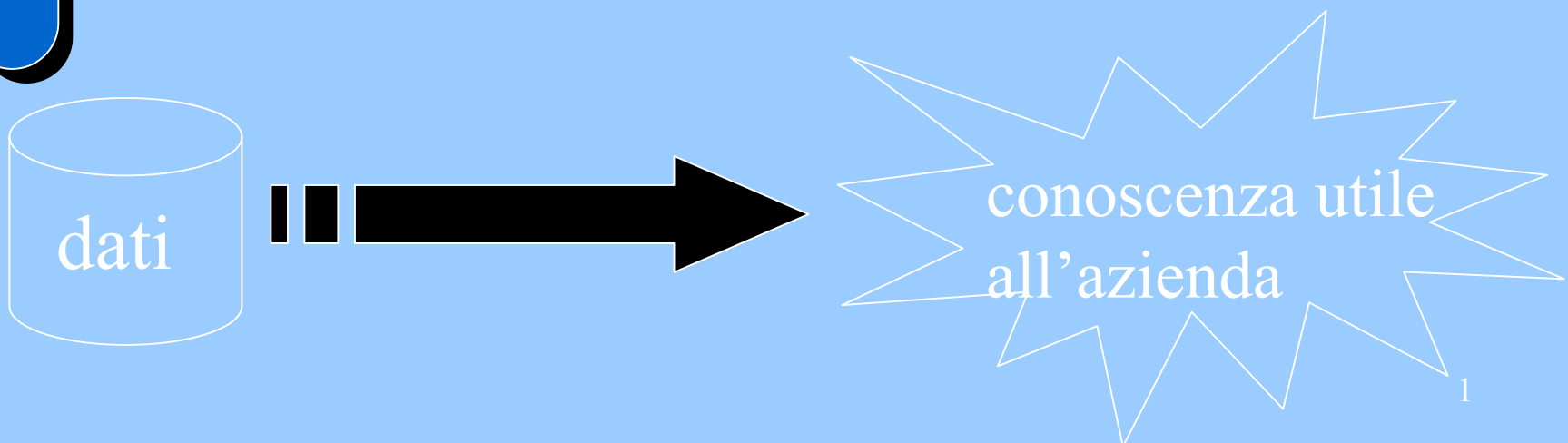
Data Warehousing (DW)

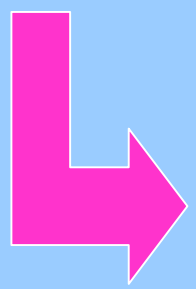
Data warehousing come processo
per memorizzare dati storici ed integrati,
eventualmente memorizzati in luoghi diversi,
da usare come supporto
al sistema di decisione aziendale



Esigenze che portano al DW

Uso di dati a fini decisionali:

- grandi quantità di dati
- più basi di dati, tra loro eterogenee
- i dati devono essere raggruppati, classificati, riassunti



Tecnologia ad hoc per OLAP (DBMS non sono sufficienti, studiati per OLTP)

Perché la tecnologia corrente non è sufficiente?

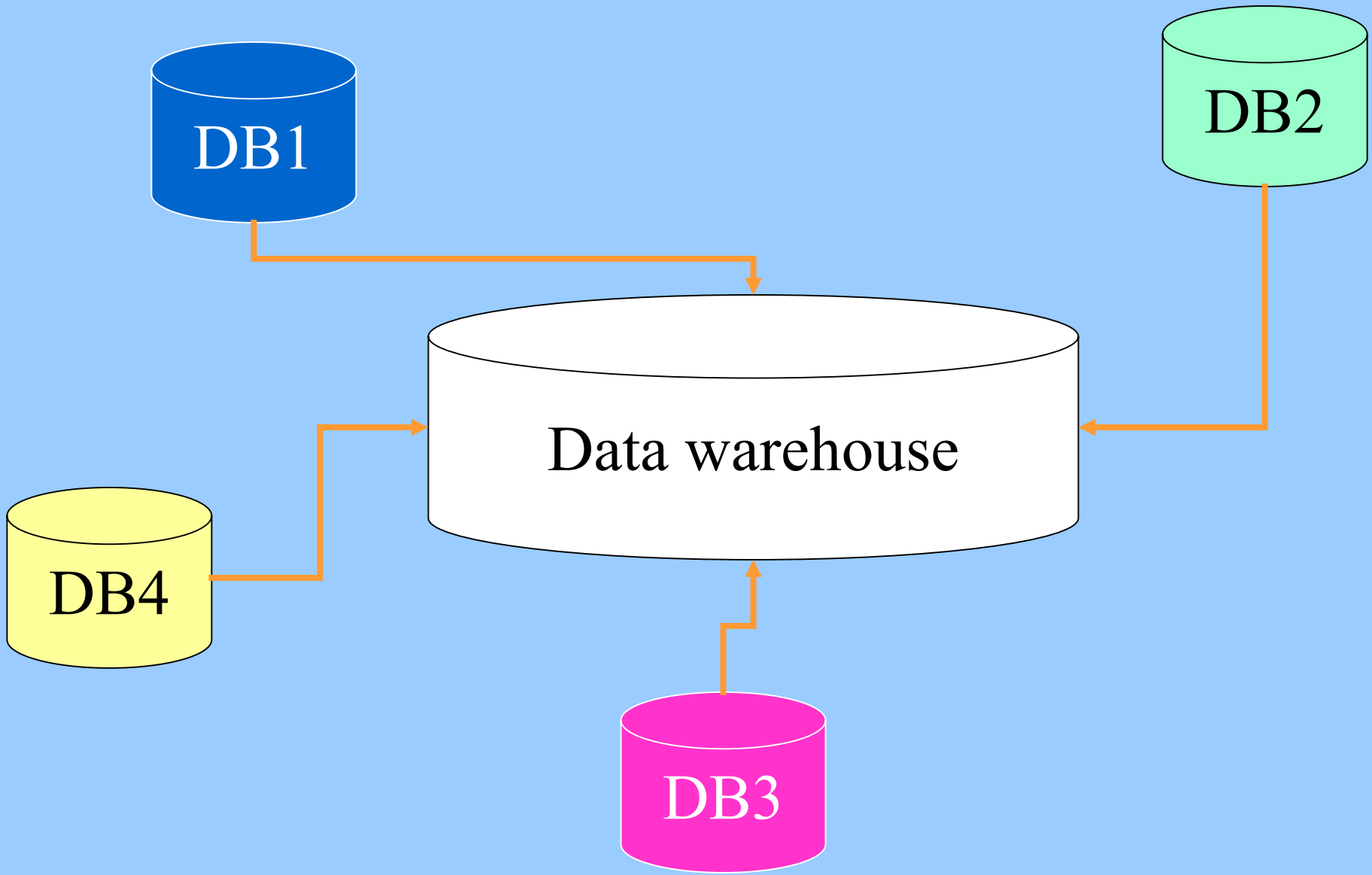
- sistemi eterogenei
- basse prestazioni
- DBMS non adeguati al supporto decisionale
- problemi di sicurezza

Soluzione

Nuova base di dati  data warehouse

Nuovo sistema di gestione dati  sistema di data warehousing

Il data warehouse



Il data warehouse

Collezione di dati

orientata ai soggetti
integrata
correlata alla variabile tempo
non-volatile
usata principalmente per il supporto alle decisioni

Il data warehouse

Orientata ai soggetti

Considera soggetto di interesse all'organizzazione e non in funzione dei processi organizzativi

Il data warehouse

Integrata

I dati arrivano da fonti diverse

diversi formati

integrazione 

dati memorizzati in formati consistenti

Esempio: sesso dell'individuo può essere rappresentato come "M" e "F" o come "0" e "1"

Il data warehouse

Correlata alla variabile tempo

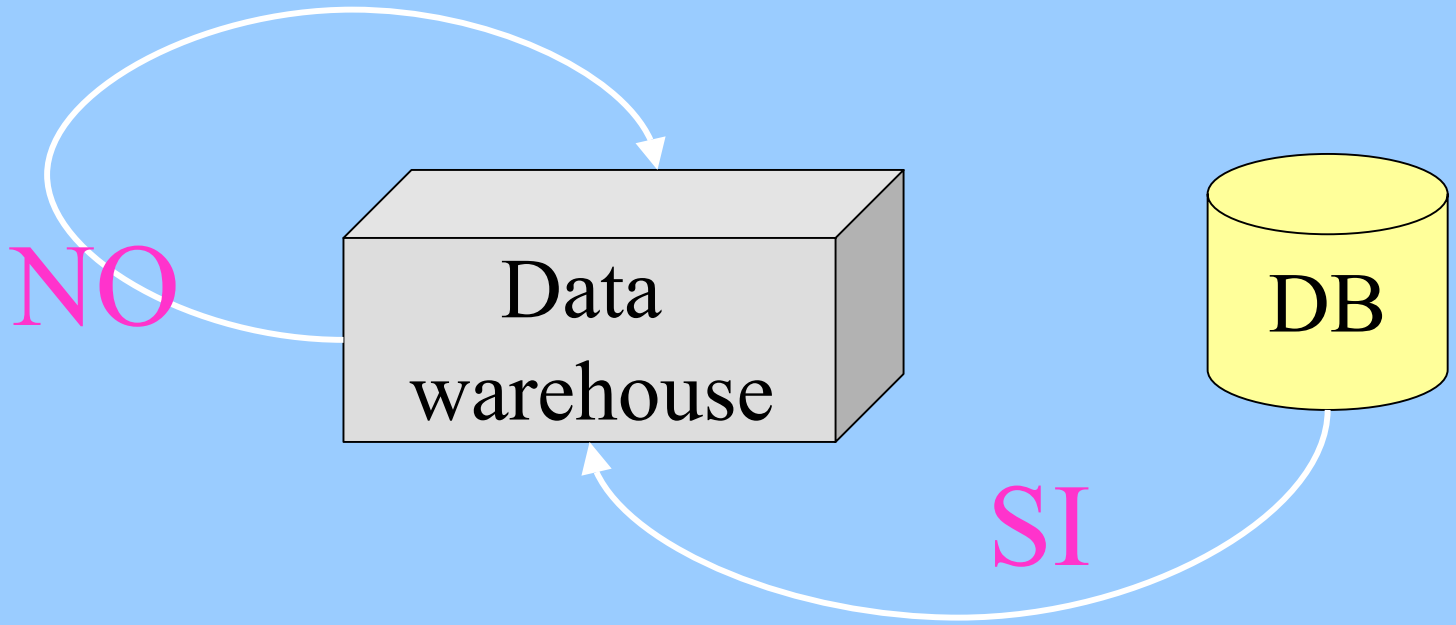
Presenza di dati storici per eseguire confronti, previsioni e per individuare tendenze

Esempio sistema bancario: dati relativi ai vecchi clienti per stabilire quanti clienti chiuderanno il conto nell'anno seguente

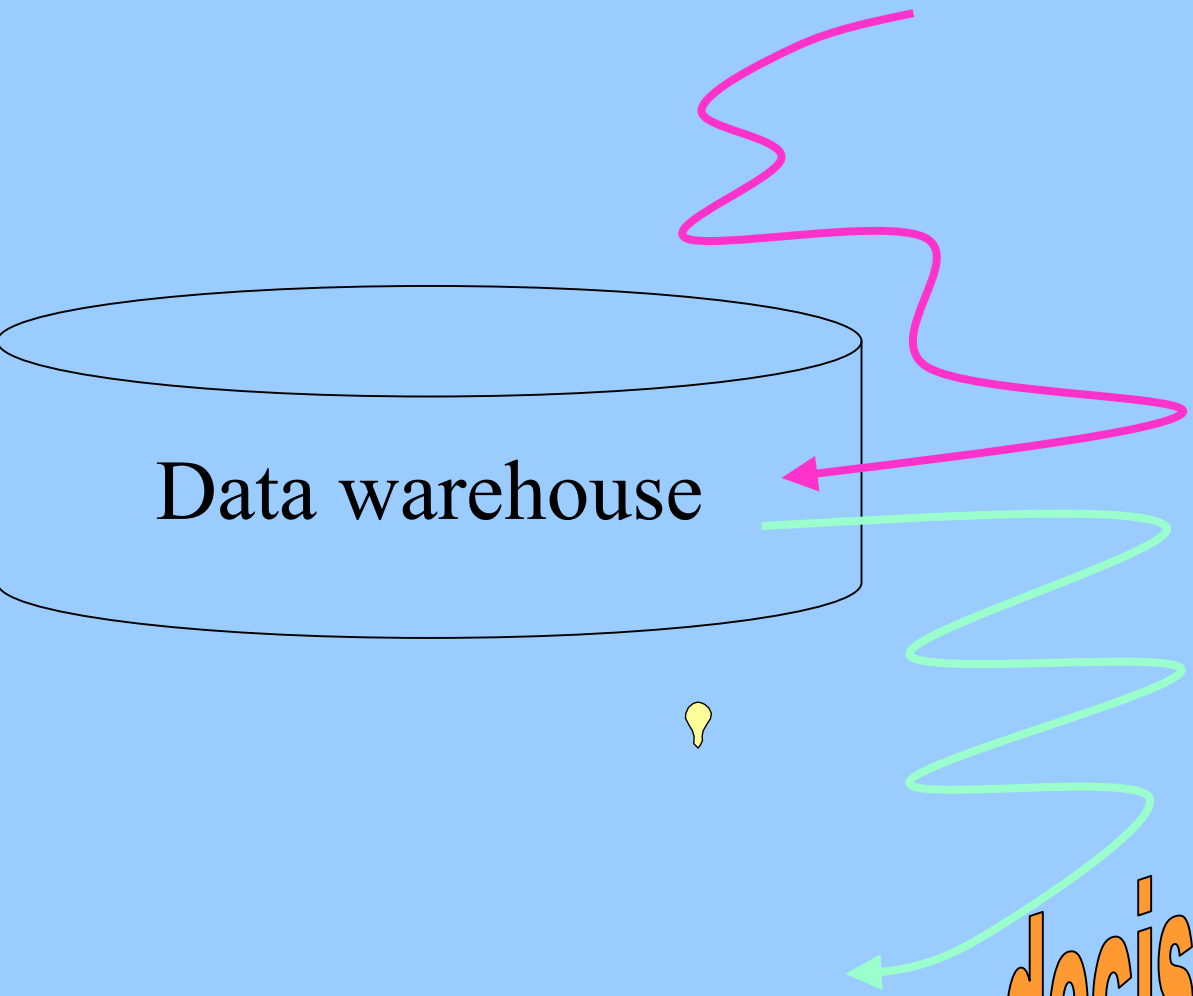
Il data warehouse

Non volatile

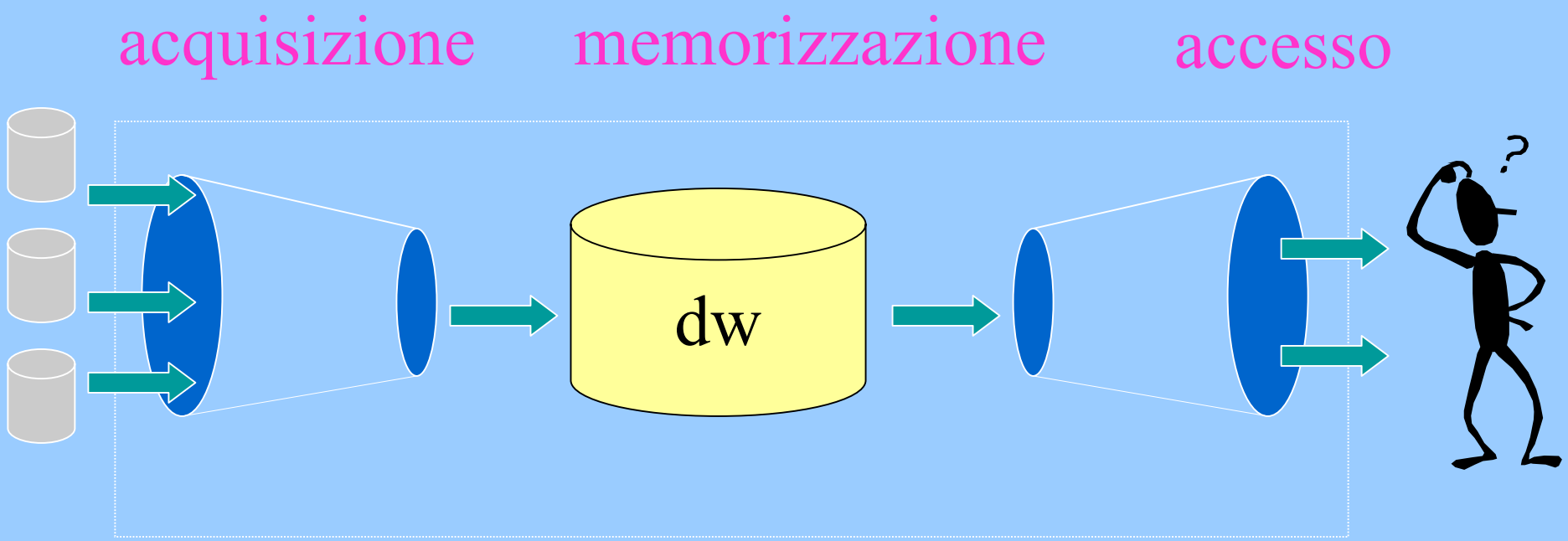
Una volta che le informazioni sono inserite nel data warehouse, non possono essere modificate ma solo ricaricate



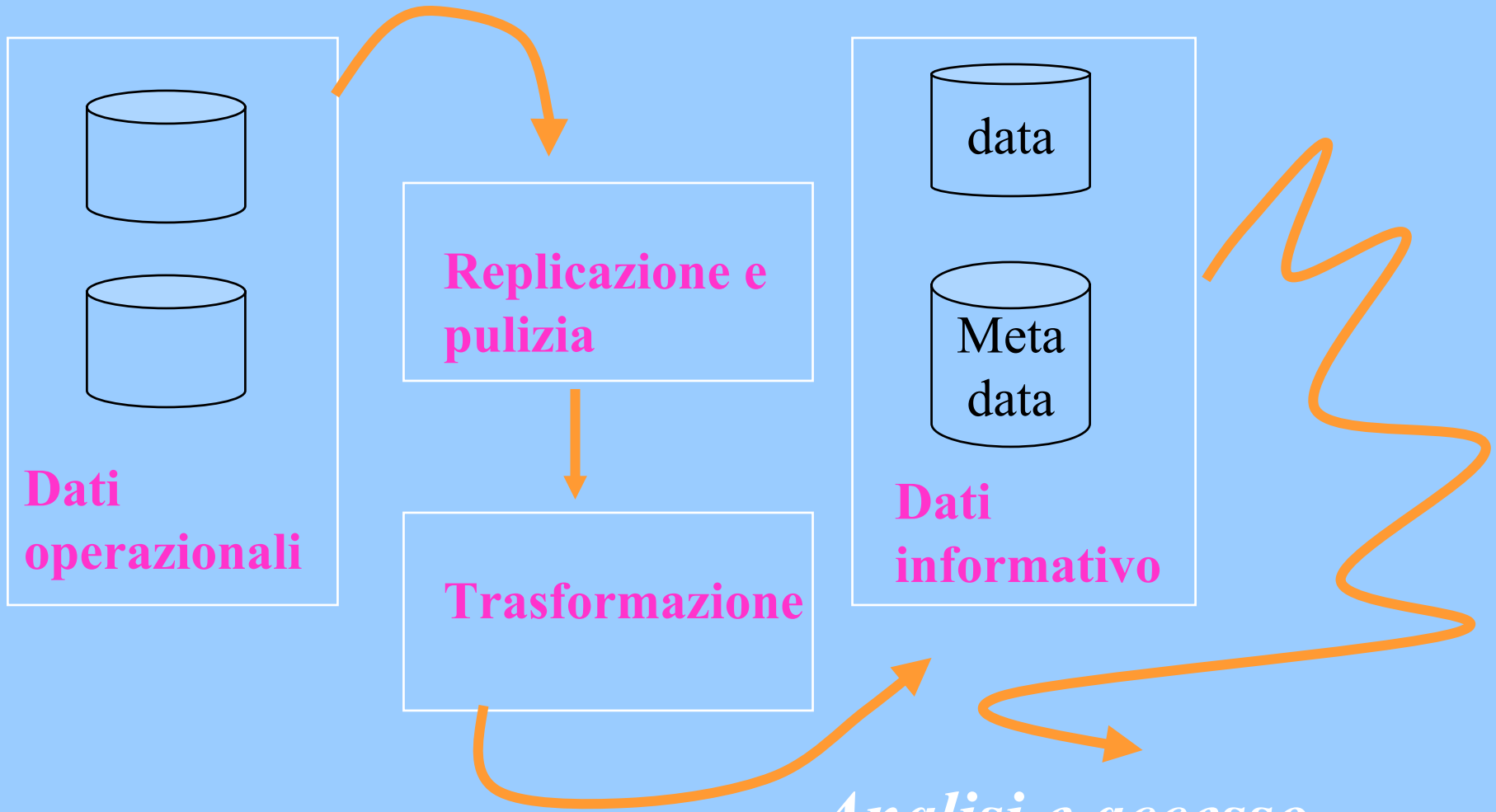
Gestione del data warehouse



Una prima architettura funzionale



Una seconda architettura funzionale



Schema di un DW

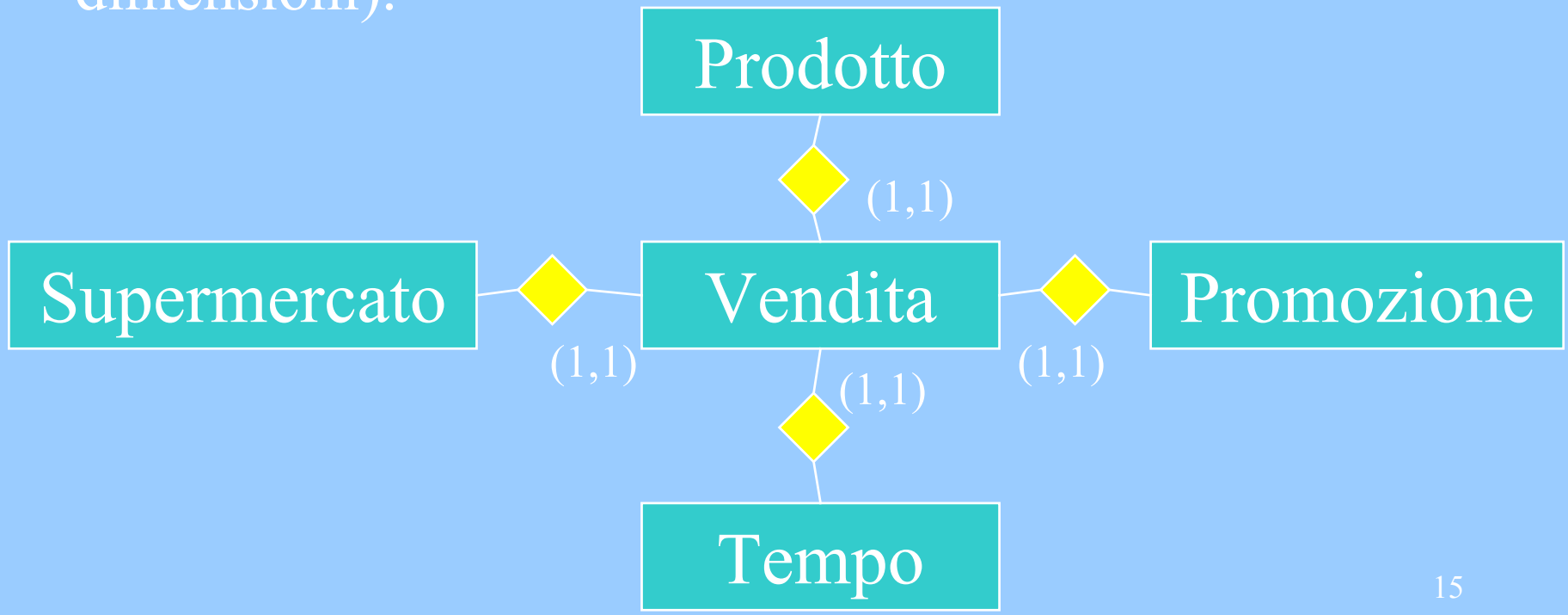
E’ strutturato in sottoparti separate dette DATA MART per le quali risulta chiaro l’obiettivo dell’analisi.

Ogni DATA MART ha uno schema concettuale che spesso si presenta con una struttura a STELLA.

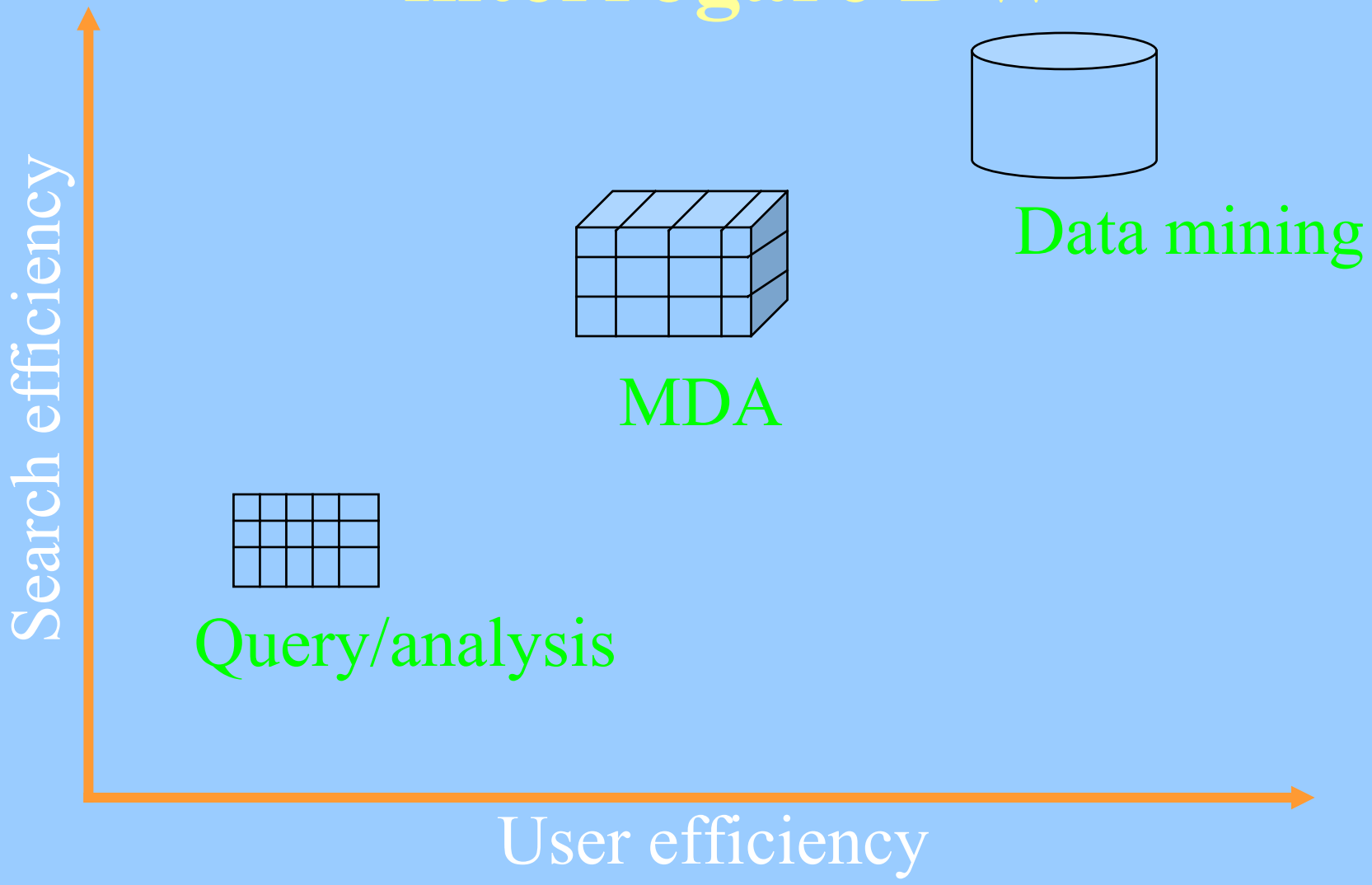
Schema di un DATA MART

Schema a *stella*:

- al centro l'entità che rappresenta i **fatti** (i cui attributi rappresentano le misure dei fatti),
- alle estremità le entità che rappresentano le **dimensioni** dell'analisi (i cui attributi rappresentano i livelli delle dimensioni).



Evoluzione degli strumenti per interrogare DW



Query e analisi

- Permettono di formulare interrogazioni senza scrivere un programma o conoscere SQL
- mettono a disposizione alcuni meccanismi di reporting: generazione di documenti (tabelle, grafici e altro) per presentare il risultato dell'analisi

Formulazione delle query via interfaccia grafica

Esempio di interfaccia

dim1 (D1.a)	dim2 (D2.b)	fatti (F.c)	fatti (F.d)
valori presenti	valori presenti		
selezione	selezione		
group by	group by	$OP1_{Aggr}$	$OP2_{Aggr.}$

Quey SQL equivalente all'interfaccia grafica

SELECT D1.a, D2.b, $OP1_{Aggr}$(F.c), $OP2_{Aggr}$(F.d)
FROM Fatti AS F, Dim1 AS D1, Dim2 AS D2
WHERE <predicati di join(F,D1)> AND
<predicati di join(F,D2)> AND <selezioni>
GROUP BY D1.a, D2.b
ORDER BY D1.a, D2.b

Strumenti
Multi Dimensional Analysis

Permettono di creare viste multidimensionali su ordinarie tabelle bidimensionali (si possono anche usare particolari DBMS multidimensionali)

MDA tools

Esempio

vendite

Dati di dettaglio (fatti)

- tempo
- cliente
- prezzo
- magazzino
- prodotto

Dimensioni (parametri analizzati)

MDA tool

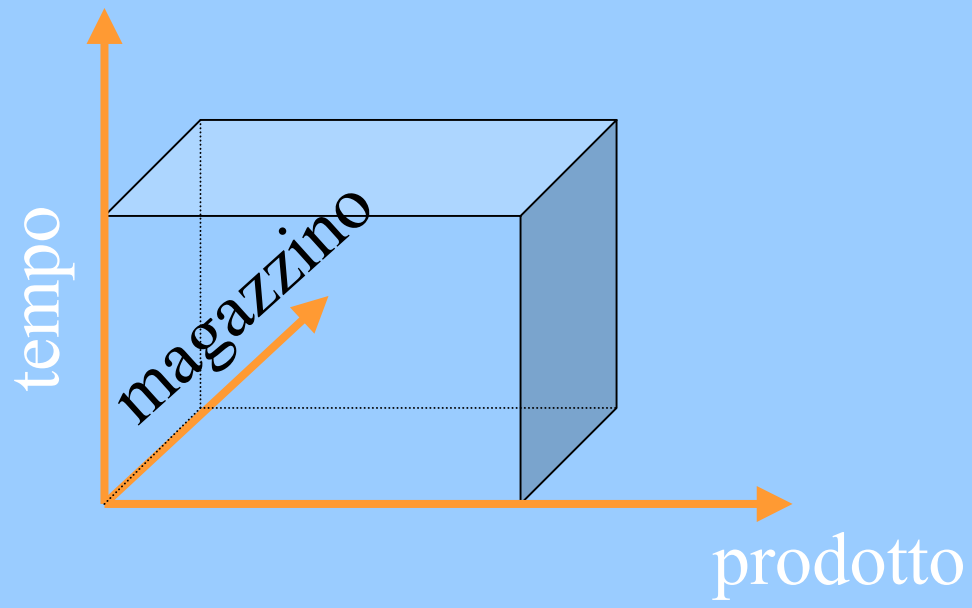
Esempio di query

Quali sono i dati sulle vendite, classificati rispetto a:

- prodotto
- magazzino
- mese?

MDA tools

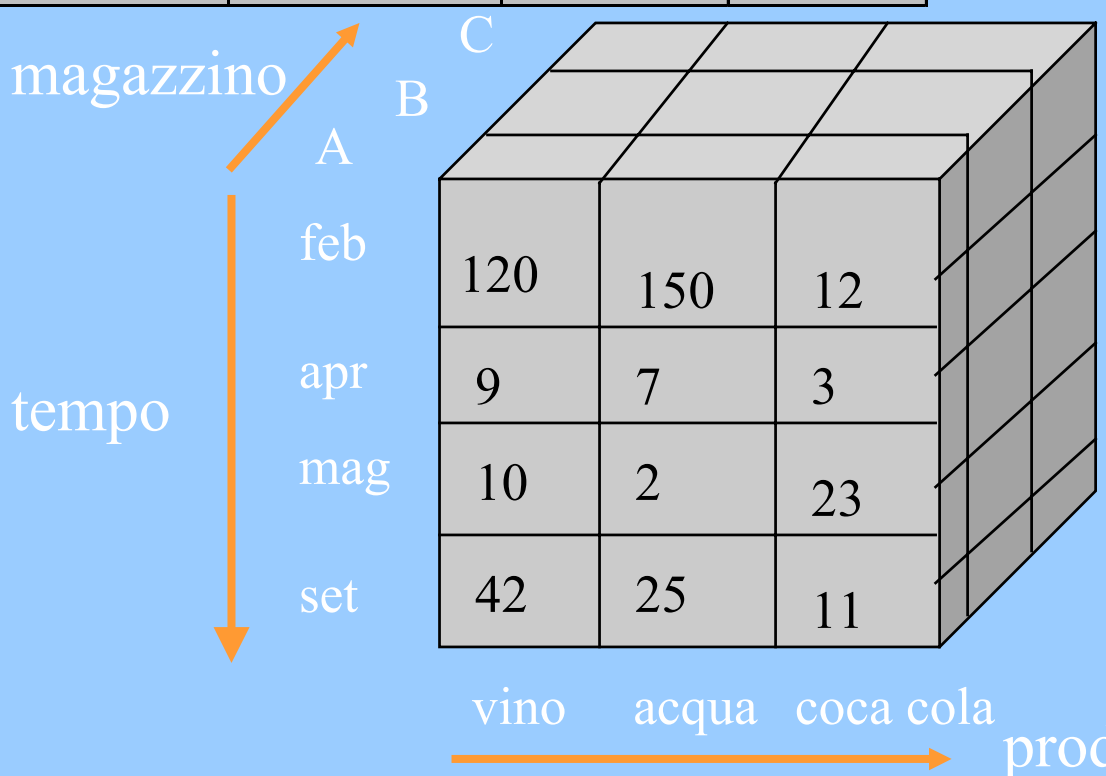
Analisi rispetto a variabili multiple



- In genere: efficienti fino a 10 variabili

MDA tools

vendite	prodotto	mese	magazzino	quantità	ricavi
1	vino	febbraio	A	120	150000
2	acqua	febbraio	B	34	34000
3	coca cola	aprile	A	245	380000
4	acqua	maggio	A	47	99000
5	acqua	settembre	C	55	71200
…	…	…	…	…	…



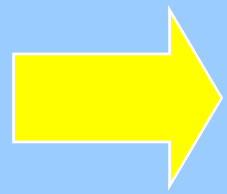
MDA tools

Le strutture dati OLAP (MDA) si possono pensare come un "CUBO DI RUBIK" di dati, che gli utenti possono modificare in modi diversi per analizzare diversi scenari del tipo "what-if"

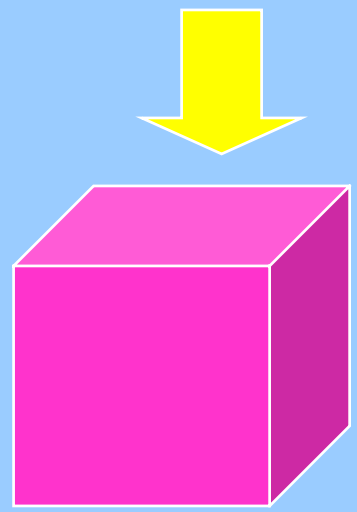
Data Cube nei sistemi

Clausola: WITH CUBE su una interrogazione SQL con il GROUP BY.

Mese	Magazzino	Prodotto	Quantità
Gen	A	Vino	10
Feb	A	Vino	20
Mar	A	Acqua	30
Gen	B	Acqua	40
Feb	B	Vino	50
Mar	B	Vino	60
Gen	C	Vino	70
Feb	C	Acqua	80
Mar	C	Acqua	90



SELECT Mese, Magazzino, Prodotto, Sum(Quantità) FROM T GROUP BY CUBE (Mese, Prodotto)



Data Cube nei sistemi

Tabella risultato della query GROUP BY CUBE del precedente lucido.

Mese	Prodotto	sum(Qta)
Gen	Vino	80
Feb	Vino	70
Mar	Vino	60
Gen	Acqua	40
Feb	Acqua	80
Mar	Acqua	120
Gen	ALL	120
Feb	ALL	150
Mar	ALL	180
ALL	Vino	210
ALL	Acqua	240
ALL	ALL	450

MDA tools

Due tipi di operazioni:

1. *Slice and dice*: si parte da un particolare aspetto e si naviga nella struttura multidimensionale, selezionando celle della struttura (cubo)

2. *Drill / Roll*

   drill-down: da informazioni più aggregate a informazioni meno aggregate (aggiunta di una dimensione o di un livello meno generale di una dimensione

   roll-up: da informazioni meno aggregate a informazioni più aggregate (togliendo una dimensione o passando a un livello più generale di una dimensione)

Data Mining

Fasi del processo di data mining

- comprensione del dominio;
- preparazione del set di dati;
- scoperta dei pattern (regole di associazione)
- valutazione dei pattern
- utilizzo dei risultati

Data Mining

Regole di associazione: si applicano ad una tabella relazionale partizionata tramite una clausola di raggruppamento (esempio della basket analysis: transazioni di acquisto).

Nella "basket analysis" una regola descrive situazioni in cui la presenza di una merce in una transazione è associata alla presenza di un'altra merce con elevata probabilità.

Data Mining

Regole di associazione

premessa $\rightarrow$ conseguenza

oppure

“corpo” $\rightarrow$ “testa”

Esempio:

giacca $\rightarrow$ camicia

Data Mining

Data un'osservazione: insieme di tuple

si dice SUPPORTO di una regola:

la probabilità che siano presenti nell'osservazione sia la testa che il corpo della regola;

si dice CONFIDENZA di una regola:

la probabilità che in una osservazione dove sia presente il corpo sia presente anche la testa.

Data Mining

CodTrans	Data	Oggetto	Quantità	Prezzo
1	17/12/03	Pantaloni-sci	1	140
1	17/12/03	Scarponi	1	180
2	18/12/03	Maglietta	2	25
2	18/12/03	Giacca	2	300
2	18/12/03	Stivali	1	70
3	18/12/03	Giacca	1	300
4	20/12/03	Giacca	1	320
4	20/12/03	maglietta	2	27

Trovare SUPPORTO e CONFIDENZA delle principali regole associative presenti in questa tabella dei fatti.