

Corso di Laurea Magistrale in Economia

Data Science

A.A. 201/2019

Lez. 4 – Data Warehousing

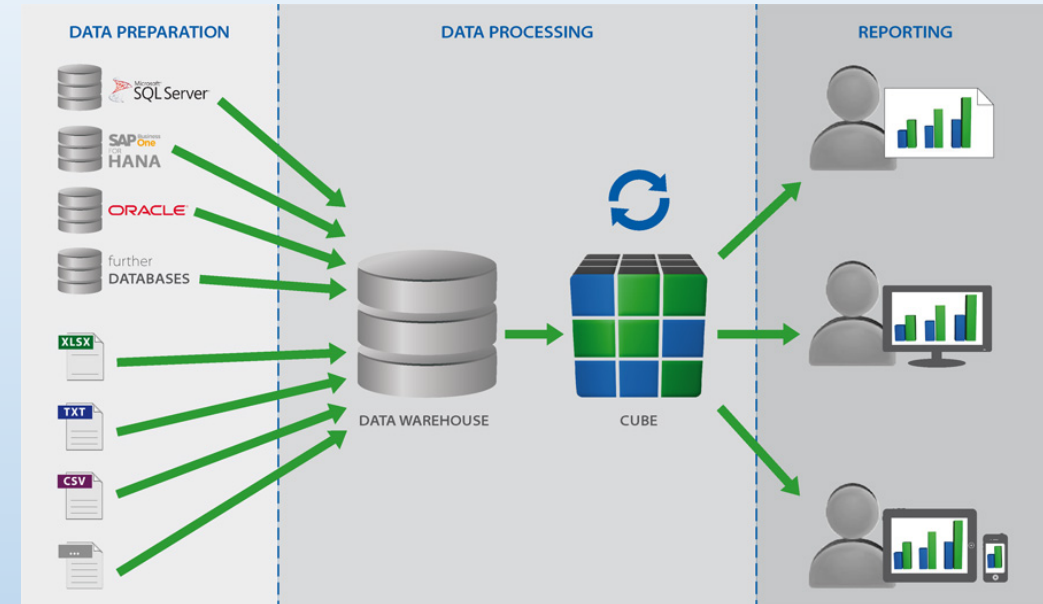
Definizione Data Warehouse

«Deposito di dati» disponibili per sviluppare analisi di *Business Intelligence*

Collezione di dati di supporto per il processo decisionale che presenta le seguenti caratteristiche:

- Orientata ai soggetti di interesse
- Integrata e consistente
- Rappresentativa dell'evoluzione temporale e non volatile

Data Warehousing: complesso di attività riguardanti la progettazione, la realizzazione e l'utilizzo di *data warehouse*



Fabbisogno di Data Warehouse

- Abbiamo montagne di dati ma non possiamo accedervi!!!
- Come è possibile che persone che svolgono lo stesso ruolo presentino risultati sostanzialmente diversi???

Tipologie di dati

- **Interni**

- Conservati nei «sistemi operazionali» interni
- Raccolti mediante programmi applicativi gestionali (ERP)
- Riguardano le principali entità che intervengono nei processi aziendali (clienti, ordini, fornitori, ecc.)
- Provengono da sistemi di *back-office*, *front-office*, web

- **Esterni**

- Fonti eterogenee che estendono i dati interni (vendite e quote mercato per settori, indagini di mercato)
- Sistemi informativi geografici (GIS)

- **Personalizzati**

- Dati non condivisi
- Obiettivo del *knowledge management* è il recupero e l'integrazione di tali dati con quelli interni ed esterni (strutturati)

Data Warehouse vs. database aziendali

I Data Warehouse sostengono le applicazioni di *Business Intelligence*, anche dette *on-line analytical processing* (OLAP).

Perché realizzare DW distinti dai DB aziendali:

- **Integrazione** di dati provenienti da diverse fonti, necessari per i DSS
- **Qualità** dei dati trasferiti dai sistemi operazionali, che vengono esaminati e corretti
- **Efficienza** delle interrogazioni in termini di risorse di calcolo e tempo di elaborazione
- **Estensione** nel tempo dei dati, che nei sistemi operazionali vengono rimossi e conservati su supporti distinti

Caratteristiche Data Warehouse

Collezione di dati a supporto dei processi decisionali e delle analisi di BI

- **Orientata alle entità**
- **Integrata**
- **Tempificata**
 - I dati sono accompagnati da un'etichetta temporale
- **Persistente**
 - Una volta inseriti in un DW i dati non vengono modificati
- **Consolidata**
 - Spesso dati ottenuti come somme parziali di dati elementari
 - Riduzione spazio
 - Dati più coerenti con le analisi di BI
- **Denormalizzata**
 - Dati non strutturati in forma normale ma con ridondanze per consentire tempi di risposta rapidi

Data Mart

Sistema che raccoglie i dati richiesti da una specifica funzione aziendale (marketing, logistica, amministrazione, ecc.)

Data Mart = Data Warehouse «dipartimentale»

- Dimensioni contenute
- Specifico per la funzione
- Stessa matrice tecnologica
- Spesso le aziende preferiscono realizzare diversi Data Mart in modo incrementale piuttosto che un Data Warehouse centrale

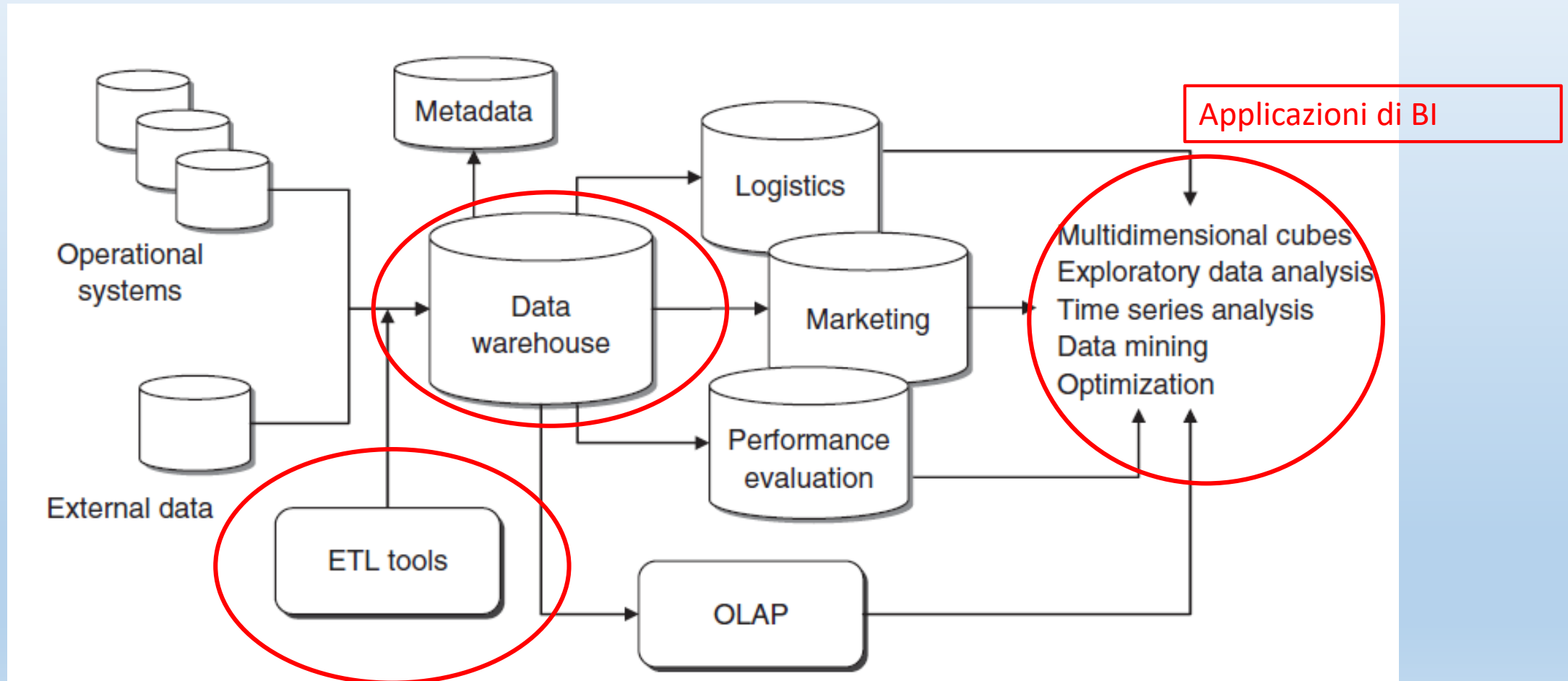
Qualità dei dati

Problem	Cause	Remedy
incorrect data	data collected without due care data entered incorrectly uncontrolled modification of data	systematic checking of input data data entry automation implementation of a safety program for access and modifications
data not updated	data collection does not match user needs	timely updating and collection of data retrieval of updated data from the web
missing data	failure to collect the required data	identification of data needed via preliminary analysis and estimation of missing data

Fattori che influenzano la qualità dei dati

- Accuratezza
- Completezza
- Consistenza
- Attualità
- Non ridondanza
 - Sprechi di memoria
 - Inconsistenze
 - Eccezioni per specifiche esigenze di interrogazione
- Rilevanza
- Interpretabilità
- Accessibilità

Architetture di Data Warehouse



Livelli tecnologici

- Livello delle fonti dei dati e degli strumenti di ETL, presenti su uno o più server
- Livello del Data Warehouse e/o dei Data Mart, su uno o più server distinti da quelli delle fonti, in cui sono presenti anche i metadati che documentano l'origine e il significato delle informazioni
- Livello di analisi, con applicativi per il supporto decisionale, su server distinti o sui client degli analisti

Logiche per la realizzazione di un DW

- Top-down
 - Disegno complessivo del Data Warehouse
 - Tempi più elevati
 - Maggiori rischi di realizzazione
- Bottom-up
 - Per ampliamenti successivi
 - Più rapida
 - Manca la visione d'insieme
- Mista
 - Progetto complessivo
 - Realizzazione di prototipi successivi per parti diverse del sistema

Strumenti ETL

- **Estrazione**

- Prima estrazione che riempie un DW vuoto
- Estrazioni incrementali per l'aggiornamento dei dati
- Selezione dei dati che dipende dalle esigenze delle analisi di BI

- **Trasformazione**

- Serve a migliorare la qualità dei dati
- «Pulitura» con regole automatiche per la correzione di errori
- Conversione dei dati
- Calcoli per aggregazione dei dati

- **Caricamento**

- Secondo lo schema previsto

Metadati

Struttura informativa necessaria a documentare:

- La struttura del Data Warehouse
 - Schema
 - Viste logiche
 - Dimensioni
 - Gerarchie
 - Localizzazione dei Data Mart
- La genealogia dei dati
 - Origine
 - Trasformazioni subite
- Le statistiche relative all'utilizzo del Data Warehouse
- Il significato del Data Warehouse rispetto al contesto applicativo
 - Definizione dei termini impiegati
 - Le proprietà dei dati
 - Le politiche di caricamento

Il modello relazionale

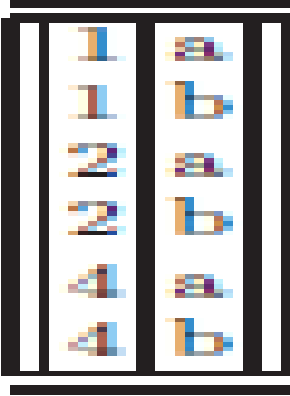
- Dati due insiemi D_1 e D_2 , si chiama prodotto cartesiano di D_1 e D_2 , e si indica con $D_1 \times D_2$, l'insieme delle coppie ordinate (v_1, v_2) tali che $v_1 \in D_1$ e $v_2 \in D_2$.
- Ad esempio, dati gli insiemi $A = \{1, 2, 4\}$ e $B = \{a, b\}$, il prodotto cartesiano $A \times B$ è dato da:

$$A \times B = \{(1, a), (1, b), (2, a), (2, b), (4, a), (4, b)\}$$

- Una relazione matematica su D_1 e D_2 (chiamati domini della relazione) è un sottoinsieme di $D_1 \times D_2$
- Ad esempio, dati gli insiemi A e B di cui sopra, una possibile relazione matematica su A e B è costituita dall'insieme delle coppie $\{(1, a), (1, b), (4, b)\}$


Rappresentazione tabellare

- Prodotto cartesiano $A \times B$:



1	a
1	b
2	a
2	b
4	a
4	b

- Relazione matematica su A e B:



1	a
1	b
4	b

Relazioni su n domini

- Dati $n > 0$ insiemi D_1, D_2, \dots, D_n , non necessariamente distinti, il prodotto cartesiano di D_1, D_2, \dots, D_n , indicato con $D_1 \times D_2 \times \dots \times D_n$ è costituito dall'insieme delle n -uple (v_1, v_2, \dots, v_n) tali che $v_i \in D_i$, per $1 \leq i \leq n$
- Una relazione matematica sui domini D_1, D_2, \dots, D_n è un sottoinsieme del prodotto cartesiano $D_1 \times D_2 \times \dots \times D_n$
- Il numero n dei domini coinvolti in un prodotto cartesiano viene detto grado del prodotto cartesiano e della relazione
- Il numero di n -uple della relazione viene denominato cardinalità della relazione

Relazioni con attributi

- Una relazione matematica è un insieme di n -uple ordinate (v_1, v_2, \dots, v_n) tali che $v_1 \in D_1, v_2 \in D_2, \dots, v_n \in D_n$; ciascuna n -upla stabilisce un legame tra i corrispondenti dati
- Dal momento che la relazione è un insieme non è definito alcun ordinamento tra le n -uple; due tabelle con le stesse righe, ma in ordine diverso, rappresentano la stessa relazione
- Le n -uple di una relazione sono distinte l'una dall'altra in quanto tra gli elementi di un insieme non possono essere presenti due elementi uguali
- Ciascuna n -upla è, al proprio interno, ordinata; l' i -esimo valore di ciascuna n -upla appartiene all' i -esimo dominio

Relazioni con attributi

- Questo ordinamento fra i domini di una relazione corrisponde ad una caratteristica insoddisfacente del concetto di relazione matematica rispetto alla possibilità di organizzare e utilizzare i dati
- In vari contesti dell'informatica si tende a privilegiare notazioni non posizionali
- Le informazioni che siamo interessati ad organizzare nelle relazioni delle nostre basi di dati hanno una struttura che si può naturalmente ricondurre a quella dei record
- Rifacendosi ai record si può associare a ciascuna occorrenza di dominio della relazione un nome, detto attributo, che descrive il ruolo giocato dal dominio stesso.
- Gli attributi di una relazione devono avere tutti nomi diversi tra di loro

Rappresentazione tabellare

- Attributi come intestazioni delle colonne:

SQUADRA DI CASA	SQUADRA OSPITATA	RETI CASA	RETI OSPITATA
Juventus	Lazio	3	1
Lazio	Milan	2	0
Juventus	Roma	1	2
Roma	Milan	0	1

- A questo punto l'ordinamento degli attributi risulta irrilevante
- Ad esempio, la relazione precedente è equivalente alla seguente relazione:

SQUADRA OSPITATA	SQUADRA DI CASA	RETI OSPITATA	RETI CASA
Lazio	Juventus	1	3
Milan	Lazio	0	2
Roma	Juventus	2	1
Milan	Roma	1	0

Relazioni con attributi

- Per formalizzare questi concetti, stabiliamo la corrispondenza tra attributi e domini per mezzo di una funzione $DOM : X \rightarrow D$, che associa a ciascun attributo $A \in X$ un dominio $DOM(A) \in D$
- Una *tupla* su un insieme di attributi X è una funzione t che associa a ciascun attributo $A \in X$ un valore del dominio $DOM(A)$
- Una relazione su X è un insieme di tuple su X
- La differenza tra questa definizione e quella tradizionale di relazione matematica risiede solo nella definizione di tupla
 - Nella relazione matematica, gli elementi delle n-uple vengono individuati per posizione
 - Nella nuova definizione di relazione gli elementi delle tuple vengono individuati per mezzo degli attributi, ovvero con una tecnica non posizionale
 - Se t è una tupla su X e $A \in X$, $t[A]$ oppure $t.A$ indica il valore di t su A
- Ad esempio, se t è la prima tupla della tabella precedente, allora si ha che $t[\text{SquadraOspitata}] = \text{Lazio}$

Relazioni e basi di dati

- Una relazione può essere utilizzata per organizzare dati rilevanti nell'ambito di un'applicazione di interesse
- Una base di dati è, in generale, costituita da più relazioni le cui tuple contengono valori comuni, ove necessario, per stabilire corrispondenze
- Consideriamo la seguente basi di dati relativa ad un'università:

STUDENTI

MATRICOLA	COGNOME	NOME	DATA DI NASCITA
276545	Rossi	Maria	25/11/1981
485745	Neri	Anna	23/04/1982
200768	Verdi	Fabio	12/02/1982
587614	Rossi	Luca	10/10/1981
997659	Bruni	Mario	01/12/1981

ESAMI

STUDENTE	VOTO	CORSO
276545	28	01
276545	27	04
997659	25	01
200768	24	04

CORSI

CODICE	TITOLO	DOCENTE
01	Analisi	Ciani
03	Chimica	Melli
04	Chimica	Belli

Relazioni e basi di dati

- La prima relazione contiene informazioni relative ad un insieme di studenti
- La terza relazione contiene informazioni su alcuni corsi
- La seconda relazione contiene informazioni relative ad esami; essa fa riferimento ai dati contenuti nelle altre due
- Il modello relazionale è basato sui valori: i riferimenti fra dati in relazioni diverse vengono rappresentati per mezzo di valori di domini che compaiono nelle tuple
- Gli altri modelli logici, reticolare e gerarchico, vengono detti modelli basati su record e puntatori

Modello relazionale

- Uno schema di relazione è costituito da un simbolo, R , detto nome della relazione, e da un insieme di attributi $X = \{ A_1, A_2, \dots, A_n \}$, il tutto di solito indicato con $R(X)$. A ciascun attributo è associato un dominio, come visto in precedenza
- Uno schema di base di dati è un insieme di schemi di relazione con nomi diversi:

$$R = \{ R_1(X_1), R_2(X_2), \dots, R_n(X_n) \}$$

- Un'istanza di relazione (o, semplicemente, relazione) su uno schema $R(X)$ è un insieme r di tuple su X
- Un'istanza di base di dati (o, semplicemente, base di dati) su uno schema $R = \{ R_1(X_1), R_2(X_2), \dots, R_n(X_n) \}$ è un insieme di relazioni $r = \{ r_1, r_2, \dots, r_n \}$, dove ogni r_i , per $1 \leq i \leq n$, è una relazione sullo schema $R_i(X_i)$
- Ad esempio, lo schema della base di dati della figura precedente è così definito:

$$R = \{ \text{STUDENTI (Matricola, Cognome, Nome, Data di Nascita),} \\ \text{ESAMI (Studente, Voto, Corso), CORSI (Codice, Titolo, Docente) } \}$$

Modello relazionale

- Nel seguito adotteremo alcune convenzioni allo scopo di favorire la sinteticità della notazione:
 - Gli attributi verranno indicati con lettere iniziali maiuscole dell'alfabeto, eventualmente con indici e/o pedici: A, B, C, A', A_1, \dots
 - Gli insiemi di attributi verranno indicati con lettere finali maiuscole dell'alfabeto X, Y, Z, X', X_1 . Un insieme di cui si vogliono evidenziare gli attributi che lo compongono si indicherà con $X = ABC$ invece che $X = \{A, B, C\}$
 - L'unione dei due insiemi verrà denotata dalla giustapposizione dei relativi nomi: scriveremo XY anziché $X \cup Y$; scriveremo, infine, XA anziché $X \cup \{A\}$
 - Per i nomi di relazione utilizzeremo la R e le lettere circostanti maiuscole: R_1, S, S', \dots ; per le relazioni utilizzeremo gli stessi simboli dei corrispondenti nomi di relazione, ma in lettere minuscole

Informazione incompleta e valori nulli

- La struttura del modello relazionale introdotta finora impone un certo grado di rigidità, in quanto le informazioni debbono essere rappresentate per mezzo di tuple di dati omogenee
- In molti casi i dati disponibili possono non corrispondere esattamente al formato previsto
- Ad esempio, nella seguente relazione, il valore dell'attributo Telefono potrebbe non essere disponibile per tutte le tuple:

PERSONE (Cognome, Nome, Indirizzo, Telefono)

- È importante notare che non sarebbe corretto utilizzare un valore del dominio per rappresentare l'assenza di informazione
- Per i numeri telefonici rappresentato per mezzo di interi potremmo utilizzare lo zero per indicare l'assenza di un valore significativo

Informazione incompleta e valori nulli

- Questa scelta non risulta soddisfacente per due motivi:
 - Essa richiede l'esistenza di un valore del dominio mai utilizzato per valori significativi
 - L'uso dei valori del dominio può generare confusione: la distinzione tra valori veri e valori fittizi è nascosta
 - Per rappresentare in modo semplice la non disponibilità di valori si utilizza un valore speciale detto NULL
- Il valore NULL è un valore aggiuntivo rispetto a quelli del dominio e ben distinto da essi
- I valori NULL, tuttavia, devono essere utilizzati con molta cautela

Informazione incompleta e valori nulli

- Consideriamo la seguente base di dati:

	MATRICOLA	COGNOME	NOME	DATA DI NASCITA
STUDENTI	276545	Rossi	Maria	NULL
	NULL	Neri	Anna	29/04/1972
	NULL	Verdi	Fabio	12/02/1972

	STUDENTE	VOTO	CORSO
ESAMI	276545	28	01
	NULL	27	NULL
	200768	24	NULL

	CODICE	TITOLO	DOCENTE
CORSI	01	Analisi	Ciani
	09	Chimica	NULL
	NULL	Chimica	Belli

Informazione incompleta e valori nulli

- Il valore nullo sulla data di nascita nella prima tupla della relazione STUDENTI è ammissibile
- Un valore nullo sul numero di matricola o sul codice di corso genera problemi maggiori
- La presenza di valori nulli nella relazione ESAMI rende, addirittura, inutilizzabili le informazioni
- La presenza di molteplici valori nulli in una relazione può, addirittura, generare dubbi sull'effettiva significatività e identità delle tuple
- È necessario, quindi, controllare opportunamente la presenza di valori nulli nelle relazioni

Introduzione ai vincoli di integrità

- Non è vero che qualsiasi insieme di tuple su uno schema rappresenta informazioni corrette per l'applicazione
- Consideriamo la seguente base di dati:

MATRICOLA	COGNOME	NOME	DATA DI NASCITA
200768	Verdi	Fabio	12/02/1972
997659	Rossi	Luca	10/10/1971
997659	Bruni	Mario	01/12/1971

STUDENTE	VOTO	LODE	CORSO
200768	36	00	05
997659	28	01	01
997659	30	01	04
276545	25	00	01

CODICE	TITOLO	DOCENTE
01	Analisi	Giani
03	Chimica	Melli
04	Chimica	Belli

Introduzione ai vincoli di integrità

- Nella prima tupla della relazione ESAMI abbiamo un voto pari a 36
- Nella seconda tupla della relazione ESAMI viene indicato che è stata attribuita la lode in un esame il cui voto è 28
- Le ultime due tuple della relazione STUDENTI contengono informazioni su due studenti diversi con lo stesso numero di matricola
- La quarta tupla della relazione ESAMI presenta, per l'attributo STUDENTE, un valore che non compare fra i numeri di matricola nella relazione STUDENTI
- Analogamente la prima tupla presenta un codice di corso che non compare nella relazione CORSI
- Per evitare situazioni come quelle descritte è stato introdotto il concetto di vincolo di integrità
- Ciascun vincolo può essere visto come un predicato che associa ad ogni istanza il valore TRUE o FALSE

Introduzione ai vincoli di integrità

- In generale, ad uno schema di basi di dati associamo un insieme di vincoli e consideriamo corrette le istanze che soddisfano tutti i vincoli
- È possibile classificare i vincoli in quattro categorie, a seconda degli elementi di una base di dati che ne sono coinvolti:
 - Un vincolo di dominio, o vincolo sui valori, viene definito con riferimento ai singoli valori
 - Un vincolo di tupla può essere valutato su ciascuna tupla indipendentemente dalle altre
 - Un vincolo è intrarelazionale se il suo soddisfacimento è definito rispetto alle singole relazioni della base di dati
 - Un vincolo è interrelazionale se coinvolge più relazioni; i più importanti vincoli interrelazionali sono i vincoli di integrità referenziale
 - Un vincolo di dominio è un caso particolare di vincolo di tupla che, a sua volta, è un caso particolare di vincolo intrarelazionale

Vincoli di dominio e vincoli di tupla

- I vincoli di dominio esprimono condizioni sui valori di un attributo di una tupla, indipendentemente dai valori degli altri attributi della tupla stessa
- I vincoli di tupla esprimono condizioni sui valori degli attributi di una tupla, indipendentemente dai valori degli attributi delle altre tuple
- Una possibile sintassi per esprimere vincoli di questo tipo è quella che permette di definire espressioni booleane
- I vincoli violati individuati nei primi due esempi della base di dati precedente potrebbero essere descritti con le seguenti espressioni:
 - $(\text{Voto} \geq 18) \text{ AND } (\text{Voto} \leq 30)$
 - $(\text{NOT } (\text{Lode} = 01)) \text{ OR } (\text{Voto} = 30)$
 - Il primo vincolo è un vincolo di dominio
- Il secondo vincolo è un vincolo di tupla

Vincoli di dominio e vincoli di tupla

- La definizione di vincolo di tupla che abbiamo dato ammette anche espressioni più complesse, purché definite sui valori delle singole tuple
- Ad esempio, data la seguente relazione:

PAGAMENTI (Data, Importo, Ritenute, Netto)

è possibile definire il seguente vincolo:

$$\text{Netto} = \text{Importo} - \text{Ritenute}$$

- La presenza di un vincolo di tupla aritmetico può essere indice di una progettazione errata della base di dati

Vincoli chiave

- Consideriamo la seguente figura:

MATRICOLA	COGNOME	NOME	DATA DI NASCITA	CORSO
4928	Rossi	Luigi	29/04/79	Ing. Informatica
6328	Rossi	Dario	29/04/79	Ing. Informatica
4766	Rossi	Luca	01/05/81	Ing. Meccanica
5996	Neri	Luca	05/09/78	Ing. Meccanica
4856	Neri	Luca	01/05/81	Ing. Meccanica

- I valori delle varie tuple sull'attributo MATRICOLA sono tutti diversi l'uno dall'altro
- Se la relazione si riferisce ad una sola università il valore della matricola identifica univocamente gli studenti
- Nella relazione non vi sono coppie di tuple con gli stessi valori su ciascuno dei tre attributi Cognome, Nome e Data di Nascita
- Anche altri insiemi di attributi identificano univocamente le tuple della relazione, ad esempio Matricola e Corso

Vincoli chiave

- Tutto questo ragionamento può essere formalizzato come di seguito specificato:
 - Un insieme K di attributi è superchiave per una relazione r se r non contiene due tuple distinte t_1 e t_2 con $t_1[K] = t_2[K]$
 - Un insieme K di attributi è chiave per r se è una superchiave minimale (ovvero non esiste un'altra superchiave K' di r tale che K' è un sottoinsieme proprio di K)
 - In base a questa definizione possiamo asserire che:
 - L'insieme {Matricola} è superchiave ed essendo minimale è anche chiave
 - L'insieme {Cognome, Nome, DataNascita} è superchiave ed essendo minimale è anche chiave
 - L'insieme {Matricola, Corso} è superchiave ma, non essendo minimale, non è chiave
 - L'insieme {Nome, Corso} non è superchiave
 - L'insieme di attributi {Nome, DataNascita, Corso} è casualmente una chiave per la relazione precedentemente

Vincoli chiave

- I vincoli sono definiti a livello di schema e devono essere rispettati da tutte le istanze dello schema
- Un'istanza corretta può poi soddisfare altri vincoli oltre a quelli definiti sullo schema
- Ad esempio, ad uno schema

STUDENTI (Matricola, Cognome, Nome, DataNascita, Corso)

vanno associati i vincoli che impongono come chiavi i due insiemi di attributi:

{Matricola}

{Cognome, Nome, DataNascita}

- La relazione vista sopra soddisfa entrambi i vincoli
- Essa casualmente soddisfa anche il vincolo secondo cui {Nome, DataNascita, Corso} è un'altra chiave

Vincoli chiave

- Ciascuna relazione e ciascuno schema di relazione hanno sempre una chiave
- Per ogni relazione $r(x)$, l'insieme x di tutti gli attributi su cui è definita è senz'altro una superchiave per essa
- Se l'insieme di tutti i suoi attributi è minimale, essa costituisce anche una chiave
- Se, invece, tale insieme non è minimale vuol dire che esiste un'altra superchiave in essa contenuta; allora è possibile procedere ricorsivamente
- Lo stesso ragionamento vale a livello di schema di relazione in quanto l'insieme di tutti i suoi attributi è superchiave per ciascuna relazione lecita
- La presenza delle chiavi garantisce l'identificabilità di tutti i valori di una base di dati

Vincoli chiave

- Le chiavi permettono, inoltre, di stabilire efficacemente quelle corrispondenze fra dati contenuti in relazioni diverse che caratterizzano il modello relazionale come “modello basato sui valori”
- I valori attraverso cui vengono stabilite le corrispondenze fra tuple di relazioni diverse sono proprio i valori delle chiavi delle relazioni

Vincoli chiave e valori nulli

- In presenza di valori nulli non è più vero che i valori delle chiavi permettono di identificare univocamente le tuple delle relazioni e di stabilire riferimenti tra tuple di relazioni diverse
- Consideriamo la seguente relazione che dovrebbe avere come chiavi {Matricola} e {Cognome, Nome, DataNascita}

MATRICOLA	COGNOME	NOME	DATA DI NASCITA	CORSO
NULL	ROSSI	Mario	NULL	Informatica
4766	Rossi	Luca	01/05/61	Civile
4856	Neri	Luca	NULL	NULL
NULL	Neri	Luca	05/09/58	Civile

Vincoli chiave e valori nulli

- La prima tupla ha valori nulli su Matricola e DataNascita e, quindi, su almeno un attributo di ciascuna chiave; questa tupla non è identificabile in alcun modo
- Non è possibile, in altre relazioni della stessa base di dati, fare riferimento a questa tupla, visto che ciò andrebbe fatto attraverso il valore di una chiave
- Le ultime due tuple della figura presentano un problema: nonostante ciascuna abbia una chiave completamente specificata, la presenza di valori nulli rende impossibile capire se le due tuple fanno riferimento allo stesso studente oppure no
- L'esempio ci suggerisce, quindi, la necessità di porre dei limiti alla presenza di valori nulli nelle chiavi delle relazioni

Vincoli chiave e valori nulli

- Pertanto, sulla chiave primaria si vieta la presenza di valori nulli; sulle altre i valori nulli sono ammessi
- Gli attributi che costituiscono la chiave primaria vengono generalmente sottolineati
- La maggior parte dei riferimenti tra relazioni vengono realizzati attraverso i valori della chiave primaria
- In quasi tutti i casi reali è sempre possibile trovare tra gli attributi di una relazione una chiave primaria
- Quando ciò non accade è necessario introdurre una chiave surrogata.

Vincoli di integrità referenziale

INFRAZIONI

CODICE	DATA	AGENTE	ART.	TARCA
149256	25/10/02	567	44	CE 543 TY
987554	26/10/02	456	94	CE 543 TY
987557	26/10/02	456	94	BD 764 AJ
690876	15/10/02	456	59	CF 764 KR
599856	12/10/02	567	44	CF 764 KR

AGENTI

MATRICOLA	CF	COGNOME	NOME
567	RSSM...	Rossi	Mario
456	NREL...	Neri	Luigi
638	NREL...	Neri	Piero

AUTO

TARCA	PROPRIETARIO	INDIRIZZO
BD 764 AJ	Verdi Piero	Via Tigli
AJ 224 TY	Verdi Piero	Via Tigli
CE 543 TY	Bini Luca	Via Aceri
CF 764 KR	Luci Gino	Via Aceri

Vincoli di integrità referenziale

- Le informazioni della relazione INFRAZIONI sono rese significative e complete attraverso il riferimento alle altre due relazioni
- I riferimenti sono significativi in quanto i valori della relazione INFRAZIONI sono uguali a valori effettivamente presenti nelle altre due relazioni
- Per formalizzare tutto il ragionamento precedente il modello relazionale prevede il concetto di vincolo di integrità referenziale
- Un vincolo di integrità referenziale fra un insieme di attributi X di una relazione R_1 e un'altra relazione R_2 è soddisfatto se i valori su X di ciascuna tupla dell'istanza di R_1 compaiono come valori della chiave primaria dell'istanza di R_2

Vincoli di integrità referenziale

- Dato un insieme di attributi $X = A_1 A_2 \dots A_p$ di una relazione R_1 e un'altra relazione R_2 , un vincolo di integrità referenziale fra X ed R_2 è soddisfatto se, per ogni tupla t_1 in R_1 , esiste una tupla t_2 in R_2 con $t_1[A_i] = t_2[B_i]$, per ogni i compreso tra 1 e p
- Sullo schema della base di dati precedente ha senso definire i vincoli di integrità referenziale:
 - Fra l'attributo Agente della relazione INFRAZIONI e la relazione AGENTI
 - Fra l'attributo Targa della relazione INFRAZIONI e la relazione AUTO

Vincoli di integrità referenziale

INFRAZIONI	CODICE	DATA	AGENTE	ART.	TARCA
	987554	26/10/02	456	94	BD 764 AJ
	690876	15/10/02	456	53	CW 122 AK

AGENTI	MATRICOLA	CF	COGNOME	NOME
	567	RSSM...	Rossi	Mario
	698	NREL...	Neri	Piero

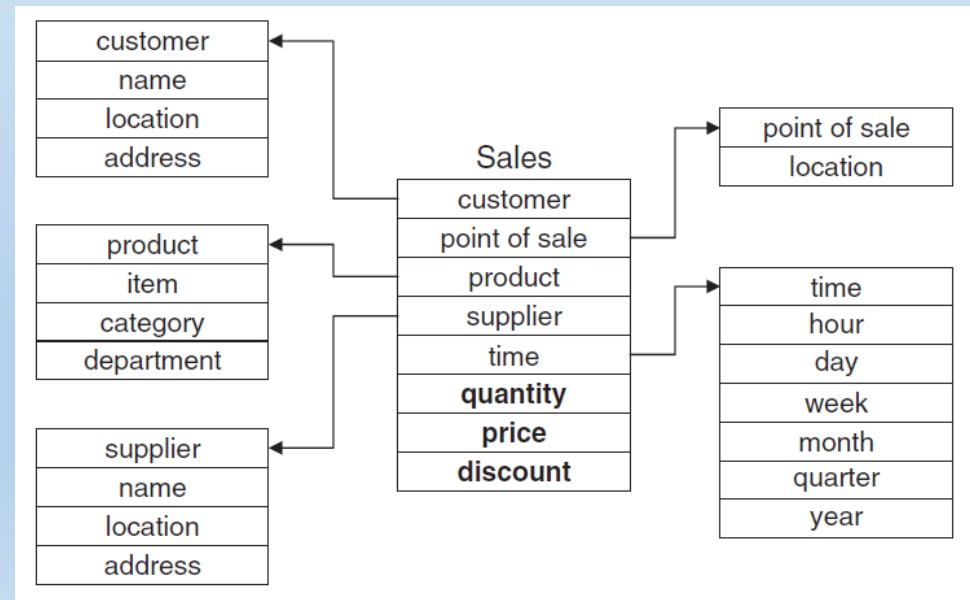
AUTO	TARCA	PROPRIETARIO	INDIRIZZO
	AJ 224 TY	Verdi Piero	Via Tigh
	CX 123 KA	Bini Luca	Via Aceri
	CF 764 KR	Luci Gino	Via Noci

- Tale base di dati viola il vincolo tra Agente della relazione INFRAZIONI e la relazione AGENTI
- Essa viola il vincolo tra l'attributo Targa di INFRAZIONI e la relazione AUTO

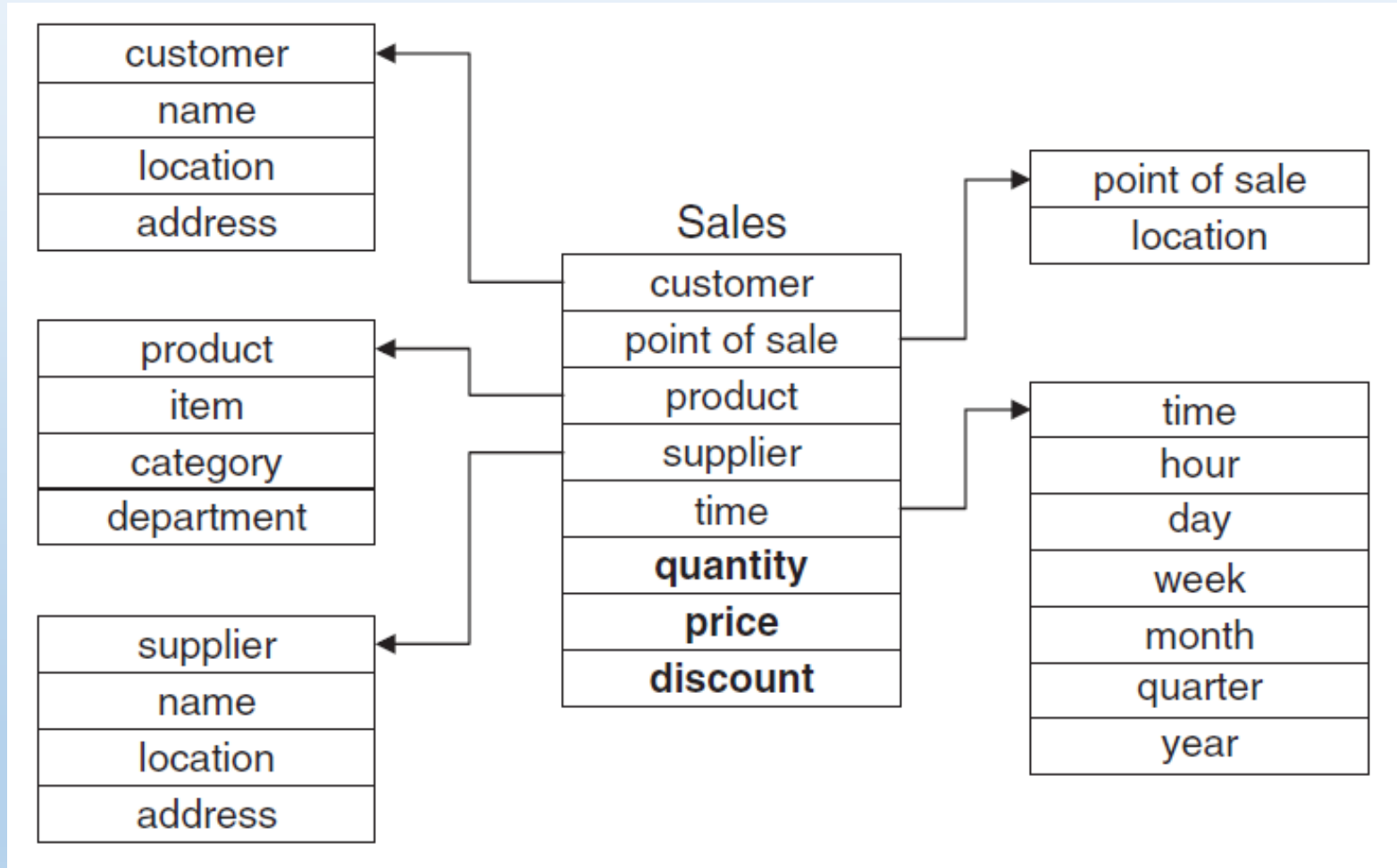
Cubi e analisi multidimensionali

La progettazione si basa su un paradigma di rappresentazione multidimensionale, per due motivi:

- Dal punto di vista funzionale, per garantire tempi di risposta rapidi a fronte di interrogazioni complesse
- Sul piano logico, per garantire la corrispondenza delle dimensioni con i criteri di analisi utilizzati



Schema a stella



Tipi di tabelle

- **Tabelle delle dimensioni**

- Dimensione: entità rilevanti nei processi aziendali (clienti, prodotti, ecc.)
- Ciascuna tabella è strutturata secondo relazioni gerarchiche

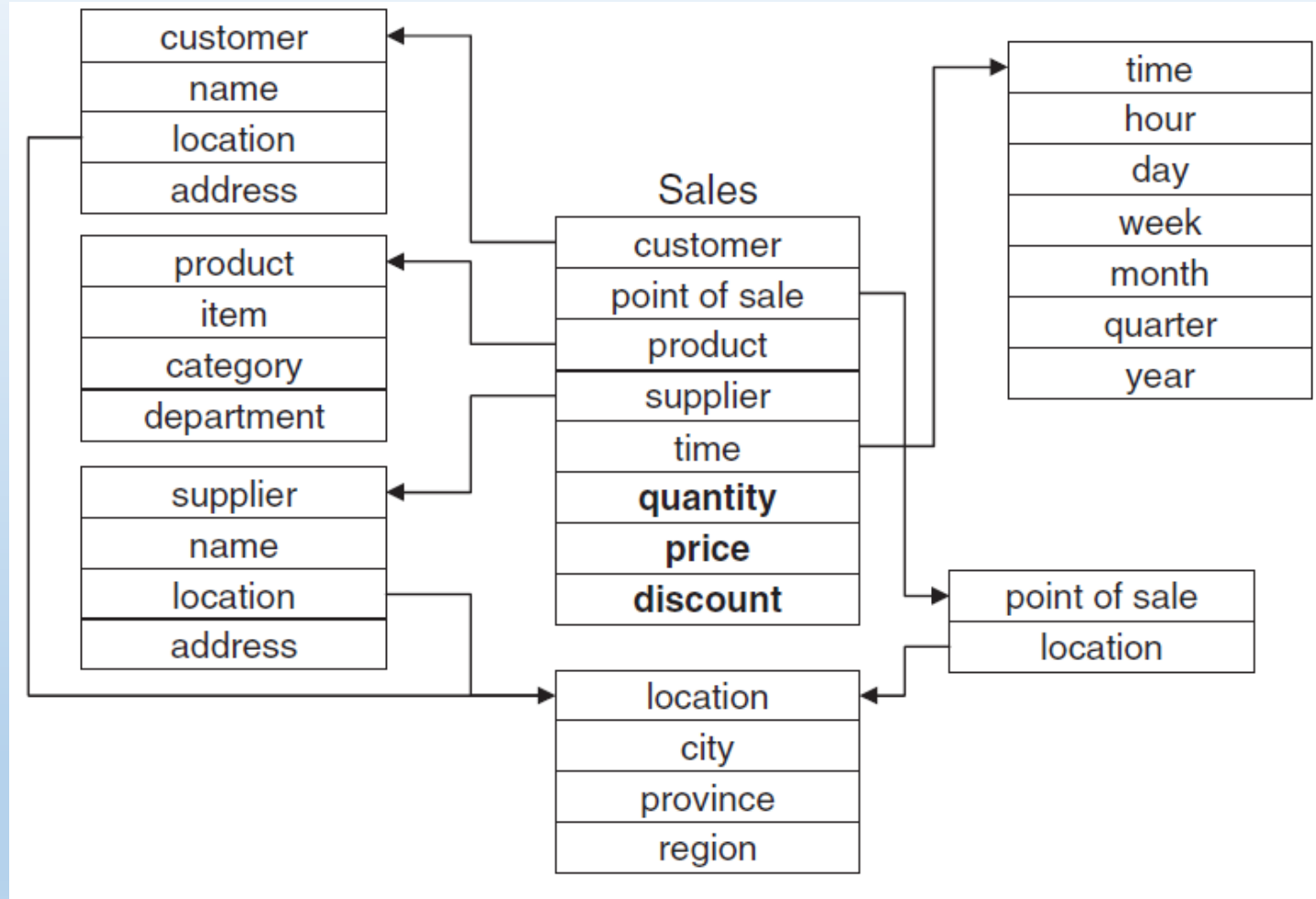
- **Tabelle dei fatti**

- Si riferiscono a transazioni che servono a collocare i riferimenti appropriati per le informazioni contenute in ciascuna tabella dei fatti
- Contengono i valori numerici degli attributi che caratterizzano le transazioni e costituiscono l'oggetto delle successive analisi

Esempio:

- Vendite: tabella dei fatti
- Clienti, punti vendita, prodotti, ecc.: tabelle delle dimensioni
- La tabella dei fatti sta al centro ed è collegata alle tabelle delle dimensioni mediante riferimenti

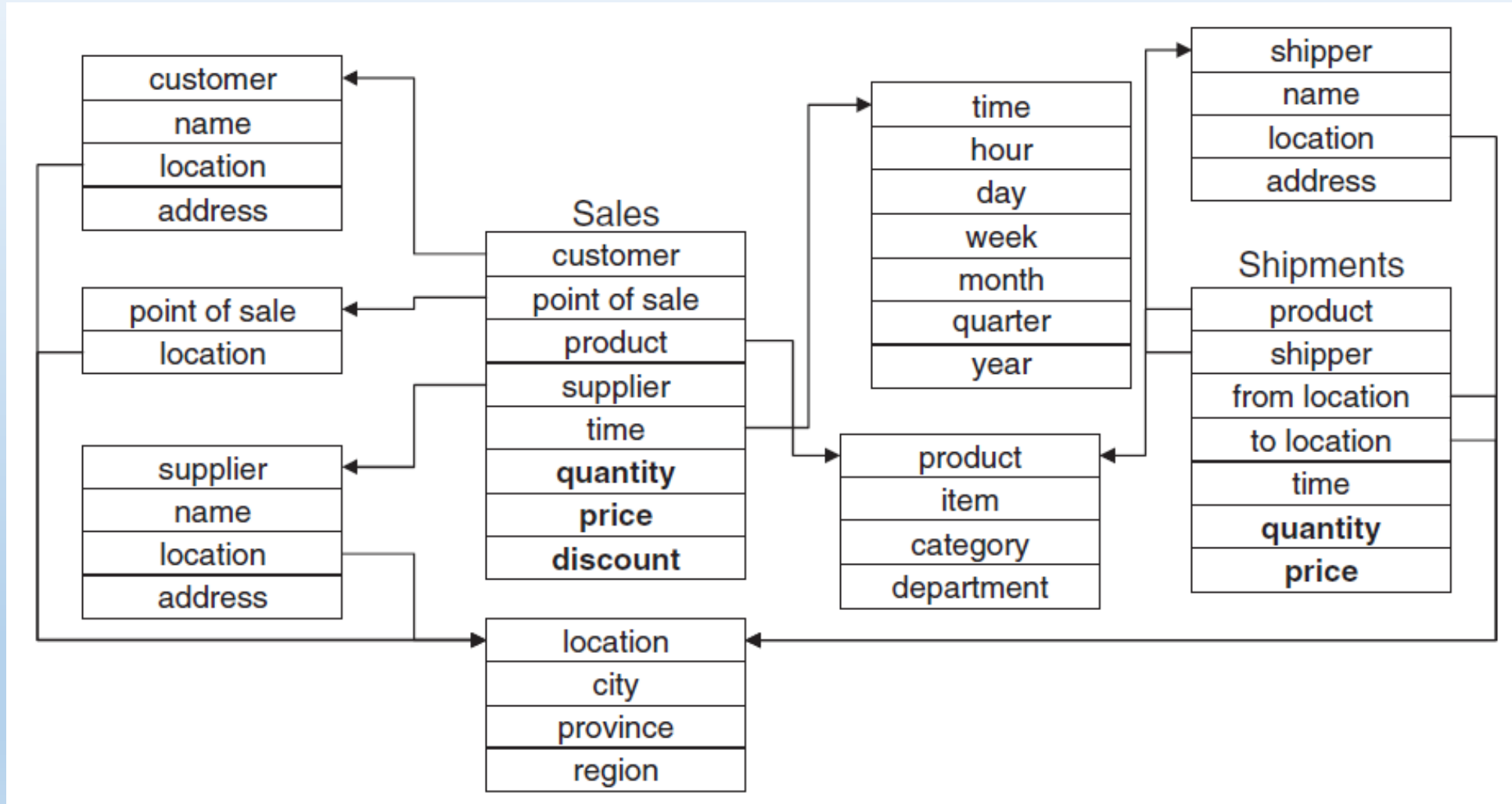
Schema a fiocco di neve



Schema a fiocco di neve

- Le tabelle delle dimensioni sono collegate ad altre tabelle delle dimensioni, mediante un processo di parziale standardizzazione dei dati, per ridurre la memoria richiesta
- Esempio
 - Le tabelle con l'attributo «località» sono collegate alla tabella di dimensione «località» che contiene le informazioni di natura geografica
- Se un Data Warehouse è costituito da numerose tabelle dei fatti collegate a tabelle delle dimensioni, a loro volta collegate ad altre tabelle di dimensioni, si parla di «schema a galassia»

Schema a galassia



Cubo di dati n -dimensionale

- Una tabella dei fatti collegata a n tabelle delle dimensioni si rappresenta mediante un cubo di dati n -dimensionale, dove ogni asse rappresenta una dimensione
- Naturale estensione dei fogli elettronici (cubi bidimensionali)

Esempio di cubo 3-dimensionale

- Tabella dei fatti:
vendita
- Tabelle delle
dimensioni:
 - Tempo
 - Prodotto
 - Regione

Table 3.3 Two-dimensional view of sales data in the USA

region = USA			
time	product		
	TV	PC	DVD
Q1	980	546	165
Q2	765	456	231
Q3	879	481	192
Q4	986	643	203

Table 3.4 Two-dimensional view of sales data in Asia

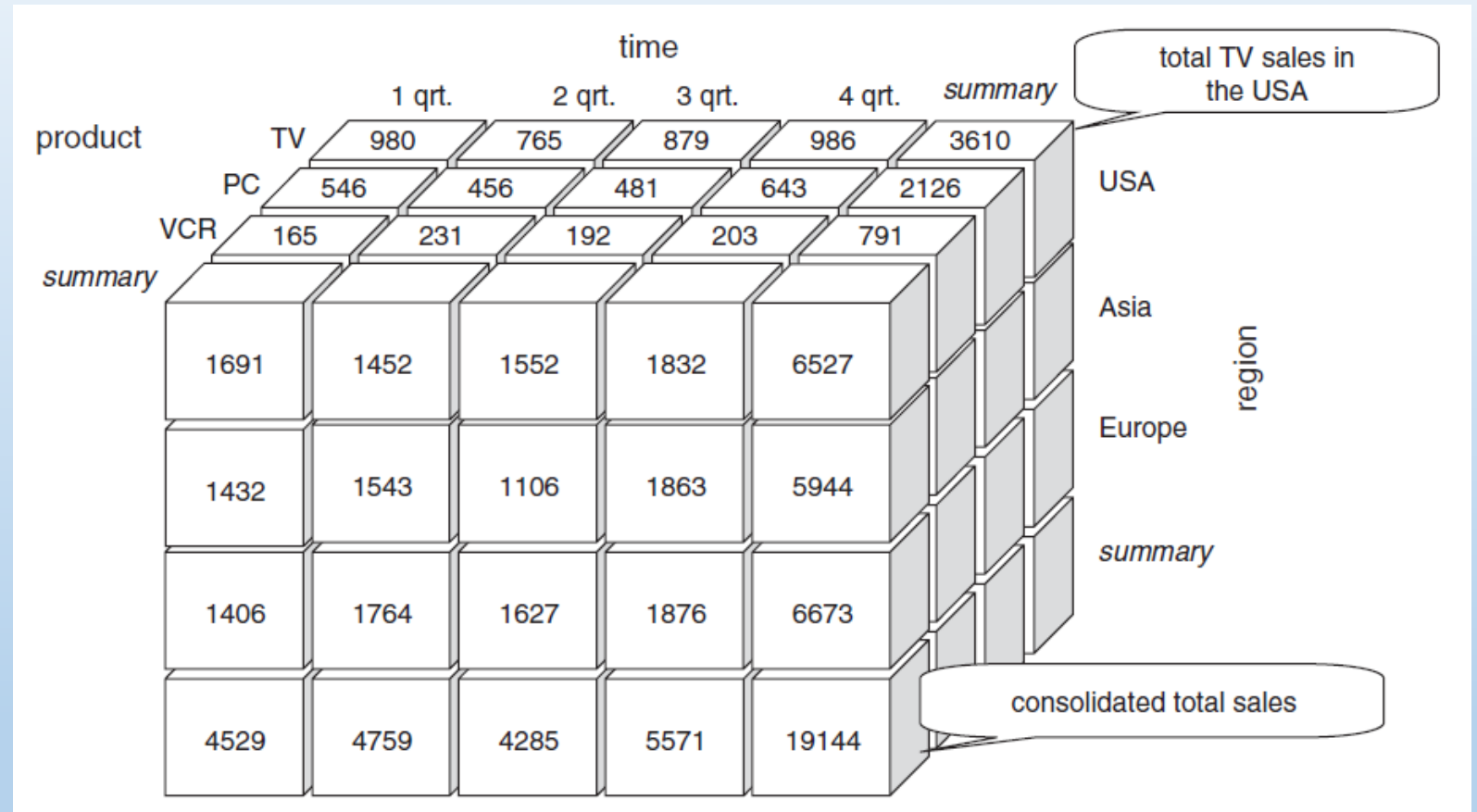
region = Asia			
time	product		
	TV	PC	DVD
Q1	789	456	187
Q2	654	732	157
Q3	623	354	129
Q4	756	876	231

Table 3.5 Two-dimensional view of sales data in Europe

region = Europe			
time	product		
	TV	PC	DVD
Q1	638	576	192
Q2	876	723	165
Q3	798	675	154
Q4	921	754	201

Esempio di cubo 3-dimensionale

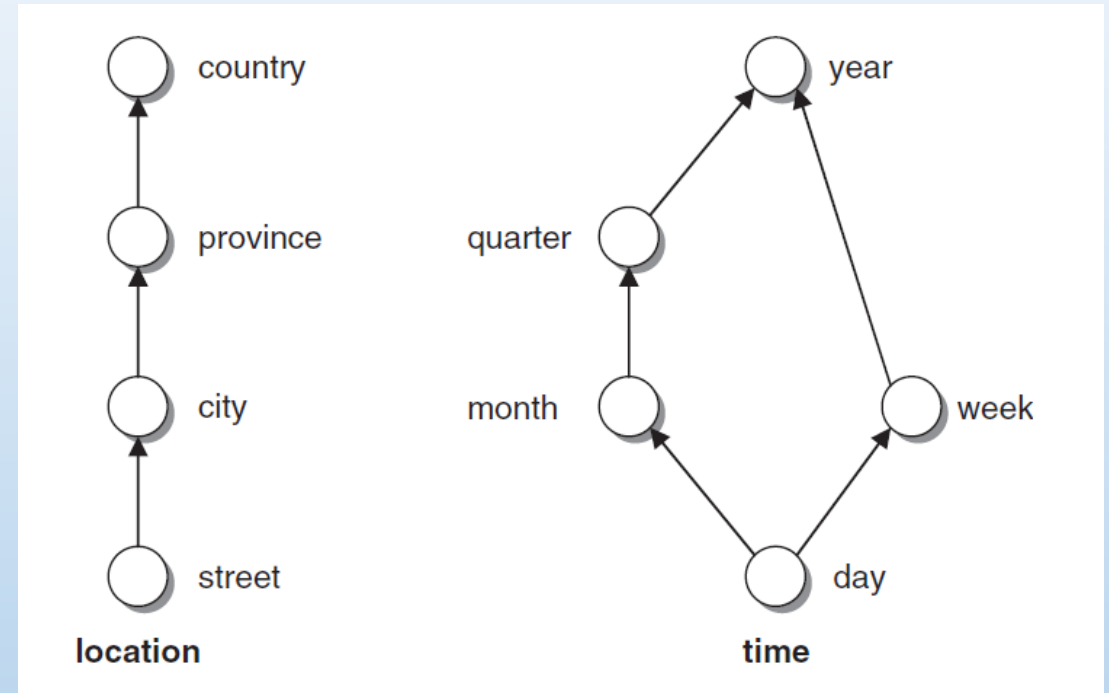
- Tabella dei fatti:
vendita
- Tabelle delle dimensioni:
 - Tempo
 - Prodotto
 - Regione



Gerarchie di concetti

Le analisi OLAP impiegano gerarchie di concetti per creare viste logiche lungo le dimensioni del Data Warehouse.

Una *gerarchia di concetti* definisce un insieme di corrispondenze da un insieme di concetti a livello inferiore verso concetti a livello superiore



Gerarchie di concetti e operazioni di visualizzazione

- **Roll-up (drill-up):** aggregazione nel cubo di dati
 - Mediante il passaggio a un livello superiore lungo una gerarchia su una dimensione. Ad esempio, nella dimensione «località» si passa da livello «comune» a livello «provincia» e si raggruppano i dati con somme condizionate (*group by*)
 - Mediante la riduzione di una dimensione. Ad esempio, si rimuove la dimensione tempo e si consolidano i dati tramite la somma su tutti i periodi temporali
- **Drill-down (roll-down):** operazione inversa del «roll-up», disaggregazione
 - Mediante il passaggio al livello inferiore lungo la gerarchia su una dimensione
 - Mediante l'aggiunta di una dimensione
- **Slice and dice**
 - Selezione il valore di un attributo lungo una dimensione
 - Recupero di un cubo in un sottospazio, selezionando più di una dimensione simultaneamente
- **Pivot (rotazione):** si ruotano gli assi scambiando tra loro alcune dimensioni per ottenere una vista diversa sul cubo dei dati

Calcolo dei cubi di dati

Per garantire tempi di risposta adeguati è utile progettare un Data Warehouse in modo che tutti i valori delle misure associate a tutti i possibili cuboidi siano pre-calcolati (**materializzazione totale**)

- In assenza di gerarchie, date n dimensioni $\rightarrow 2^n$ possibili cuboidi
- In presenza di gerarchie
 - L_i numero di livelli gerarchici associati alla dimensione $i \rightarrow$

$$T = \prod_{i=1}^n (L_i + 1).$$

- Se $n = 5$ e $L_i = 3 \rightarrow T \approx 10^3$

Materializzazione parziale

Compromesso tra:

- Esigenza di rapidità nell'accesso alle informazioni (→ materializzazione)
- Necessità di contenere l'occupazione di memoria

Vengono materializzate preventivamente soltanto i cuboidi per cui si stima un accesso più frequente

Per gli altri il calcolo viene svolto al momento delle interrogazioni