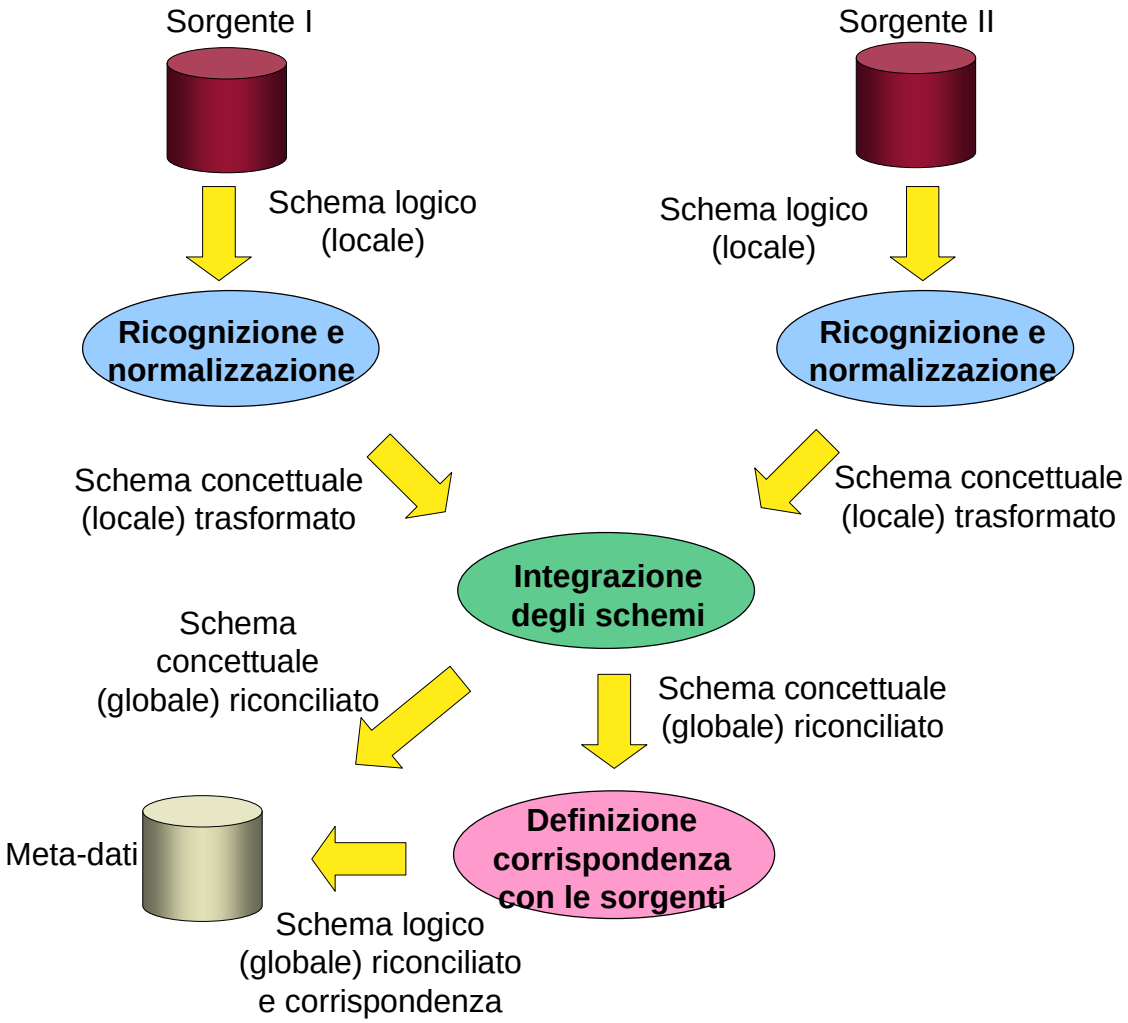
Prof. Stefano Rizzi

Progettazione del livello riconciliato



- ✓ La fase di integrazione è incentrata sulla componente intensionale delle sorgenti operazionali, ossia riguarda la consistenza degli schemi che le descrivono
- ✓ Pulizia e trasformazione dei dati operano a livello estensionale, ossia coinvolgono direttamente i dati veri e propri

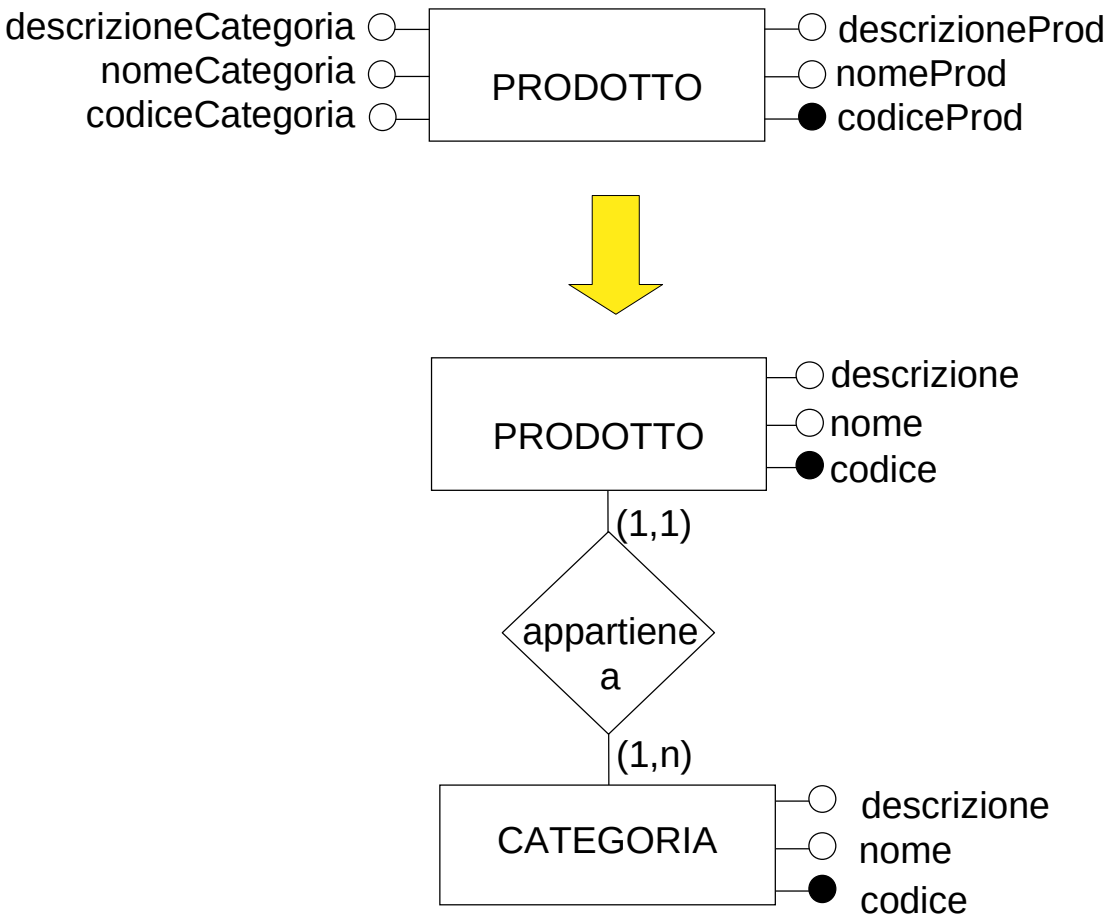
Analisi e riconciliazione delle sorgenti operazionali



Ricognizione e normalizzazione

- Il progettista, confrontandosi con gli esperti del dominio applicativo, acquisisce un'approfondita conoscenza delle sorgenti operazionali attraverso:
  - ✓ *ricognizione*, che consiste in un esame approfondito degli schemi locali mirato alla piena comprensione del dominio applicativo;
  - ✓ *normalizzazione*, il cui obiettivo è correggere gli schemi locali al fine di modellare in modo più accurato il dominio applicativo
- Ricognizione e normalizzazione devono essere svolte anche qualora sia presente una sola sorgente dati; qualora esistano più sorgenti, l'operazione dovrà essere ripetuta per ogni singolo schema

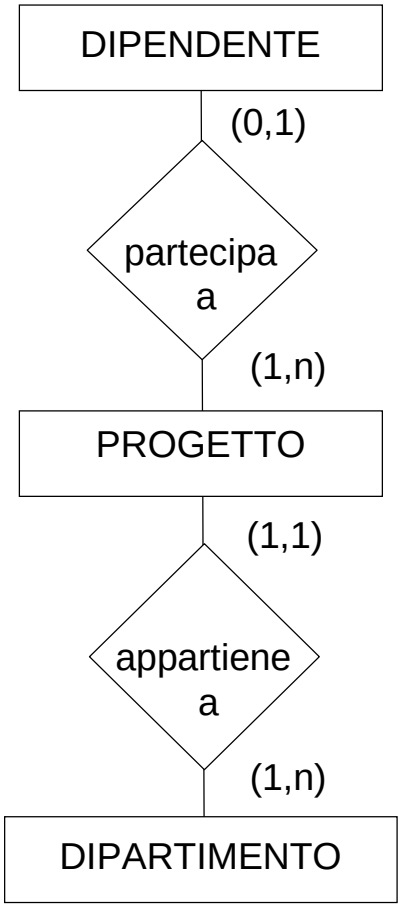
Ricognizione e normalizzazione

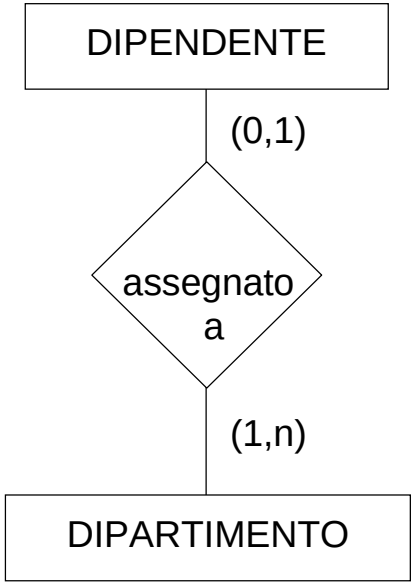


Integrazione

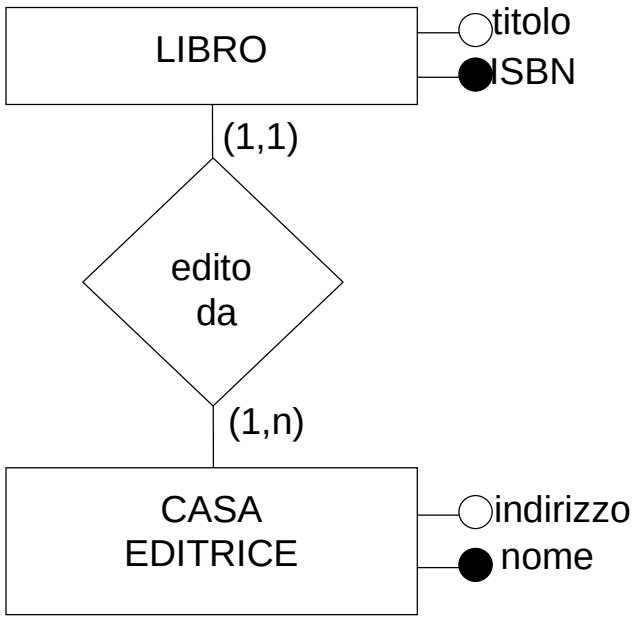
- L'integrazione di un insieme di sorgenti dati eterogenee (basi di dati relazionali, file dati, sorgenti legacy) consiste nell'individuazione delle corrispondenze tra i concetti rappresentati negli schemi locali e nella risoluzione dei conflitti evidenziati, finalizzate alla creazione di un unico schema globale i cui elementi possano essere correlati con i corrispondenti elementi degli schemi locali (*mapping*)
- La fase di integrazione non si deve limitare a evidenziare le differenze di rappresentazione dei concetti comuni a più schemi locali, ma deve anche identificare l'insieme di concetti distinti e memorizzati in schemi differenti che sono correlati attraverso proprietà semantiche (*proprietà interschema*)
- Per poter ragionare sui concetti espressi negli schemi delle diverse sorgenti dati è necessario utilizzare un unico formalismo in modo da fissare i costrutti utilizzabili e la potenza espressiva

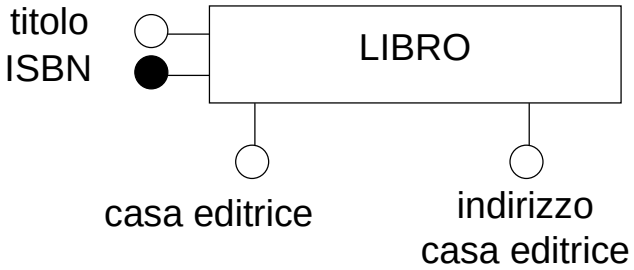
Problemi: diversa prospettiva



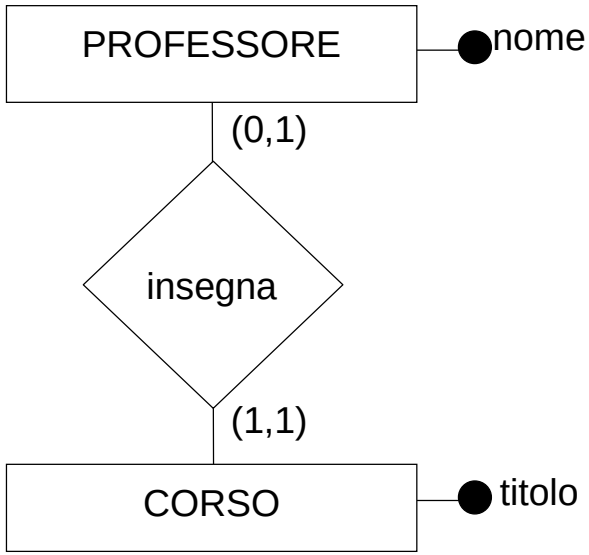


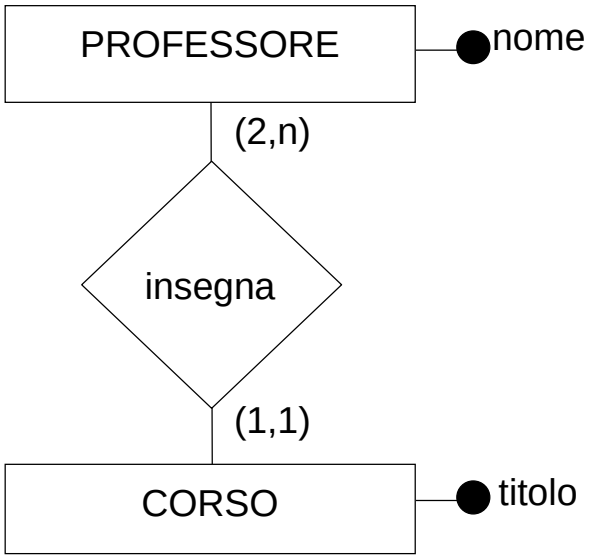
Problemi: costrutti equivalenti





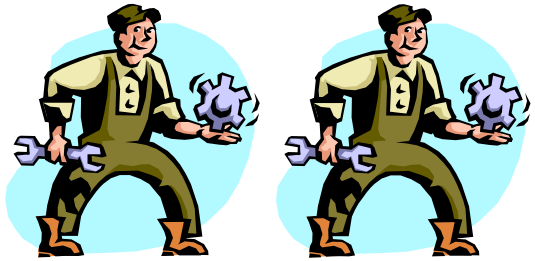
Problemi: incompatibilità

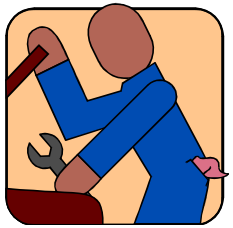




Relazioni tra concetti comuni

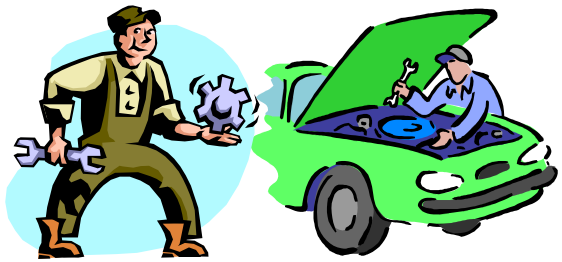
- *Identità*: vengono utilizzati gli stessi costrutti, il concetto è modellato dallo stesso punto di vista e non vengono commessi errori di specifica





- *Equivalenza*: sono stati utilizzati costrutti diversi (ma equivalenti) e non sussistono errori di specifica o diversità di percezione

- *Comparabilità*: i costrutti utilizzati e i punti di vista dei progettisti non sono in contrasto tra loro



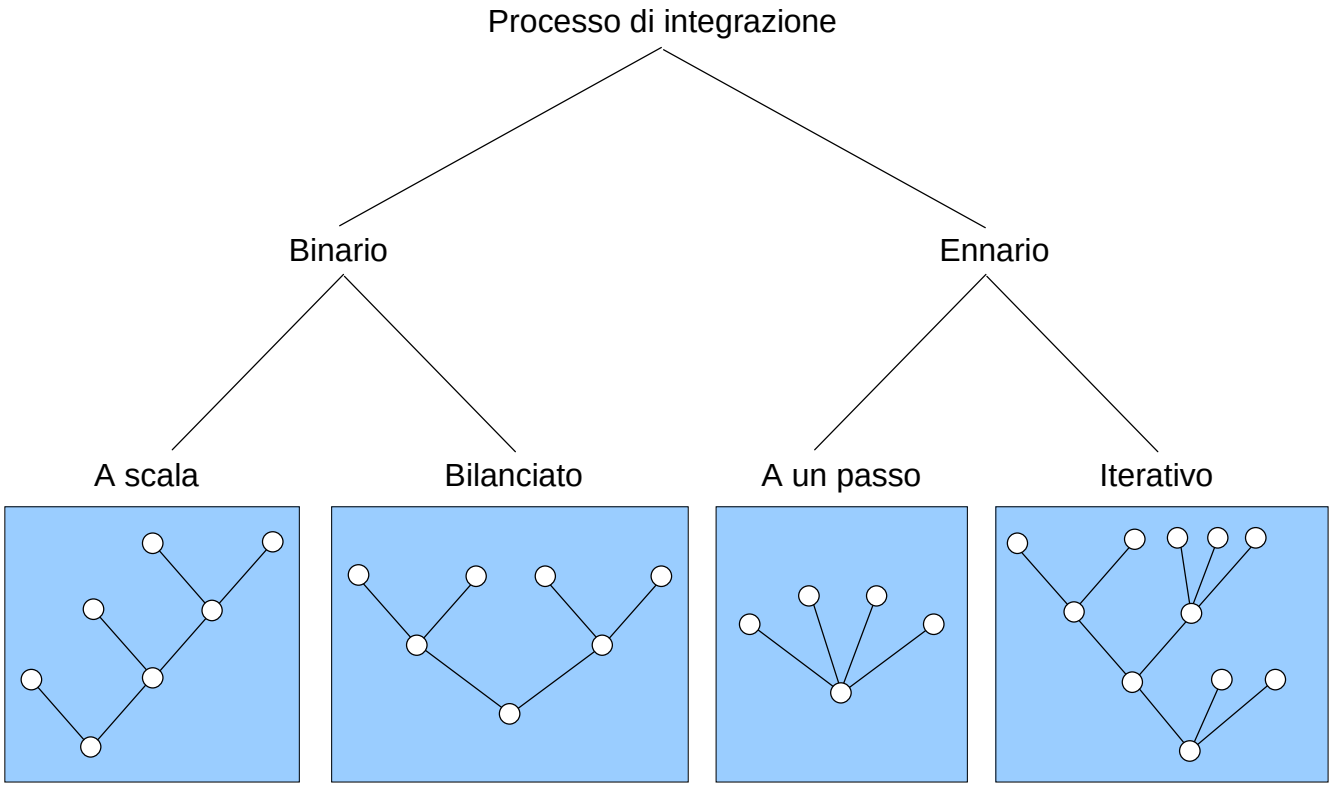
- *Incompatibilità*: gli schemi sono in contrasto a causa dell'incoerenza nelle specifiche

Fasi dell'integrazione

1. *Preintegrazione*
2. *Comparazione degli schemi*
3. *Allineamento degli schemi*
4. *Fusione e ristrutturazione degli schemi*

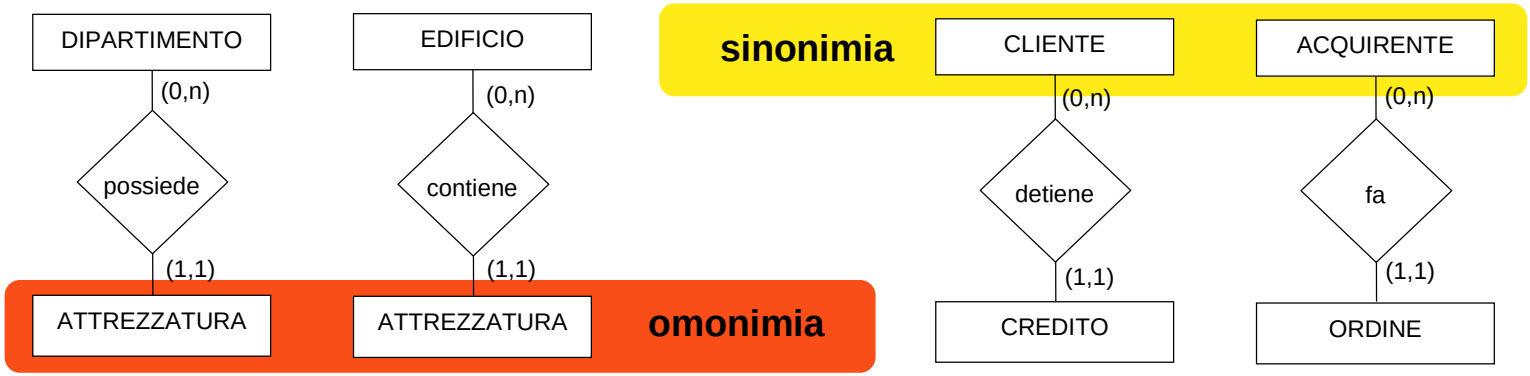
1. Preintegrazione

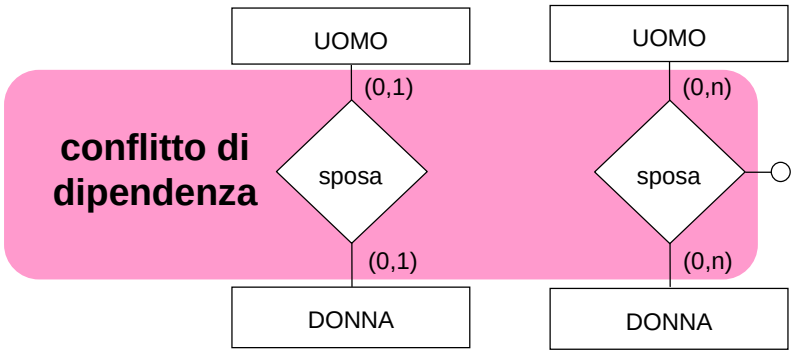
- Viene definita la strategia di integrazione



2. Comparazione degli schemi

- Un'analisi comparativa dei diversi schemi che mira a identificare le correlazioni e i conflitti tra i concetti in essi espressi





3. Allineamento degli schemi

- Scopo di questa fase è la risoluzione dei conflitti evidenziatisi al passo precedente, che si ottiene applicando primitive di trasformazione agli schemi sorgenti o allo schema riconciliato temporaneamente definito
  - ✓ Tipiche primitive di trasformazione riguardano il cambio dei nomi e dei tipi degli attributi, la modifica delle dipendenze funzionali e dei vincoli esistenti sugli schemi
  - ✓ Non sempre i conflitti possono essere risolti, poiché derivano da inconsistenze di base del sistema informativo; in questo caso la soluzione deve essere discussa con gli utenti che dovranno fornire indicazioni su qual è la più fedele interpretazione del mondo reale
  - ✓ In caso di incertezza si preferiscono le trasformazioni che avvantaggiano gli schemi ritenuti centrali nella struttura del data mart

4. Fusione degli schemi

- Gi schemi allineati vengono fusi a formare un unico schema riconciliato; l'approccio più diffuso è quello di sovrapporre i concetti comuni a cui saranno collegati tutti i rimanenti concetti provenienti dagli schemi locali.
- Dopo questa operazione si renderanno necessarie ulteriori trasformazioni mirate a migliorare la struttura dello schema riconciliato rispetto a:
  - ✓ Completezza
  - ✓ Minimalità
  - ✓ Leggibilità

Definizione delle corrispondenze

// DB1 Magazzino
ORDINI2001(chiaveO, chiaveC, data ordine, impiegato)
CLIENTE(chiaveC, nome, indirizzo, città, regione, stato)
.....

// DB2 Amministrazione
CLIENTE(chiaveC, partitaIva, nome, telefono, fatturato)
FATTURE(chiaveF, data, chiaveC, importo, iva)
STORICO_ORDINI2000(chiaveO, chiaveC, data ordine, impiegato)
.....

```
CREATE VIEW CLIENTE AS
SELECT CL1.chiaveC, CL1.nome, CL1.indirizzo, CL1.città, CL1.regione,
                                    CL1.stato, CL2.partitaIva, CL2.telefono, CL2.fatturato
FROM     DB1.CLIENTE AS CL1, DB2.CLIENTE AS CL2
WHERE    CL1.chiaveC = CL2.chiaveC;

CREATE VIEW ORDINI AS
SELECT * FROM DB1.ORDINI2001
UNION
SELECT * FROM DB2.STORICO_ORDINI2000;
```

Analisi dei requisiti

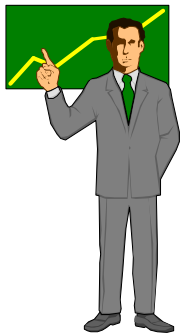
Prof. Stefano Rizzi

Obiettivi

- La fase di analisi dei requisiti ha l'obiettivo di raccogliere le esigenze di utilizzo del data mart espresse dai suoi utenti finali
- Essa ha un'importanza strategica poiché influenza le decisioni da prendere riguardo:
  - ✓ lo schema concettuale dei dati
  - ✓ il progetto dell'alimentazione
  - ✓ le specifiche delle applicazioni per l'analisi dei dati
  - ✓ l'architettura del sistema
  - ✓ il piano di avviamento e formazione
  - ✓ le linee guida per la manutenzione e l'evoluzione del sistema.

Fonti

- La “fonte” principale da cui attingere i requisiti sono i futuri utenti del data mart (*business users*)
  - ✓ La differenza nel linguaggio usato da progettisti e utenti, e la percezione spesso distorta che questi ultimi hanno del processo di warehousing, rendono il dialogo difficile e a volte infruttuoso

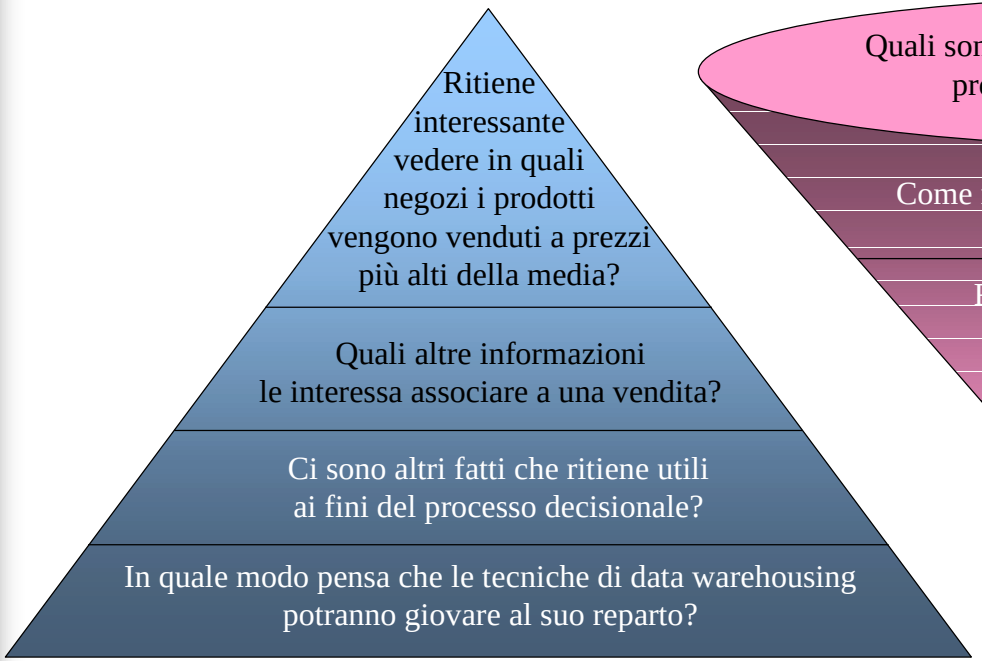


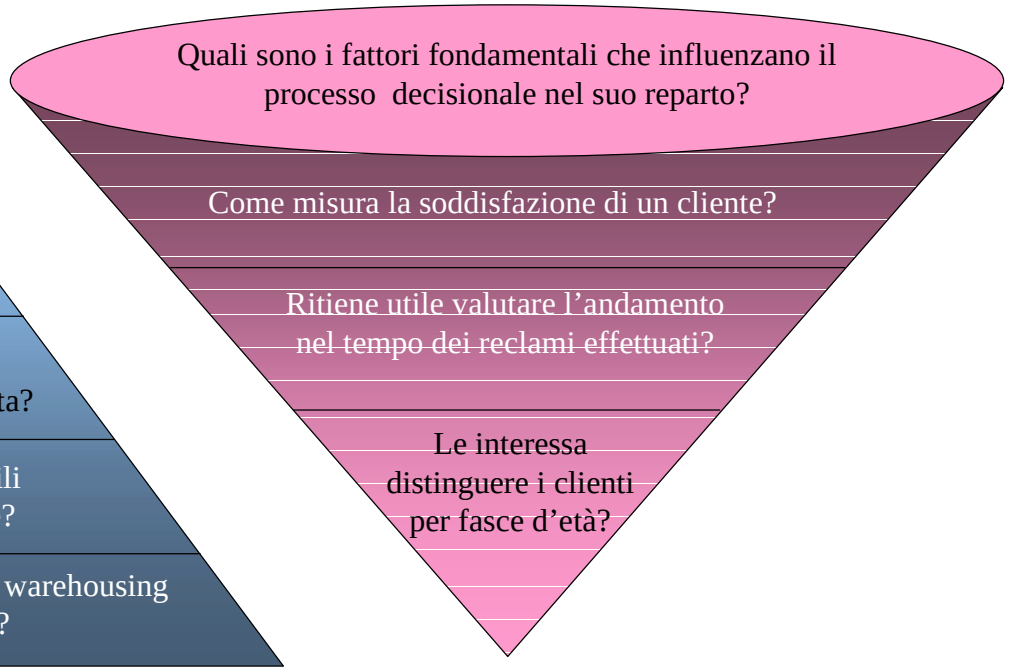
- Per gli aspetti più tecnici, saranno gli amministratori del sistema informativo e/o i responsabili del CED a fungere da riferimento per il progettista
  - ✓ In questo caso, i requisiti che dovranno essere catturati riguardano principalmente vincoli di varia natura imposti sul sistema di data warehousing

Le interviste

- A piramide. Approccio induttivo: l'intervistatore parte da domande molto dettagliate per poi ampliare l'argomento dell'intervista mediante domande aperte che richiedono risposte più generali.
  - ✓ Questo tipo di intervista permette di superare la riluttanza di un intervistato scettico poiché inizialmente non richiede un forte coinvolgimento da parte dell'intervistato.
- A imbuto. Approccio deduttivo: l'intervistatore parte da domande molto generali per poi restringere l'argomento dell'intervista a temi specifici
  - ✓ Questo approccio è utile nel caso in cui l'intervistato sia emozionato o eccessivamente deferente, poiché il fatto che le domande di carattere generale (normalmente in forma aperta) non prevedano una risposta "sbagliata" allevia la tensione dell'intervistato.

Le interviste





Le domande

Ruolo	*Domande chiave*
Dirigente	Quali sono gli obiettivi aziendali? Come misuri il successo della tua azienda? Quali sono oggi i principali problemi dell'azienda? In che modo ti aspetti che una maggiore disponibilità di informazioni possa migliorare la situazione aziendale?
Direttore di reparto	Quali sono gli obiettivi del tuo reparto? Come misuri il successo del tuo reparto? Descrivi i soggetti coinvolti nel tuo settore di interesse. Ci sono colli di bottiglia nell'accesso ai dati? Che analisi di routine esegui? Che tipi di analisi ti piacerebbe poter eseguire? A che livello di dettaglio occorre vedere le informazioni? Quanta informazione storica è necessaria?
Amministratore del sistema informativo	Illustra le caratteristiche delle principali fonti dati disponibili. Che strumenti vengono usati per analizzare i dati? Come vengono gestite le richieste di analisi ad hoc? Quali sono i principali problemi di qualità dei dati?

I fatti

- I fatti sono i concetti su cui gli utenti finali del data mart baseranno il processo decisionale; ogni fatto descrive una categoria di eventi che si verificano in azienda
  - ✓ Fissare le dimensioni di un fatto è importante poiché significa determinarne la granularità, ovvero il più fine livello di dettaglio a cui i dati saranno rappresentati. La scelta della granularità di un fatto nasce da un delicato compromesso tra due esigenze contrapposte: quella di raggiungere un'elevata flessibilità d'utilizzo e quella di conseguire buone prestazioni
  - ✓ Per ogni fatto occorre definire l'intervallo di storicizzazione, ovvero l'arco temporale che gli eventi memorizzati dovranno coprire

I fatti

	Data mart	Fatti
commerciale/ manufatturiero	approvvigionamenti	acquisti, inventario di magazzino, distribuzione
	produzione	confezionamento, inventario, consegna, manifattura
	gestione domanda	vendite, fatturazione, ordini, spedizioni, reclami
	marketing	promozioni, fidelizzazione, campagne pubblicitarie
finanziario	bancario	conti correnti, bonifici, prestiti ipotecari, mutui
	investimenti	acquisto titoli, transazioni di borsa
	servizi	carte di credito, domiciliazioni bollette
sanitario	scheda di ricovero	ricoveri, dimissioni, interventi chirurgici, diagnosi
	pronto soccorso	accessi, esami, dimissioni
	medicina di base	scelte, revoche, prescrizioni
trasporti	merci	domanda, offerta, trasporti
	passeggeri	domanda, offerta, trasporti
	manutenzione	interventi
telecomunicazioni	traffico	traffico in rete, chiamate
	CRM	fidelizzazione, reclami, servizi
turismo	gestione domanda	biglietteria, noleggi auto, soggiorni
	CRM	frequent-flyers, reclami
gestionale	logistica	trasporti, scorte, movimentazione
	risorse umane	assunzioni, dimissioni, promozioni, incentivi
	budgeting	budget commerciale, budget di marketing
	infrastrutture	acquisti, opere

Glossario dei requisiti

Fatto	*Possibili dimensioni*	*Possibili misure*	*Storicità*
inventario di magazzino	prodotto, data, magazzino	quantità in magazzino	1 anno
vendite	prodotto, data, negozio	quantità venduta, importo, sconto	5 anni
linee d'ordine	prodotto, data, fornitore	quantità ordinata, importo, sconto	3 anni

Il carico di lavoro preliminare

- Il riconoscimento di fatti, dimensioni e misure è strettamente collegato all'identificazione di un *carico di lavoro preliminare*.
  - ✓ Oltre che dall'interazione diretta con l'utente, indicazioni al riguardo potranno essere ricavate da un esame della reportistica correntemente in uso in azienda.
  - ✓ In questa fase il carico di lavoro può essere espresso in linguaggio naturale; esso sarà comunque utile per valutare la granularità dei fatti e le misure di interesse, nonché per iniziare ad affrontare il problema dell'aggregazione

Il carico di lavoro preliminare

Fatto	*Interrogazione*
inventario di magazzino	Quantità media di ciascun prodotto presente mensilmente in tutti i magazzini. Prodotti per i quali è stata esaurita la scorta contemporaneamente in tutti i magazzini in almeno un'occasione durante la settimana passata. Andamento giornaliero delle scorte complessive per ciascun tipo di prodotto.
vendite	Quantità totali di ciascun tipo di prodotto vendute durante l'ultimo mese. Incasso totale giornaliero di ciascun negozio. Per un dato negozio, incassi relativi alle diverse categorie di prodotti durante un certo giorno. Riepilogo annuale degli incassi per regione relativamente a un dato prodotto.
linee d'ordine	Quantità totale ordinata annualmente presso un certo fornitore. Importo giornaliero ordinato nell'ultimo mese per un certo tipo di prodotto. Sconto massimo applicato da ciascun fornitore durante l'ultimo anno per ciascuna categoria di prodotto.

Altri requisiti

- Vincoli di progettazione logica e fisica (spazio disponibile)
- Progetto dell'alimentazione (periodicità dell'alimentazione)
- Architettura del sistema di data warehousing (tipo di architettura da implementare, numero dei livelli, presenza di data mart dipendenti o indipendenti, materializzazione del livello riconciliato)
- Applicazioni per l'analisi dei dati (disamina delle tipologie di interrogazioni e dei rapporti analitici normalmente richiesti)
- Piano di avviamento
- Piano di formazione

Progettazione concettuale

Prof. Stefano Rizzi

Quale formalismo?

- Mentre è universalmente riconosciuto che un DW si appoggia sul modello multidimensionale, non c'è accordo sulla metodologia di progetto 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*.

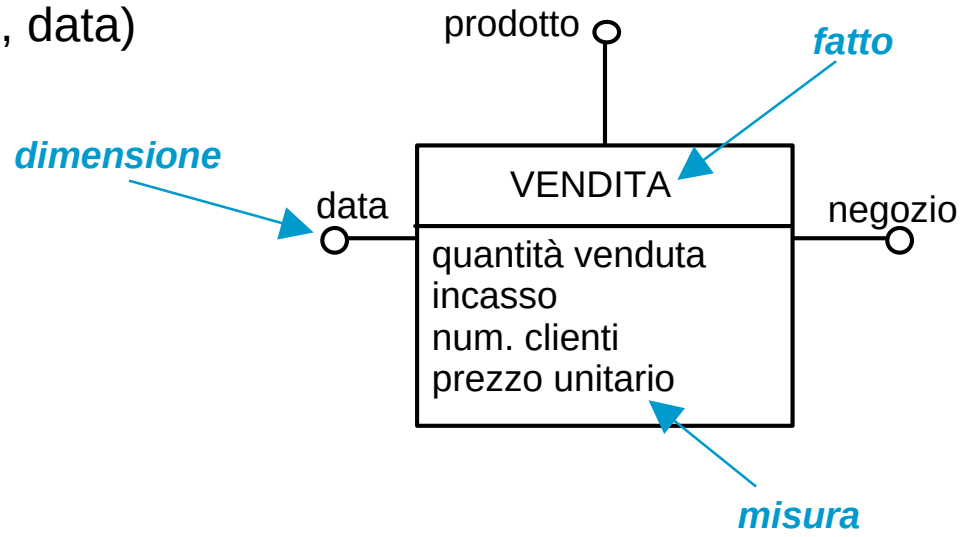
Il Dimensional Fact Model

- Il DFM è 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

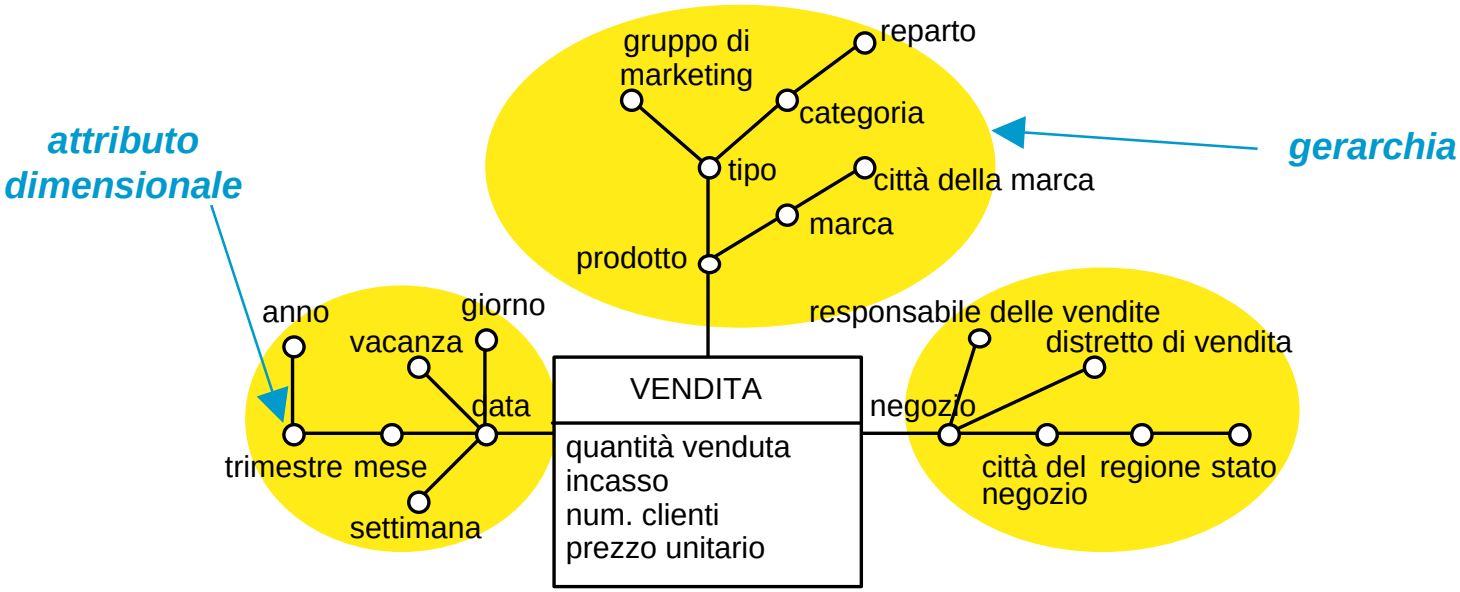
- Un *fatto* è un concetto di interesse per il processo decisionale; tipicamente modella un insieme di eventi che accadono nell'impresa (ad esempio: vendite, spedizioni, acquisti, ...). È essenziale che un fatto abbia aspetti dinamici, ovvero evolva nel tempo
- Una *misura* è una proprietà numerica di 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

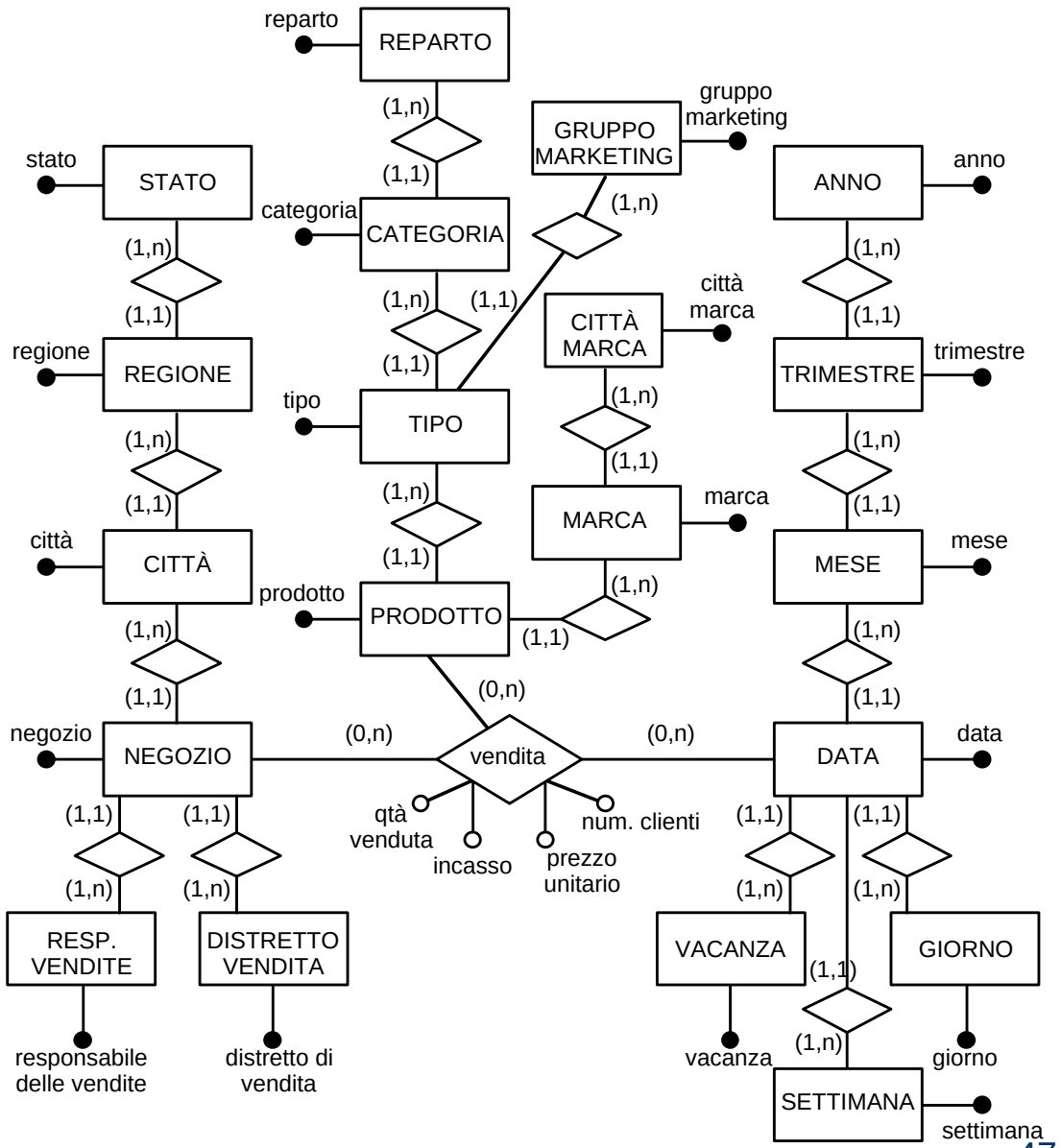


Il DFM: costrutti di base

- Con il termine generale *attributi dimensionali* si intendono le dimensioni e gli eventuali altri attributi, sempre a valori discreti, 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 modellano associazioni molti-a-uno tra coppie di attributi dimensionali. Essa racchiude una dimensione, posta alla radice dell'albero, e tutti gli attributi dimensionali che la descrivono



Il DFM: corrispondenza con l'E/R



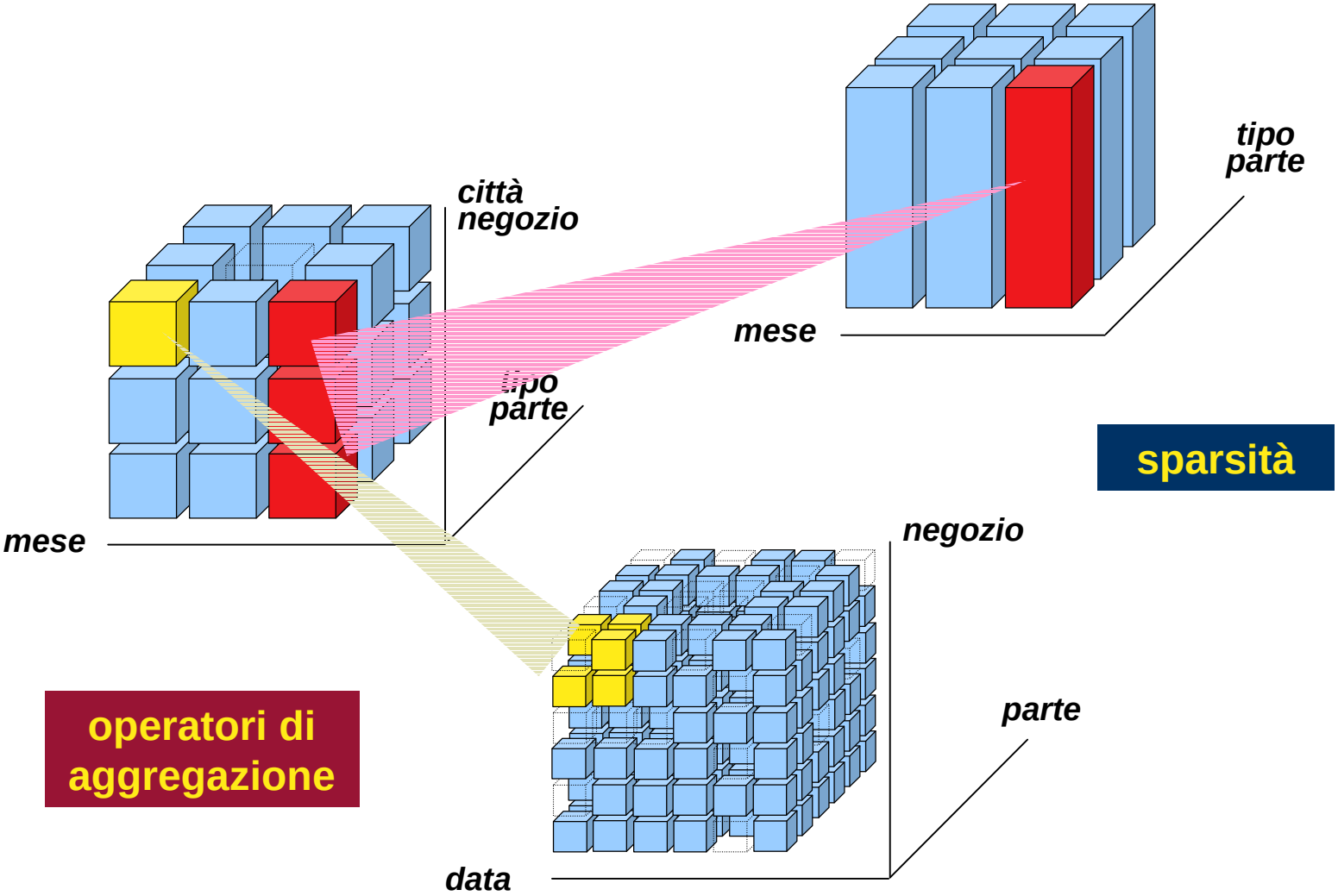
"Naming conventions"

- Tutti gli attributi dimensionali in ciascuno schema di fatto devono avere nomi diversi
- Eventuali nomi uguali devono 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*
- Attributi con lo stesso significato in schemi diversi devono avere lo stesso nome

Eventi e aggregazione

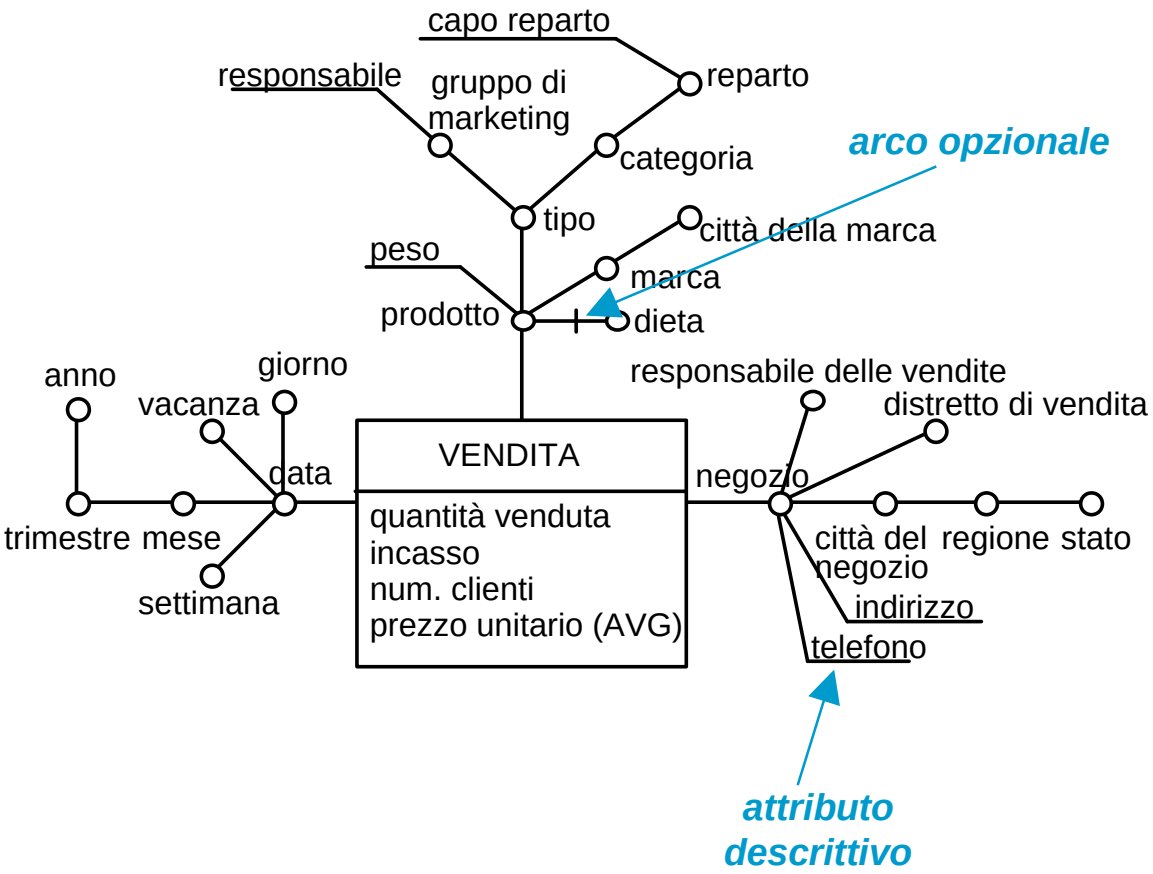
- Un *evento primario* è una particolare occorrenza di un fatto, individuata da una ennupla costituita da un valore per ciascuna dimensione. A ciascun evento primario è associato un valore per ciascuna misura
  - ✓ Con riferimento alle vendite, un possibile evento primario registra per esempio che, il 10/10/2001, nel negozio NonSoloPappa sono state vendute 10 confezioni di detersivo Brillo per un incasso complessivo pari a 25 euro
- 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

Eventi e aggregazione

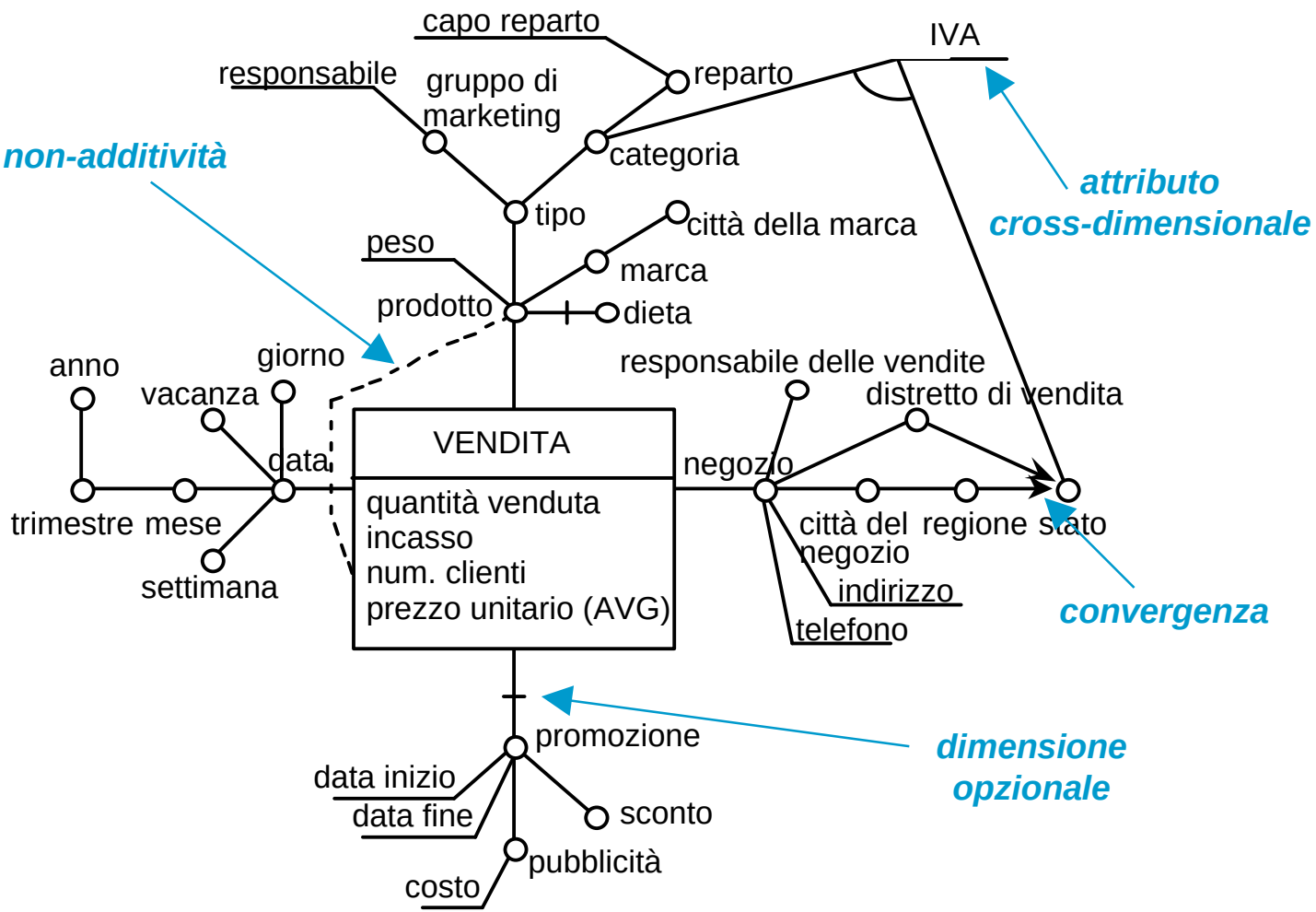


Il DFM: costrutti avanzati

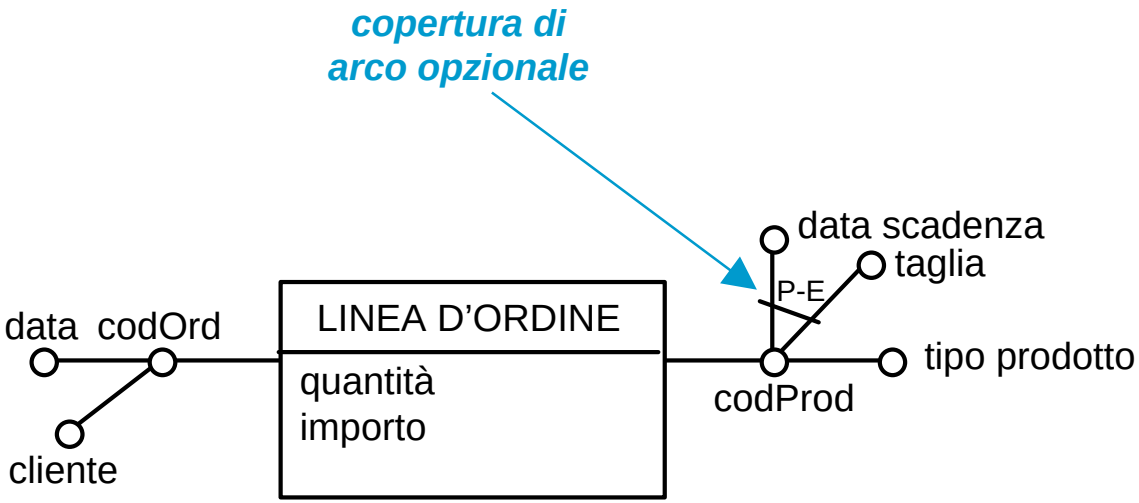
- Un *attributo descrittivo* contiene informazioni aggiuntive su un attributo dimensionale di una gerarchia, a cui è connesso da una associazione -a-uno. Non viene usato per l'aggregazione poiché ha valori continui e/o poiché deriva da un'associazione uno-a-uno
- Alcuni archi dello schema di fatto possono essere *opzionali*



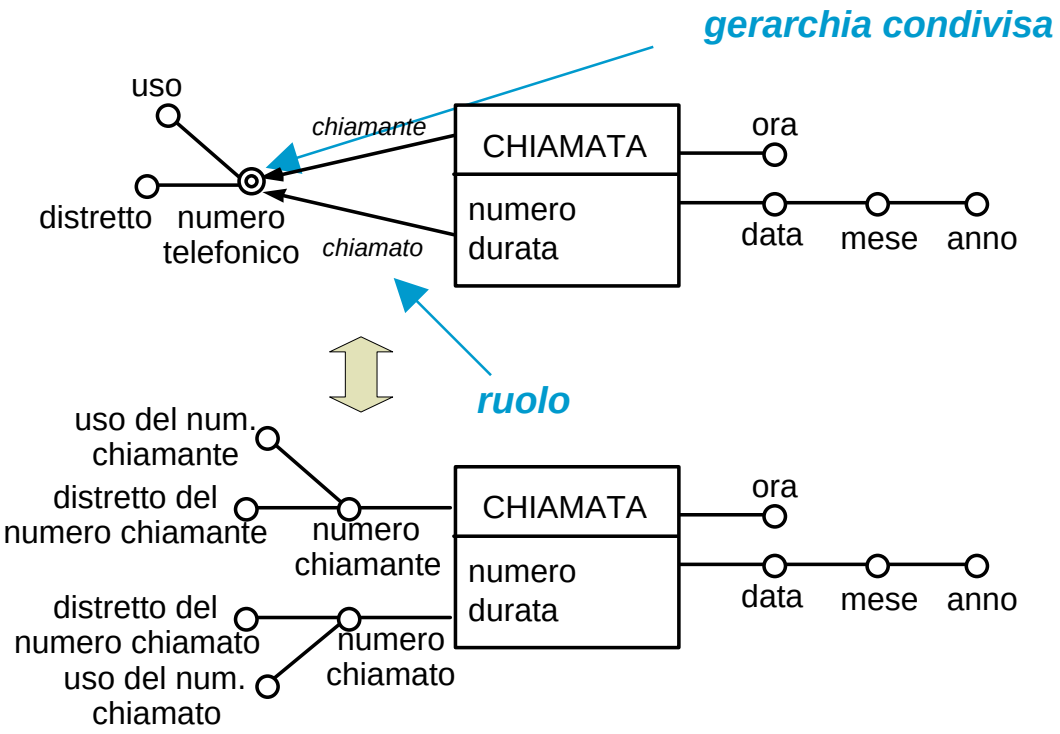
Il DFM: costrutti avanzati

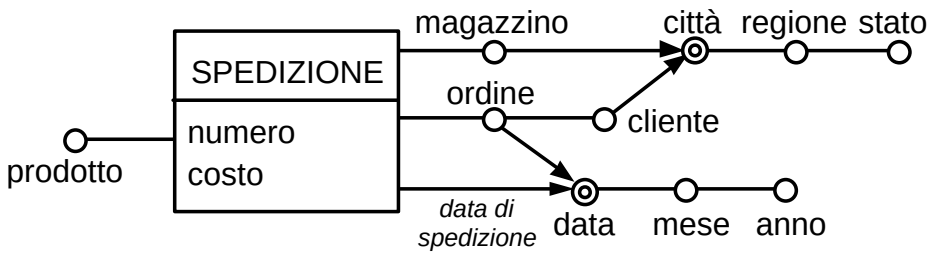


Il DFM: costrutti avanzati

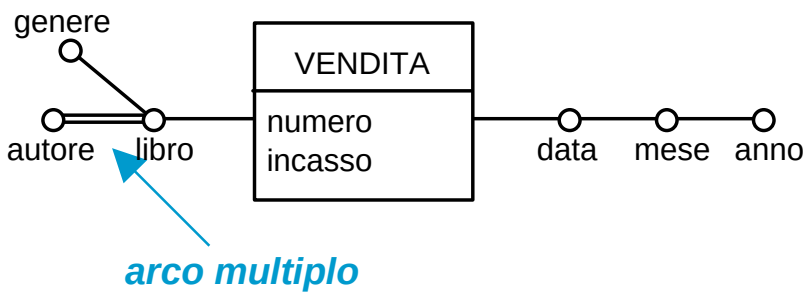


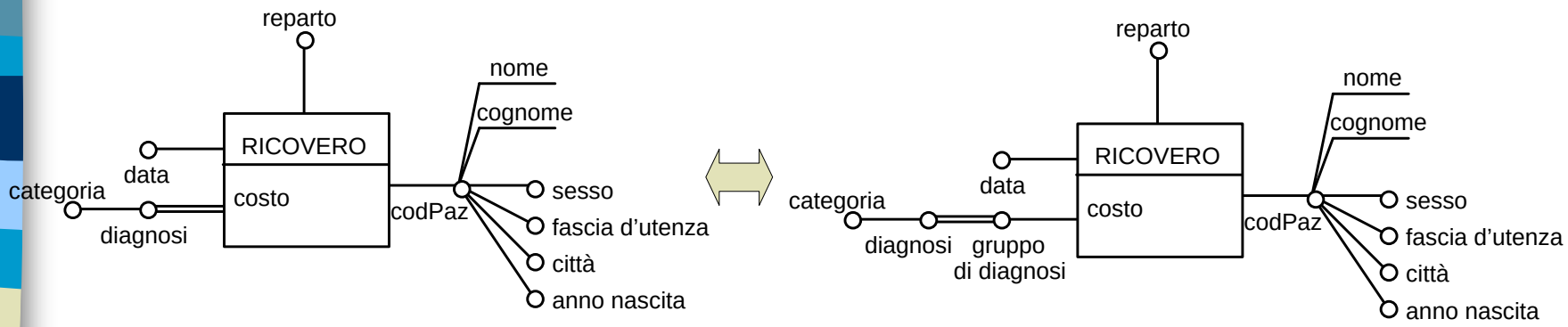
Il DFM: costrutti avanzati





Il DFM: costrutti avanzati





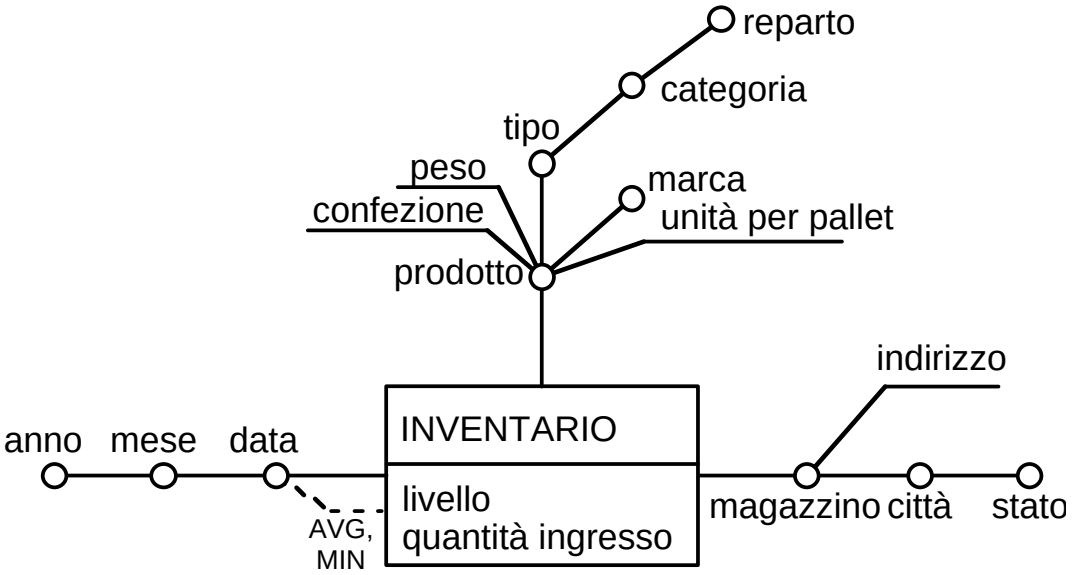
Additività

- L'aggregazione richiede di definire un operatore adatto per comporre i valori delle misure che caratterizzano gli eventi primari in valori da abbinare a ciascun evento secondario
- Da questo punto di vista è possibile distinguere tre categorie di misure:
  - ✓ Misure di flusso: si riferiscono a un periodo, al cui termine vengono valutate in modo cumulativo (il numero di prodotti venduti in un giorno, l'incasso mensile, il numero di nati in un anno)
  - ✓ Misure di livello: vengono valutate in particolari istanti di tempo (il numero di prodotti in inventario, il numero di abitanti di una città)
  - ✓ Misure unitarie: vengono valutate in particolari istanti di tempo, ma sono espresse in termini relativi (il prezzo unitario di un prodotto, la percentuale di sconto, il cambio di una valuta)

	Gerarchie temporali	Gerarchie non temporali
Misure di flusso	SUM, AVG, MIN, MAX	SUM, AVG, MIN, MAX
Misure di livello	AVG, MIN, MAX	SUM, AVG, MIN, MAX
Misure unitarie	AVG, MIN, MAX	AVG, MIN, MAX

Additività

- Una misura è detta additiva su una dimensione se i suoi valori possono essere aggregati lungo la corrispondente gerarchia tramite l'operatore di somma, altrimenti è detta non-additiva. Una misura non-additiva è non-aggregabile se nessun operatore di aggregazione può essere usato su di essa



Misure additive

categoria	tipo	prodotto
pulizia casa	detersivo	Brillo
		Sbianco
		Lucido
	sapone	Manipulite
		Scent
alimentari	latticino	Latte F Slurp
		Latte U Slurp
		Yogurt Slurp
	bibita	Bevimi
		Colissima

anno	1999				2000			
trim.	I '99	II '99	III '99	IV '99	I '00	II '00	III '00	IV '00
	100	90	95	90	80	70	90	85
	20	30	20	10	25	30	35	20
	60	50	60	45	40	40	50	40
	15	20	25	30	15	15	20	10
	30	35	20	25	30	30	20	15
	90	90	85	75	60	80	85	60
	60	80	85	60	70	70	75	65
	20	30	40	35	30	35	35	20
	20	10	25	30	35	30	20	10
	50	60	45	40	50	60	45	40

anno	1999				2000			
trim.	I'99	II'99	III'99	IV'99	I'00	II'00	III'00	IV'00
categoria								
pulizia cas	225	225	220	200	190	185	215	170
alimentari	240	270	280	240	245	275	260	195

anno	1999	2000
categoria		
pulizia casa	870	760
alimentari	1030	975

categoria	tipo	1999	2000
pulizia casa	detersivo	670	605
	sapone	200	155
alimentari	latticino	750	685
	bibita	280	290

Misure non-additive

categoria	tipo	prodotto
pulizia casa	detersivo	Brillo
		Sbianco
		Lucido
	sapone	Manipulite
		Scent

anno trim.	1999			
	I'99	II'99	III'99	IV'99
	2	2	2,2	2,5
	1,5	1,5	2	2,5
	–	3	3	3
	1	1,2	1,5	1,5
	1,5	1,5	2	–

categoria	tipo
pulizia casa	detersivo
	sapone
	media:

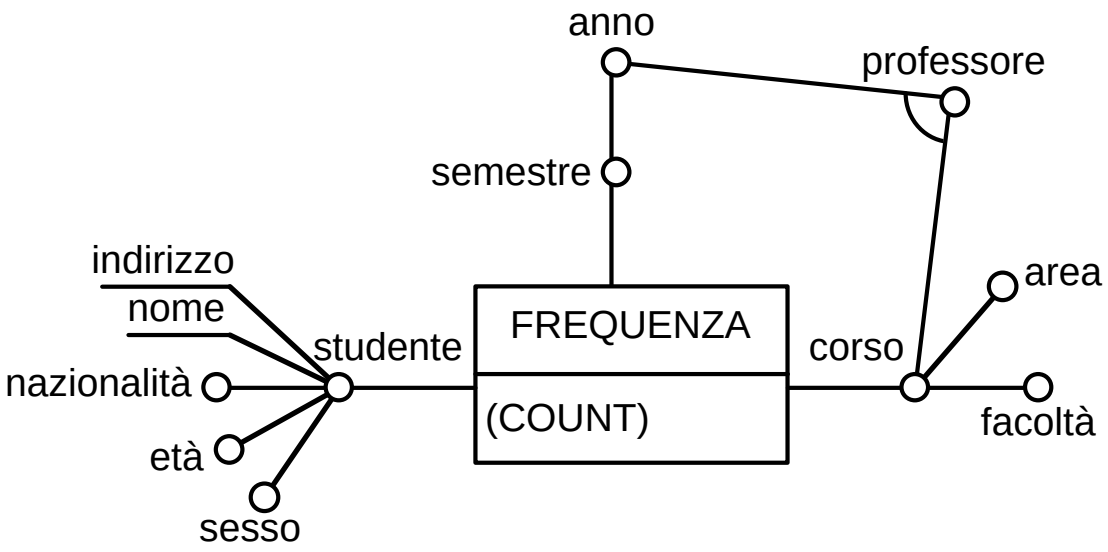
anno trim.	1999			
	I'99	II'99	III'99	IV'99
	1,75	2,17	2,40	2,67
	1,25	1,35	1,75	1,50
	1,50	1,76	2,08	2,09

categoria
pulizia casa

anno trim.	1999			
	I'99	II'99	III'99	IV'99
	1,50	1,84	2,14	2,38

Schemi di fatto vuoti

- Uno schema di fatto si dice **vuoto** se non ha misure
  - ✓ In questo caso, il fatto registra solo il verificarsi di un evento



Progettazione concettuale: approcci

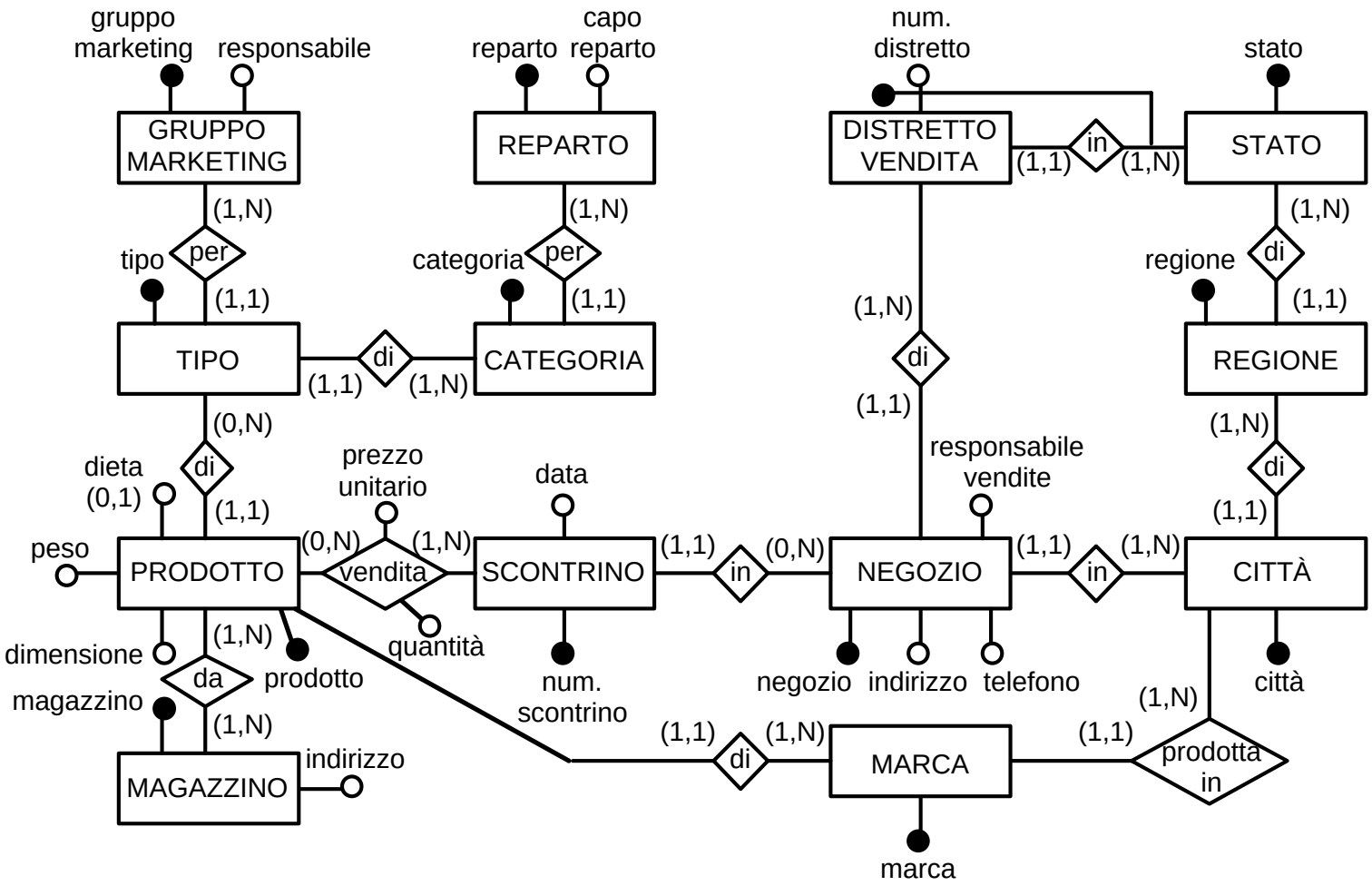
- Basata sui requisiti
  - ✓ Il progettista deve essere in grado di enucleare, dalle interviste condotte presso l'utente, un'indicazione precisa circa i fatti da rappresentare, le misure che li descrivono e le gerarchie attraverso cui aggregarli utilmente. Il problema del collegamento tra lo schema concettuale così determinato e le sorgenti operazionali viene affrontato in un secondo tempo
- Basata sulle sorgenti
  - ✓ È possibile definire lo schema concettuale in funzione della struttura delle sorgenti, evitando il complesso compito di stabilire il legame con esse a posteriori. Inoltre, è possibile derivare uno schema concettuale prototipale dagli schemi operazionali in modo pressoché automatico

Progettazione concettuale: come

- La progettazione concettuale viene effettuata a partire dalla documentazione relativa al database riconciliato:
  - ✓ Schemi E/R
  - ✓ Schemi Relazionali
  - ✓ Schemi XML
  - ✓ .....

- Passi di progettazione:
  - ❶ Definizione dei fatti
  - ❷ Per ogni fatto:
    1. Costruzione di un *albero degli attributi*
    2. Editing dell'albero degli attributi
    3. Definizione delle dimensioni
    4. Definizione delle misure
    5. Creazione dello schema di fatto

L'esempio delle vendite (da E/R)



L'esempio delle vendite (da schema logico)

PRODOTTI (prodotto, peso, dimensione, dieta,
diMarca:MARCHE, diTipo:TIPI)

NEGOZI (negozio, indirizzo, telefono, respVendite,
(numDistr, stato):DISTRETTI, inCittà:CITTÀ)

SCONTRINI (numScontrino, data, negozio:NEGOZI)

VENDITE (prodotto:PRODOTTI, numScontrino:SCONTRINI,
quantità, prezzoUnitario)

MAGAZZINI (magazzino, indirizzo)

CITTÀ (città, regione:REGIONI)

REGIONI (regione, stato:STATI)

STATI (stato)

DISTRETTI (numDistr, stato:STATI)

PROD_IN_MAGAZZ (prodotto:PRODOTTI, magazzino:MAGAZZINI)

MARCHE (codMarca, prodottaIn:CITTÀ)

TIPI (tipo, gruppoMarketing:GRUPPIMARK,
categoria:CATEGORIE)

GRUPPIMARK (gruppoMarketing, responsabile)

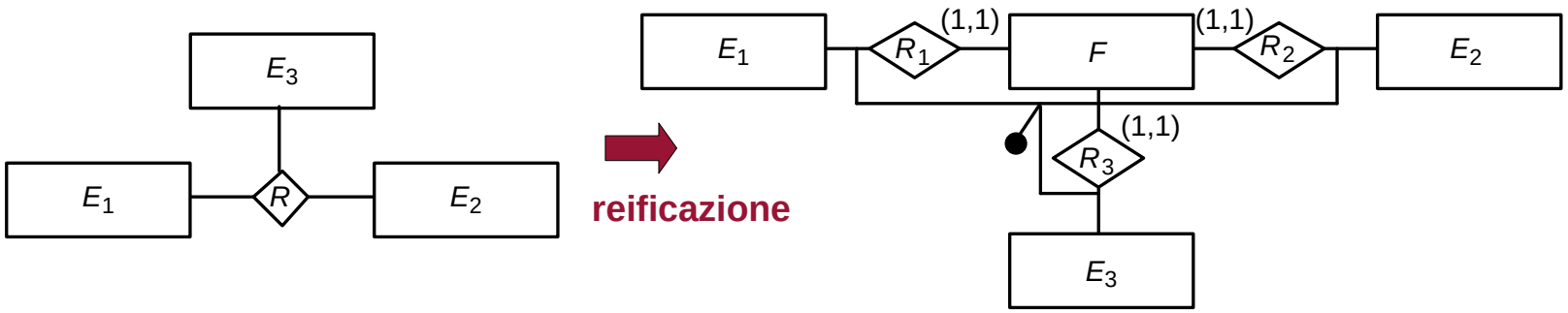
CATEGORIE (categoria, reparto:REPARTI)

REPARTI (reparto, capoReparto)

Definizione dei fatti

I fatti sono concetti di interesse primario per il processo decisionale; tipicamente, corrispondono a eventi che accadono dinamicamente nel mondo aziendale

- Sullo schema E/R un fatto può corrispondere o a un'entità F o a un'associazione n-aria R tra le entità E1, E2..., En

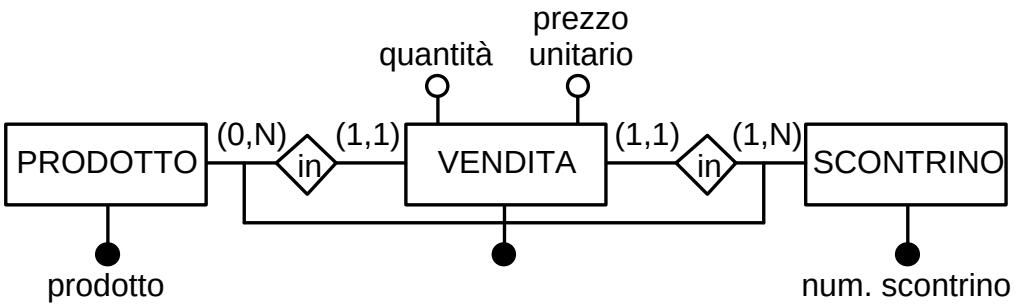


- Sullo schema relazionale un fatto corrisponde a una relazione F

Definizione dei fatti

Le entità o relazioni che rappresentano archivi frequentemente modificati (come VENDITA) sono buoni candidati per definire fatti; quelli che rappresentano archivi quasi-statici (come NEGOZIO e CITTÀ) no

- ✓ Nell'esempio delle vendite si sceglie come fatto l'associazione VENDITA, corrispondente alla relazione VENDITE.



- Ogni fatto identificato diviene la radice di un nuovo schema

Costruzione dell'albero degli attributi

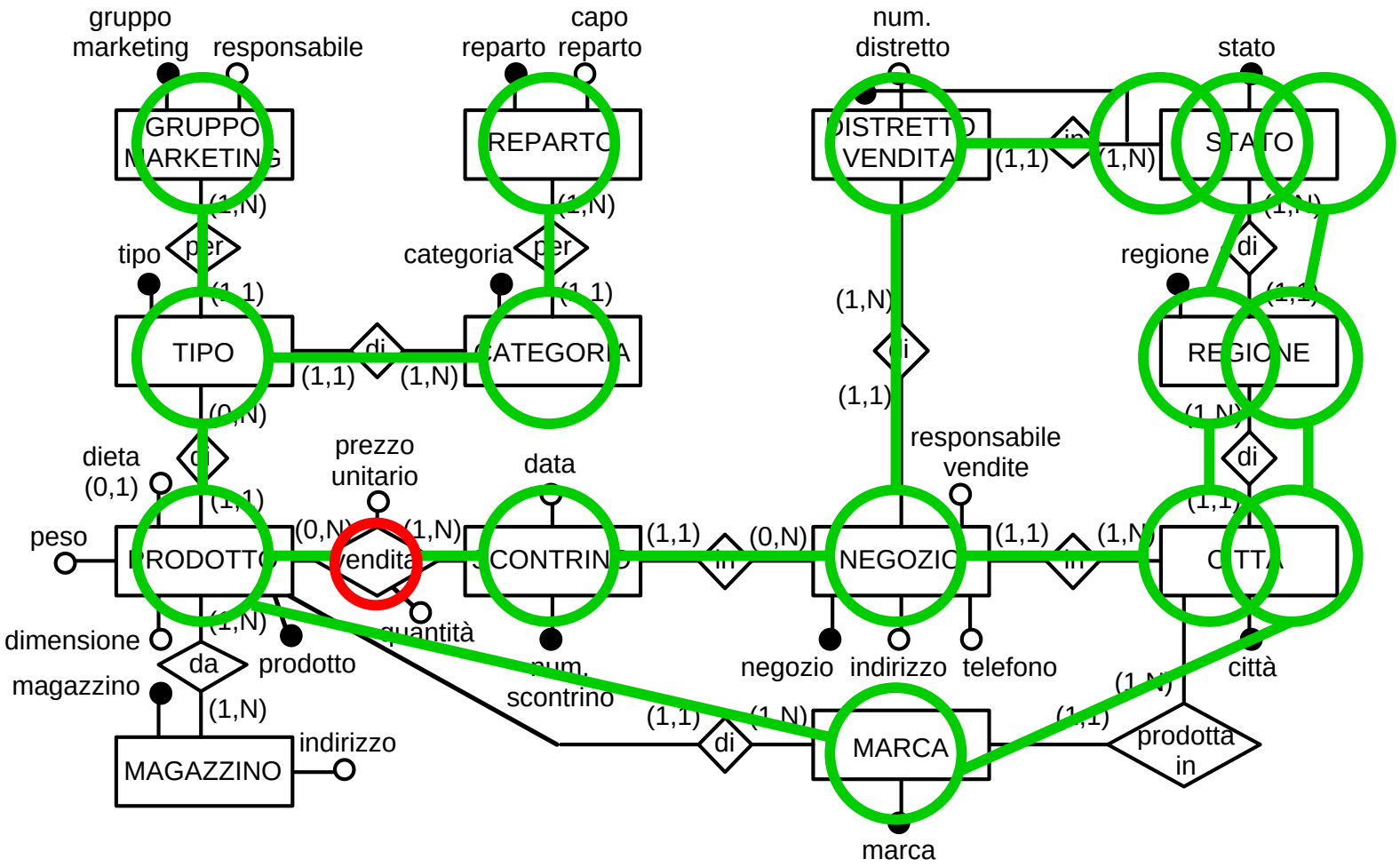
- L'albero degli attributi è un albero in cui:
  - ✓ ogni vertice corrisponde a un attributo - semplice o composto - dello schema sorgente;
  - ✓ la radice corrisponde all'identificatore (chiave primaria) di F;
  - ✓ per ogni vertice v, l'attributo corrispondente determina funzionalmente tutti gli attributi corrispondenti ai discendenti di v
- L'albero degli attributi corrispondente a F può essere costruito in modo automatico applicando una procedura che naviga ricorsivamente le dipendenze funzionali espresse, nello schema sorgente, dagli identificatori e dalle associazioni a-uno

Costruzione dell'albero degli attributi

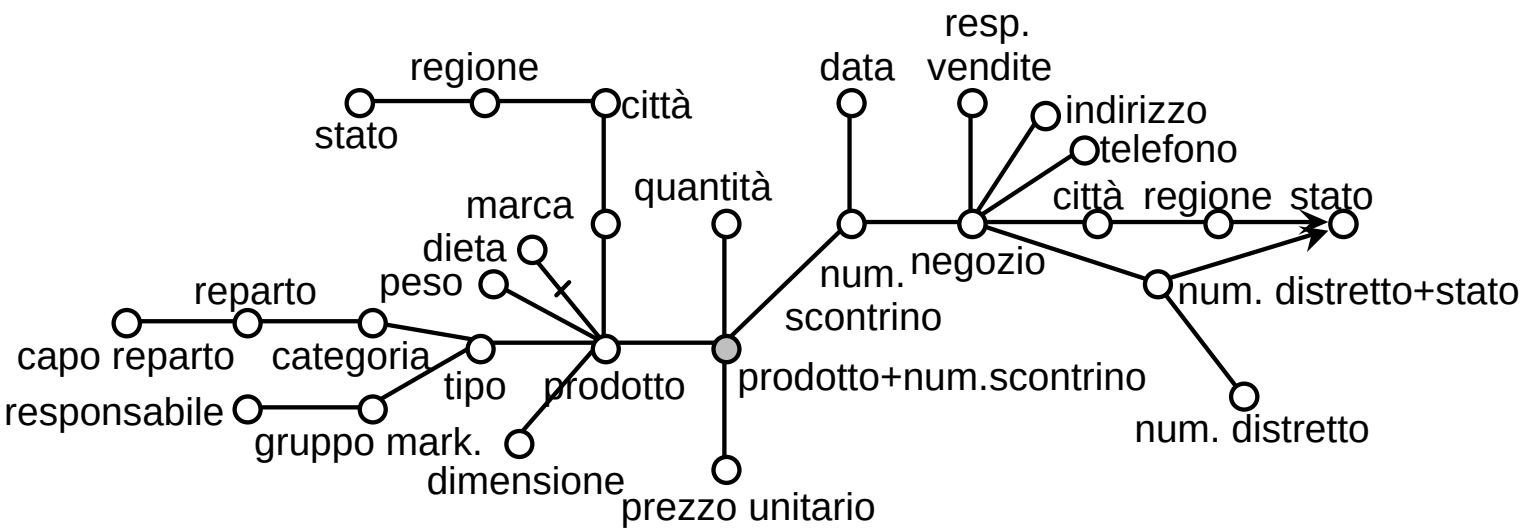
```
root=nuovoVertice(ident(F));          // ident(F) è l'identificatore di F
// la radice dell'albero è etichettata con l'identificatore dell'entità scelta come fatto
traduci(F, root);

procedura traduci(E, v):
// E è l'entità corrente dello schema sorgente, v il vertice corrente dell'albero
{   per ogni attributo a∈E tale che a≠ident(E)
            aggiungiFiglio(v, nuovoVertice(a));
            // aggiunge al vertice v un figlio a
    per ogni entità G connessa a E da un'associazione R tale che max(E,R)=1
    {       per ogni attributo b∈R
                        aggiungiFiglio(v, nuovoVertice(b));
                        // aggiunge al vertice v un figlio b
            prossimo=nuovoVertice(ident(G));
            // crea un nuovo vertice con il nome dell'identificatore di G ...
            aggiungiFiglio(v, prossimo);
            // ... lo aggiunge a v come figlio ...
            traduci(G, prossimo);
            // ... e innesca la ricorsione
    }
}
```

L'esempio delle vendite



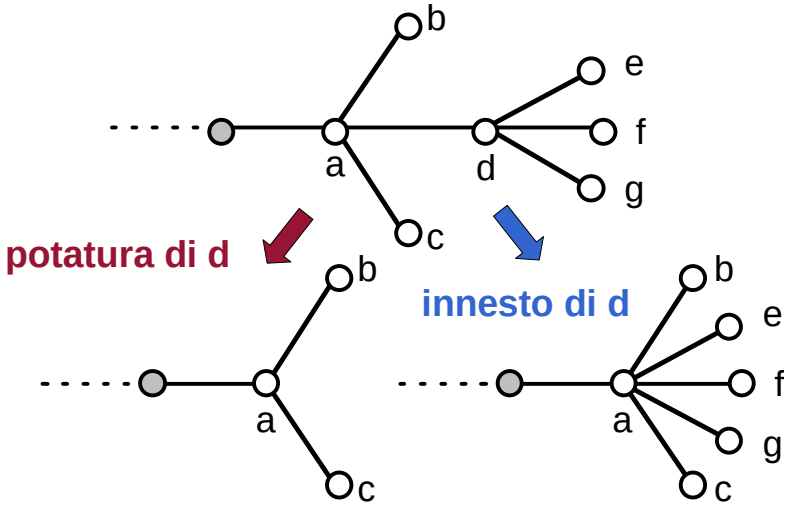
L'esempio delle vendite



Editing dell'albero

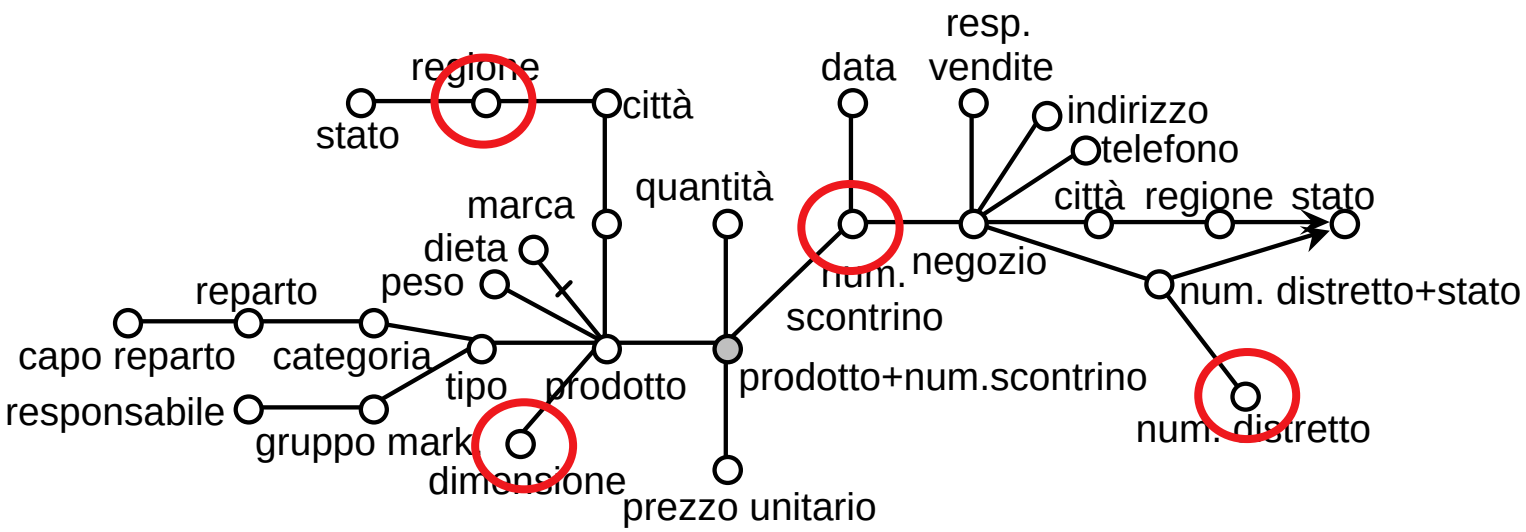
- In genere non tutti gli attributi dell'albero sono d'interesse per il data mart; quindi, l'albero può essere manipolato per eliminare i livelli di dettaglio non necessari
  - ✓ La potatura di un vertice v si effettua eliminando l'intero sottoalbero con radice in v
    - Gli attributi eliminati non verranno inclusi nello schema di fatto, quindi non potranno essere usati per aggregare i dati
  - ✓ L'innesto viene utilizzato quando, sebbene un vertice esprima un'informazione non interessante, è necessario mantenere nell'albero i suoi discendenti
    - L'innesto del vertice v, con padre v', viene effettuato collegando tutti i figli di v direttamente a v' ed eliminando v; come risultato verrà perduto il livello di aggregazione corrispondente all'attributo v ma non i livelli corrispondenti ai suoi discendenti

Editing dell'albero

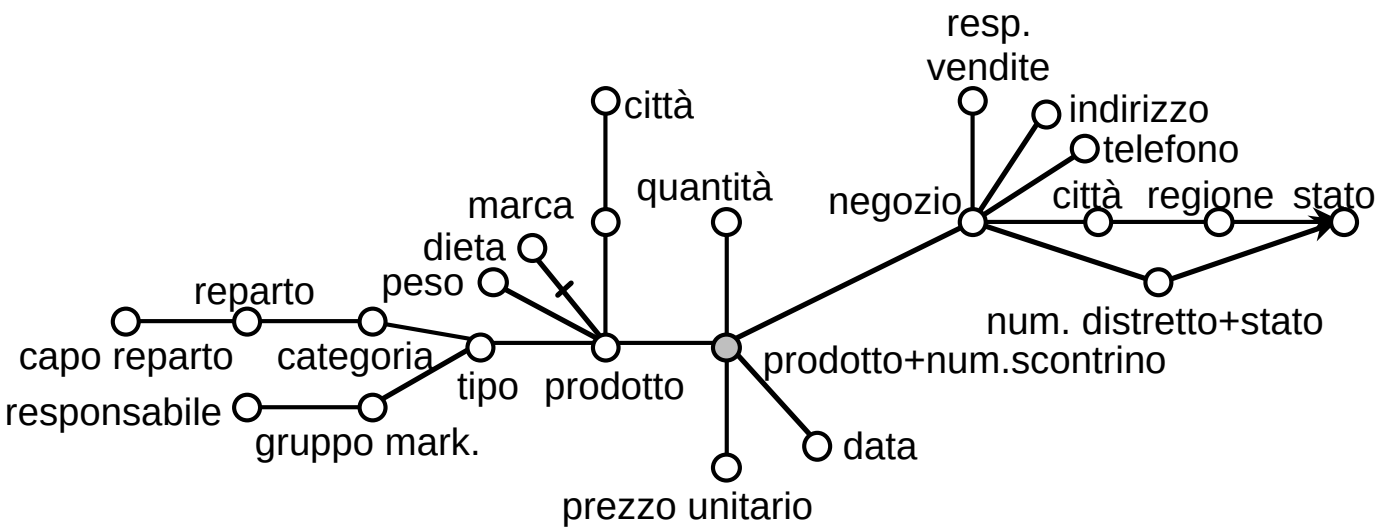


- Quando un vertice opzionale viene innestato, tutti i suoi figli ereditano il trattino di opzionalità
  - ✓ Nel caso di potatura o innesto di un vertice opzionale v con padre v' è possibile aggiungere a v' un nuovo figlio b corrispondente a un attributo booleano che esprima l'opzionalità
- Potare o innestare un figlio della radice che corrisponde, sullo schema sorgente, a un attributo incluso nell'identificatore dell'entità scelta come fatto significa rendere più grossolana la granularità del fatto
  - ✓ Se il vertice innestato ha più di un figlio, si può avere un aumento del numero di dimensioni nello schema di fatto

L'esempio delle vendite

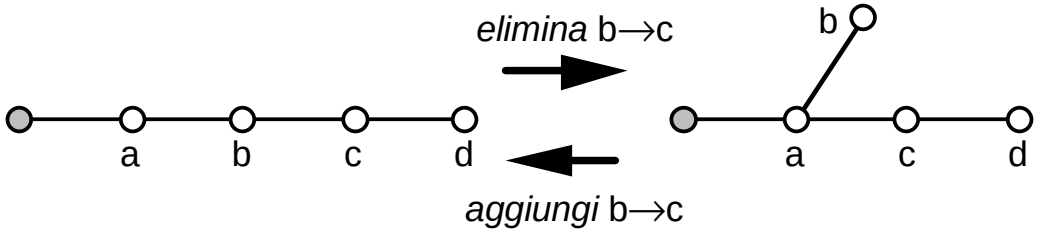


L'esempio delle vendite

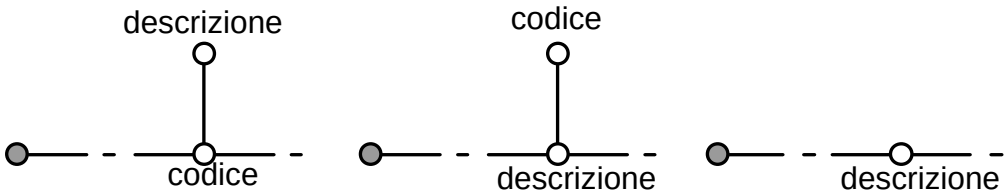


Editing dell'albero

- Nella pratica possono rendersi necessarie ulteriori manipolazioni sull'albero degli attributi
  - ✓ Può essere necessario modificarne radicalmente la struttura sostituendo il padre di un certo nodo: ciò corrisponde ad aggiungere o eliminare una dipendenza funzionale



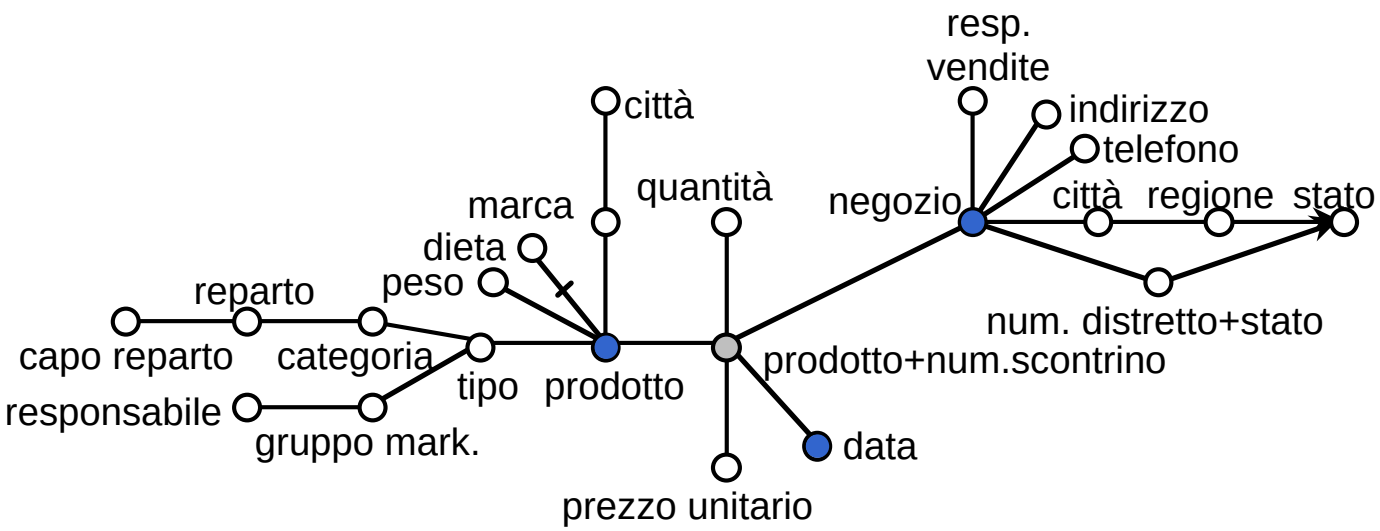
  - ✓ In presenza di un'associazione uno-a-uno sono consigliabili due soluzioni:
    - quando il vertice v determinato dall'associazione uno-a-uno ha dei discendenti di interesse lo si può eliminare dall'albero tramite innesto;
    - quando v non ha discendenti di interesse lo si può rappresentare come attributo descrittivo.
    - in alcuni casi può convenire *invertire* i due nodi coinvolti



Definizione delle dimensioni

- Le dimensioni devono essere scelte nell'albero degli attributi tra i vertici figli della radice; possono corrispondere ad attributi discreti o a intervalli di valori di attributi discreti o continui
- La loro scelta è cruciale per il progetto poiché definisce la *granularità* degli eventi primari
- Il tempo dovrebbe sempre essere una dimensione:
  - ✓ Se la sorgente è uno schema storico, il tempo è rappresentato esplicitamente come un attributo; se appare nell'albero degli attributi come figlio di un vertice diverso dalla radice, si può effettuare un innesto o eliminare una dipendenza funzionale al fine di farlo diventare un figlio diretto della radice e quindi una dimensione
  - ✓ Nelle sorgenti snapshot il tempo non viene rappresentato esplicitamente; in questo caso il tempo viene tipicamente aggiunto "manualmente" allo schema di fatto

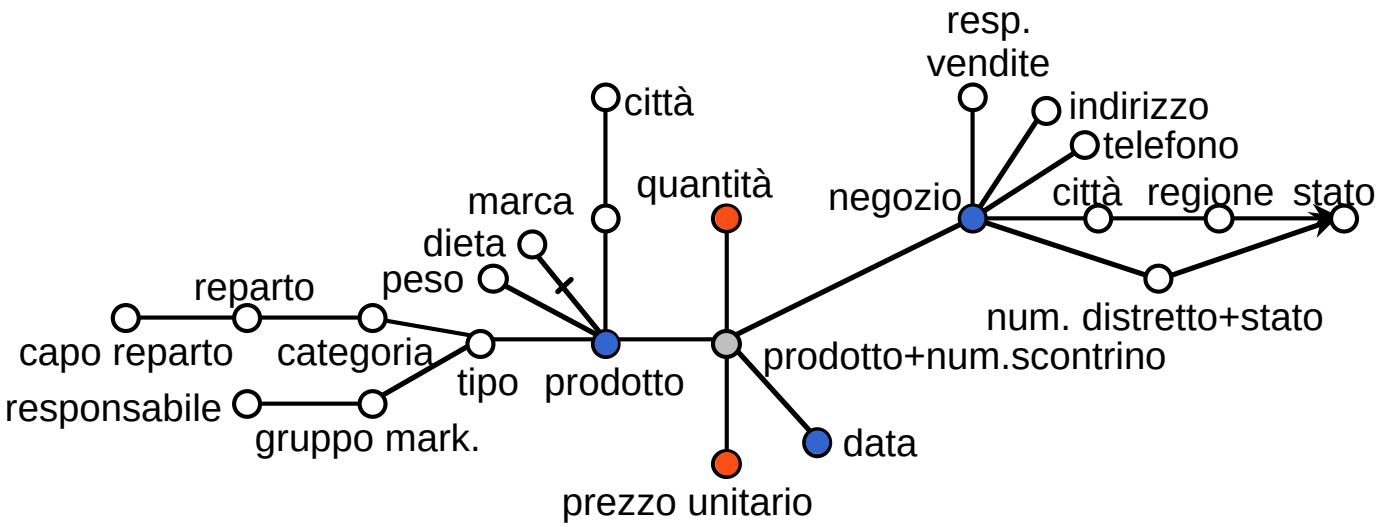
L'esempio delle vendite



Definizione delle misure

- Se tra le dimensioni compaiono tutti gli attributi che costituiscono un identificatore dell'entità fatto, allora le misure corrispondono ad attributi numerici che siano figli della radice dell'albero
- In caso contrario le misure devono essere definite applicando, ad attributi numerici dell'albero, funzioni di aggregazione che operano su tutte le istanze di F corrispondenti a ciascun evento primario (in genere si tratta di somma/media/massimo/minimo di espressioni oppure del conteggio del numero di istanze di F)
  - ✓ Un fatto può anche non avere misure
  - ✓ Qualora la granularità del fatto sia differente da quella dello schema sorgente, può essere utile definire più misure che aggregano lo stesso attributo tramite operatori diversi

L'esempio delle vendite



GLOSSARIO

quantità venduta = SUM(VENDITA.quantità)

incasso = SUM(VENDITA.quantità*VENDITA.prezzoUnitario)

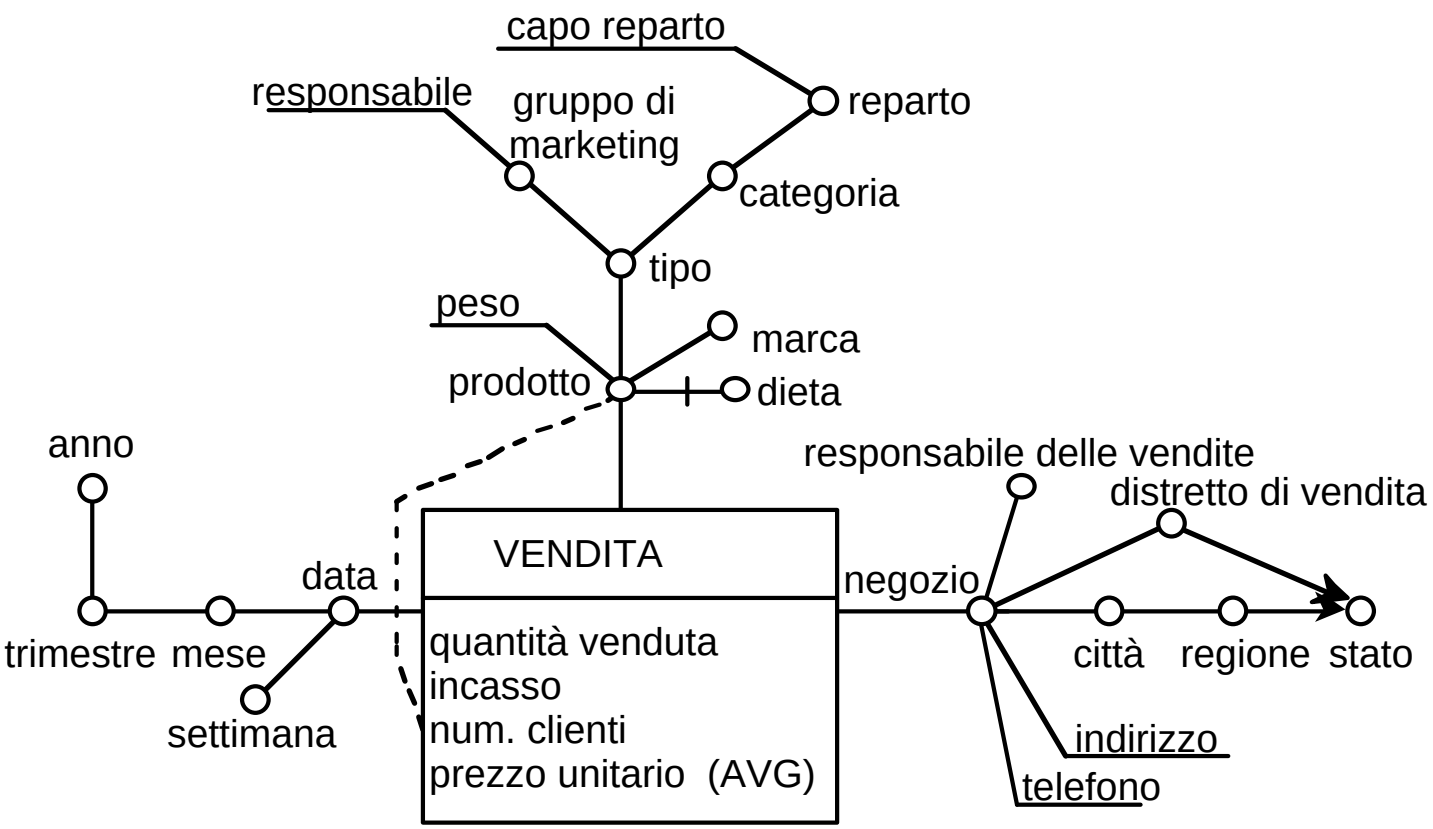
prezzo unitario = AVG(VENDITA.prezzoUnitario)

num. clienti = COUNT(*)

Creazione dello schema di fatto

- L'albero degli attributi può ora essere tradotto in uno schema di fatto che include le dimensioni e misure definite
  - ✓ le gerarchie corrispondono ai sottoalberi dell'albero degli attributi con radice nelle diverse dimensioni
  - ✓ il nome del fatto corrisponde al nome dell'entità scelta come fatto
  - ✓ È possibile potare e innestare l'albero per eliminare dettagli inutili
  - ✓ È possibile aggiungere attributi dimensionali definendo opportuni intervalli per attributi numerici (per es. sulla dimensione tempo)
  - ✓ Gli attributi che non verranno usati per l'aggregazione possono essere contrassegnati come descrittivi; tra questi compariranno in genere anche gli attributi determinati da associazioni uno-a-uno e privi di discendenti
  - ✓ Per quanto riguarda eventuali attributi alfanumerici figli della radice ma non prescelti né come dimensioni né come misure:
    - se la granularità degli eventi primari coincide con quella dell'entità F, essi possono essere rappresentati come attributi descrittivi associati direttamente al fatto, di cui descriveranno ciascuna occorrenza
    - se invece le due granularità sono differenti, essi devono necessariamente essere potati

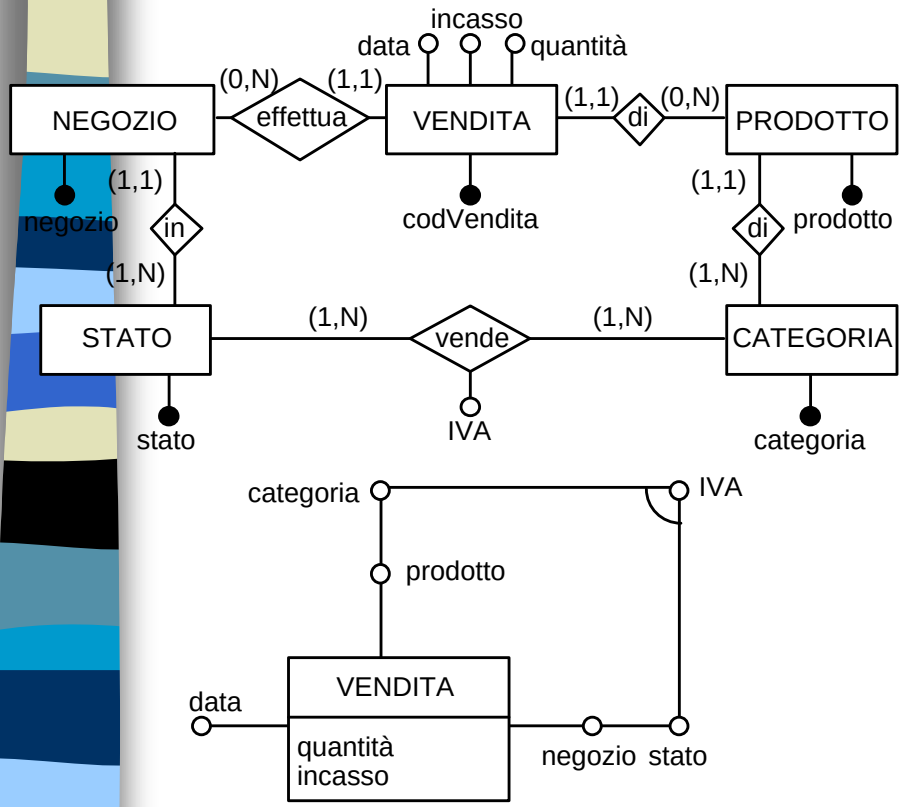
L'esempio delle vendite

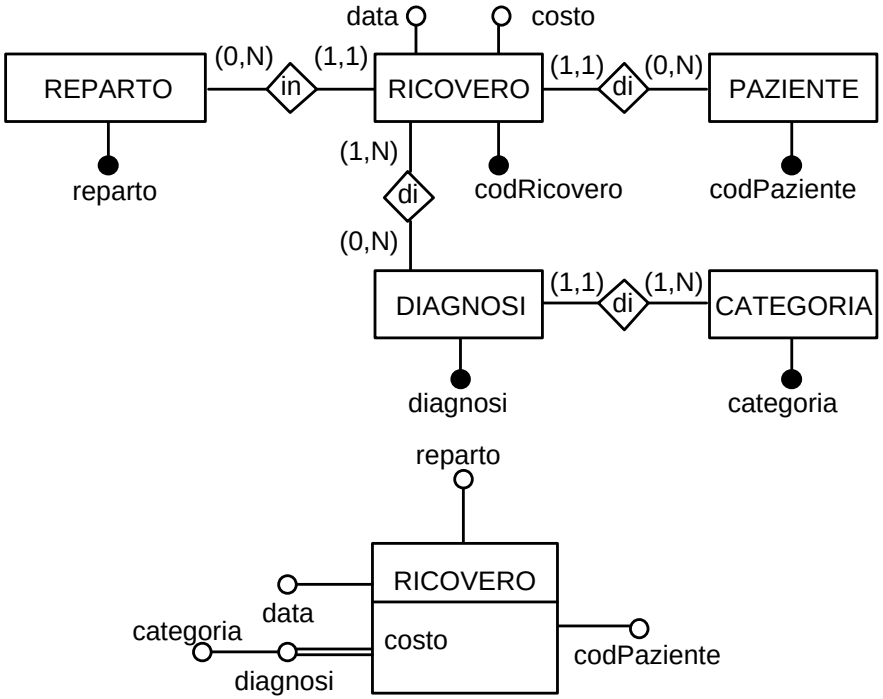


Creazione dello schema di fatto

- Eventuali attributi cross-dimensionali e archi multipli possono essere evidenziati in questa fase
  - ✓ Identificare queste tipologie di attributi a partire dallo schema sorgente è complesso, poiché richiede di navigare anche le associazioni a-molti, per cui si preferisce definirli a partire dai requisiti utente per rappresentarli solo successivamente sullo schema di fatto
    - Un attributo cross-dimensionale corrisponde in genere a un attributo posto su un'associazione molti-a-molti R dello schema E/R; i suoi padri nello schema di fatto corrisponderanno allora agli identificatori delle entità coinvolte in R
    - Un arco multiplo corrisponde a un'associazione a-molti R da un'entità E a un'entità G; nello schema di fatto, esso potrà allora connettere l'identificatore di E o il fatto con un attributo di R o di G

Creazione dello schema di fatto





Creazione dello schema di fatto

- In questa fase devono anche essere identificate le eventuali non-additività e non-aggregabilità presenti nello schema, considerando tutte le accoppiate dimensione-misura
- Dato uno schema di fatto n-dimensionale, per la dimensione d_i e la misura m_j, la domanda da porsi sarà:

  "Siano $\{val_1..., val_k\}$ i valori assunti dalla misura m_j nei k eventi primari corrispondenti a k differenti valori presi dal dominio della dimensione di e da un valore prefissato di ciascuna delle altre n–1 dimensioni. Volendo caratterizzare complessivamente i k eventi con un unico valore di m_j, quali operatori di aggregazione ha senso utilizzare?"

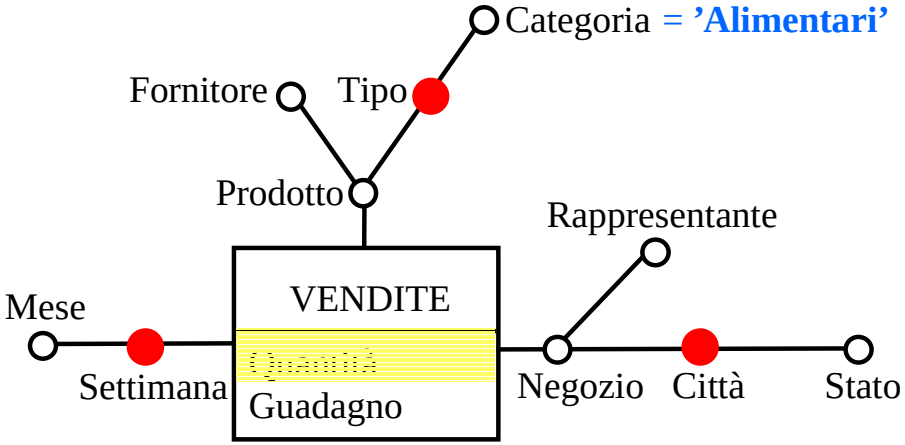
Carico di lavoro e volume dati

Prof. Stefano Rizzi

Il carico di lavoro

- Il carico di lavoro di un sistema OLAP è per sua natura estemporaneo
- È necessario identificare in fase di progettazione un carico di lavoro di riferimento
  - ✓ Reportistica standard
  - ✓ Colloqui con gli utenti
- Le interrogazioni OLAP sono facilmente caratterizzabili
  - ✓ Pattern di aggregazione
  - ✓ Misure richieste
  - ✓ Clausole di selezione

Il carico di lavoro



VENDITE(Negozio.Città, Settimana, Prodotto.Tipo;
Prodotto.Categoria='Alimentari').Quantità

Totale della quantità venduta per i diversi tipi di prodotto, in ogni settimana e città ma solo per i prodotti alimentari

Dinamicità del carico di lavoro

- Il carico di lavoro preliminare non è di per sé sufficiente a ottimizzare le prestazioni del sistema
  - ✓ L'interesse degli utenti cambia nel tempo
  - ✓ Il numero di interrogazioni aumenta al crescere della confidenza degli utenti con il sistema
- Per ottimizzare la struttura logica del data mart è necessaria una fase di tuning attuabile solo dopo che il sistema è stato messo in funzione
- Il carico di lavoro reale può essere desunto dal log delle interrogazioni sottoposte al sistema

Il volume dati

- Consiste nelle informazioni necessarie a determinare/stimare la dimensione del data mart.
  - ✓ Numero di valori distinti degli attributi nelle gerarchie
  - ✓ Lunghezza degli attributi
  - ✓ Numero di eventi di ogni fatto
- Deve essere calcolato considerando la quantità di dati necessari a coprire l'intervallo temporale deciso per il data mart.

Il volume dati

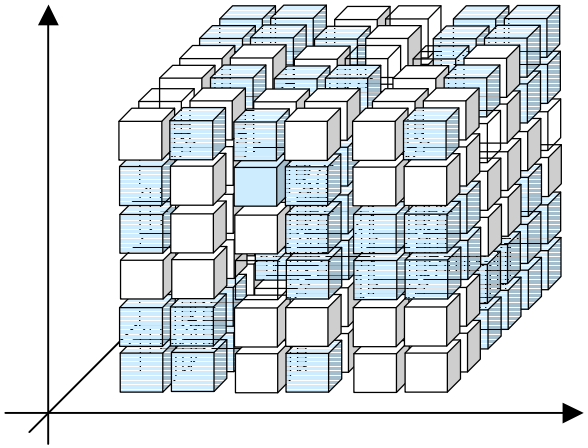
- È utilizzato sia durante la progettazione logica sia durante la progettazione fisica per determinare:
  - ✓ la dimensione delle tabelle
  - ✓ la dimensione degli indici
  - ✓ i costi di accesso
- La bontà delle stime è spesso compromessa a causa del problema della sparsità.

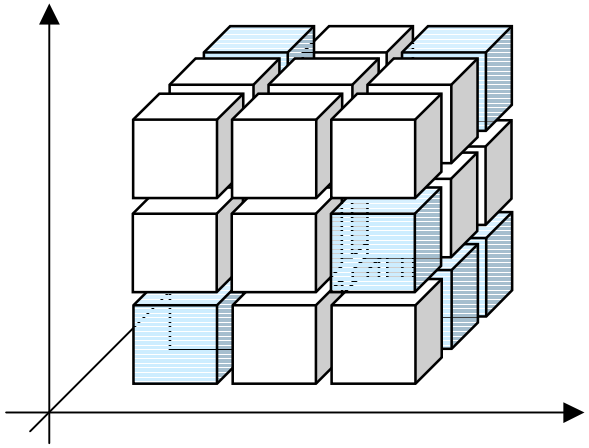
Il problema della sparsità

- Nel modello multidimensionale, a un insieme di coordinate corrisponde un possibile evento anche se questo non è realmente avvenuto
- Normalmente il numero di eventi accaduti è di gran lunga inferiore a quelli possibili
- Tenere traccia degli eventi non accaduti comporta uno spreco di risorse e riduce le prestazioni del sistema
  - ✓ ROLAP: memorizza solo gli eventi accaduti
  - ✓ MOLAP: richiede tecniche complesse per ridurre al minimo lo spazio necessario a tenere traccia degli eventi non accaduti

Il problema della sparsità

- La sparsità dei dati inficia le stime sulla cardinalità dei dati aggregati





- La sparsità si riduce all'aumentare del livello di aggregazione dei dati