Il passaggio tra il primo ed il secondo livello: gli strumenti di *extraction*, *tranformation* and *loading* (*ETL tools*)

LA RICONCILIAZIONE DEI DATI

1. Estrazione: vengono estratti dai sistemi alimentanti i dati necessari per la produzione finale delle informazioni manageriali di interesse.
2. Pulitura: si determina la qualità dei dati che verranno caricati nel DW correggendo gli errori e le inconsistenze presenti nei dati elementari (dati duplicati, dati mancanti, valori errati o nulli, inconsistenza fra valore e descrizione del campo, inconsistenza tra valori logicamente associati).
3. Trasformazione: si convertono i dati dal formato di origine a quello del DW (da numero intero a numero decimale, sistemi di misura, maiuscolo minuscolo) e si effettua l'operazione di matching che stabilisce corrispondenze tra campi equivalenti in sorgenti diverse.
4. Caricamento: si effettua il popolamento del DW che può avvenire per *sostituzione* oppure per *aggiornamento*.

Esempio di pulitura e trasformazione di un dato anagrafico

Carlo Bianchi
P.zza Grande 12
50126 Bologna (I)

Normalizzazione

Nome: Carlo
Cognome: Bianchi
Indirizzo: P.zza Grande 12
CAP: 50126
Città: Bologna
Nazione: I

Standardizzazione

Nome: Carlo
Cognome: Bianchi
Indirizzo: Piazza Grande 12
CAP: 50126
Città: Bologna
Nazione: Italia

Correzione

Nome: Carlo
Cognome: Bianchi
Indirizzo: Piazza Grande 12
CAP: 40126
Città: Bologna
Nazione: Italia

Il secondo livello dell'architettura in esame è costituito dal *datawarehouse* che rappresenta:

l'anello di collegamento tra i dati, le applicazioni ed i sistemi informativi di tipo operativo e transazionale, ed i sistemi informativi manageriali di supporto alle attività di controllo e di decisione.

Data Warehouse: una collezione di dati in supporto al processo decisionale del management…

- Orientata al soggetto
- Integrata
- Variante nel tempo
- Non volatile

Bill Inmon 1996

ORIENTATA AL SOGGETTO – Subject oriented:

- Il DW è orientato a temi specifici dell'azienda (clienti, prodotti, ecc.) piuttosto che alle applicazioni o funzioni (inventario, fatturazione, gestione ordini, ecc.). Esso include tutti i dati che possono essere utilizzati nel processo di controllo e di decisione, raggruppandoli per aree, fatti o temi di interesse (specifico fenomeno o fatto aziendale rilevante).
- Nel DW i dati vengono archiviati in modo che possano essere facilmente letti o elaborati dagli utenti cioè in modo da favorire la produzione di informazioni; viceversa i database operazionali sono organizzati intorno alle differenti applicazioni del dominio aziendale (sono, cioè orientati a chi li genera).

INTEGRATA:

- L’integrazione è un requisito fondamentale del DW in quanto in esso confluiscono dati provenienti da più fonti.
- Il DW è consistente rispetto ad un modello concettuale dei dati, al glossario aziendale e rispetto alle unità di misura ed alle strutture di *decodifica condivise a livello aziendale* (dati documentati).
- L’obiettivo dell’integrazione può essere raggiunto percorrendo differenti strade: mediante l’utilizzo di metodi di codifica uniformi, mediante il perseguimento di una omogeneità sistematica di tutte le variabili, mediante l’utilizzo delle stesse unità di misura, ecc.
- I dati archiviati in un DW sono certificati, omogenei e consistenti.

VARIANTE NEL TEMPO – Time variant:

- I dati archiviati all'interno di un DW hanno un orizzonte temporale molto più esteso rispetto a quelli archiviati in un sistema operazionali (passato, presente, futuro).
- Nel DW sono contenuti da "fotografie" (*snapshots*) periodiche dei fatti correnti o storici.
- I dati contenuti in un DW sono aggiornati fino ad una certa data, che nella maggior parte dei casi, è antecedente a quella in cui l'utente interroga il sistema.

NON VOLATILE:

- Il dato viene caricato periodicamente ed acceduto fuori linea cioè non può essere modificato dall'utente (l'accesso è in sola lettura).
- Si evitano le possibili anomalie dovute ad aggiornamenti e tanto meno si ricorre a strumenti complessi per gestire l'integrità referenziale o per bloccare record a cui possono accedere altri utenti in fase di aggiornamento.

Sistemi informativi direzionali – il ruolo dei datawarehouse

Tre ruoli

1. **Integratore**: il sistema funge da *hub* dei flussi di dati, consentendo di disporre a livello centrale di un patrimonio complesso di dati "semilavorati" (non elementari), integrati, omogeneizzati, certificati e documentati

2. **Disaccoppiatore**: opera una separazione fra l'ambiente operazionale e quello decisionale. Il DW consente ai SID di funzionare a velocità e con prestazioni differenziate, con dati storici e previsionali, aggiornati con frequenza diversa, più aggregati e meglio documentati all'utente

3. **Consolidatore**: nelle strutture di gruppo, nelle organizzazioni multidivisionali, multibusiness o multigeografiche, può permettere a livello centrale di omogeneizzare ed analizzare correttamente fenomeni gestionali complessi generati da politiche e procedure gestionali diversificate

Caratteristiche principali di un Datawarehouse rispetto ai Database operativi

	Database Operativi	Datawarehouse
Utenti	**tendenzialmente tutto il personale operativo aziendale**	**manager, staff**
Carico di lavoro	**transazioni predefinite**	**interrogazioni per analisi ad hoc**
Accesso	**in lettura e scrittura**	**in lettura (in scrittura nel caso della simulazione decisionale)**
Scopo	**dipende dall'applicazioni**	**supporto alle decisioni**
Dati	**elementari**	**sintetici**
Integrazione dei dati	**per applicazione**	**per soggetto di indagine**
Qualità	**in termini di integrità**	**in termini di consistenza**
Copertura temporale	**solo dati correnti**	**dati correnti, storici e previsionali**
Aggiornamento	**continuo**	**periodico**
Modello	**normalizzato**	**denormalizzato, multidimensionale**

segue

segue

	Database operativi	**Datawarehouse**
Dominio Applicativo	i sistemi transazionali sono definiti per un limitato dominio applicativo che si riferisce a una specifica applicazione	i progetti di DW forniscono un'infrastruttura di appoggio ai sistemi di supporto alle decisioni con caratteristiche di scalabilità di ampliamento e flessibilità
Sviluppo	sistemi OLTP sviluppati seguendo i requisiti del sistema esplicitati dagli utenti	i criteri di sviluppo rispondono a principi evolutivi e iterativi
Sponsorship	i sistemi transazionali tendono ad essere sponsorizzati seguendo un processo che consente di individuare il responsabile che individua a sua volta anche le gerarchie organizzative	un progetto di DW richiede una forte sponsorizzazione a causa dell'ampiezza organizzativa dello stesso

Sistemi informativi direzionali – i metadati

I dati contenuti nel *datawarehouse* devono essere "semanticamente corretti" (non devono esistere dati diversi identificati con lo stesso nome), rilevati e calcolati con criteri omogenei nel tempo (per poter confrontare i dati passati con i dati correnti) e nello spazio (uguali nelle diverse funzioni, divisioni, unità operative, magazzini o filiali dell'azienda).

Per verificare che ciò sia stato realizzato deve essere creato un *catalogo dei dati* che riassuma il significato preciso di ogni dato, le sue modalità di calcolo, di proprietà del dato e di omogeneità (unità di misura).

Il catalogo contiene, quindi, dati che descrivono altri dati, i cosiddetti *metadati*.

Classificazione dei metadati in ambito SID

- **Business**: significato e modalità di calcolo dei dati (creazione di un linguaggio direzionale comune e condiviso), viste di dati disponibili (combinazioni di dimensioni di analisi disponibili e relative ad una certa misura quantitativa), la provenienza dei dati (fonti interne ed esterne), la proprietà dei dati, i processi usati per l'estrazione o le procedure che usano i dati (*report*, fogli elettronici, ecc.).
- **Tecnici**: descrivono l'accesso ai dati di input, il trasporto e la trasformazioni di questi dai sistemi di origine all'ambiente DW, la descrizione del modello dei dati e delle aggregazioni presenti, le corrispondenze fra le fonti dei dati operativi e le tabelle di output del DW, la mappatura dei dati operativi di input, le aggregazioni ed i passaggi tra i livelli del DW, la frequenza di aggiornamento dei dati, la sicurezza, ecc.

Idealmente l'utente dovrebbe essere in grado di accedere ed operare sui dati senza conoscere dove essi risiedano, in quale forma siano stati memorizzati e quali strumenti software provvedano al loro trattamento fino alla schermata con cui l'utente interagisce.

In un ambiente di analisi e di supporto alle decisioni manageriali, le situazioni aziendali si modificano ogni giorno ed i metadati funzionano da indispensabile supporto al processo informativo, perché l'utente non è quasi mai proprietario e produttore dei dati e spazia le sue indagini su una base dati spesso molto vasta.

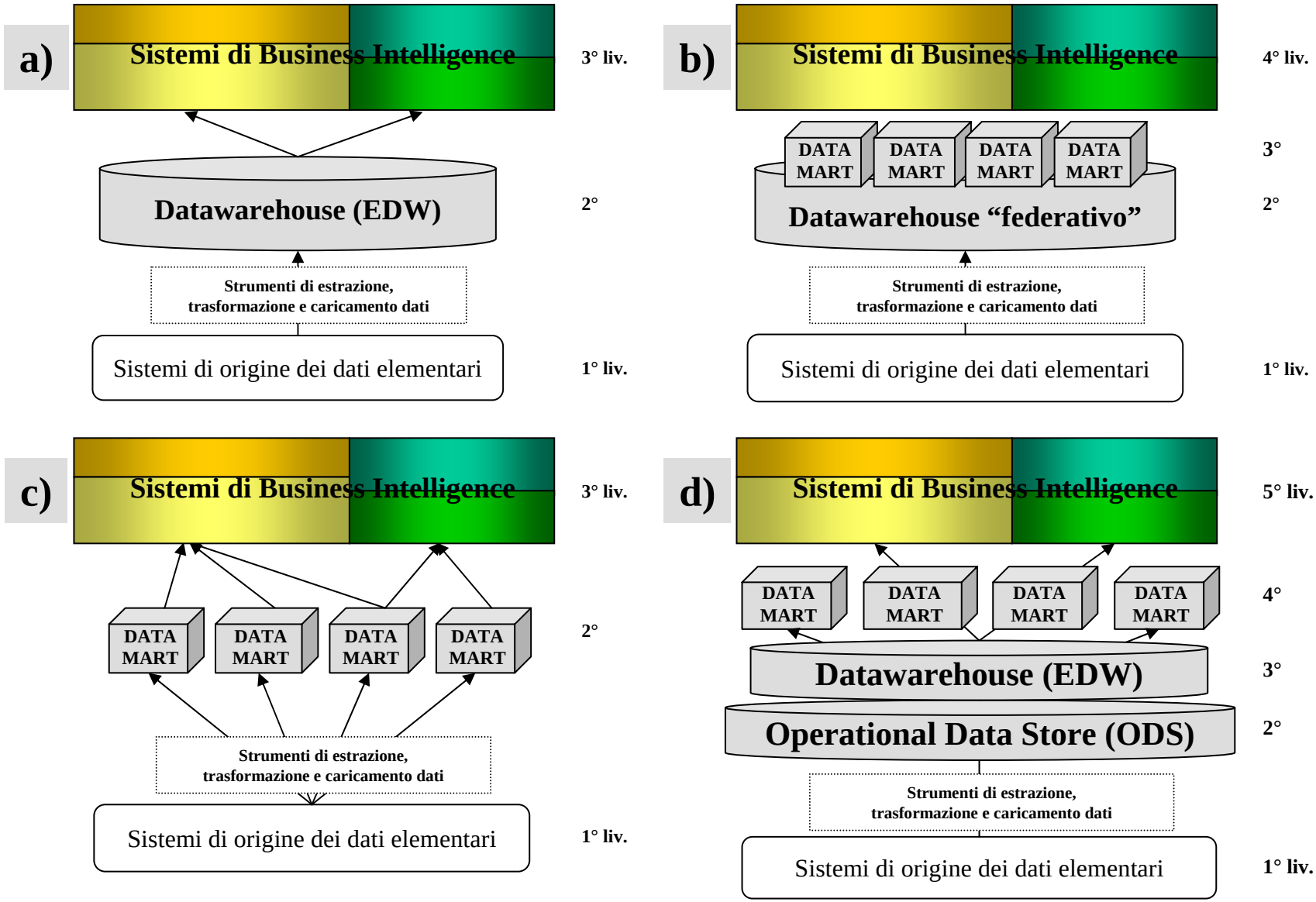
Sistemi informativi direzionali – i *datamart*

Per poter gestire e far convivere al meglio il numero di fenomeni descritti nel DW e la necessità di avere a disposizione dati di dettaglio differente è possibile utilizzare architetture di dati su più livelli (e ricorrere quindi a tecnologie differenti).

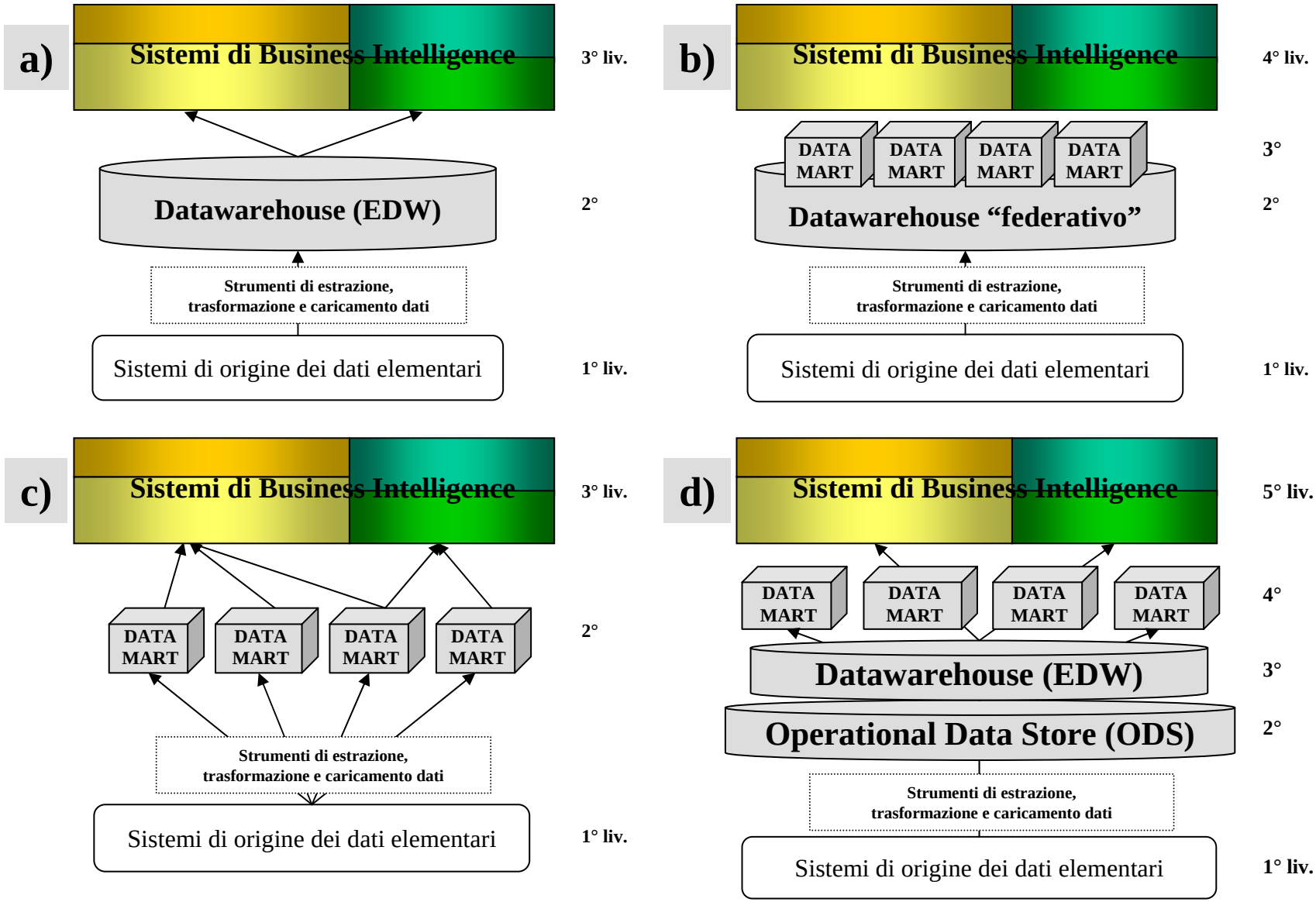
L'architettura può essere, quindi, costruita in modo da non accedere per le analisi direttamente al DW ma ai cosiddetti *datamart* che sono basi di dati con le caratteristiche generali del *datawarehouse*, ma la cui modellizzazione e l'archiviazione fisica sono rivolte ad un sottoinsieme tematico o ad un'aggregazione dei dati presenti nel DW aziendale, cioè ai dati rilevanti per una particolare area di business o funzione aziendale (vendite, ordini, scorte, ecc.).

I *datamart* permettono di ottenere prestazioni migliori essendo dimensionalmente inferiori al DW.

Architetture di DW



Architetture di DW



Sistemi informativi direzionali – l'architettura di tipo b)

Il DW viene diviso su due livelli:

1. DW realizzato con database relazionali (modello logico per la memorizzazione dei dati sotto forma di tabelle a due dimensioni) che contiene il massimo livello di dettaglio dei dati necessari.
2. Serie di *datamart* realizzati con database multidimensionali con dati di sintesi, informazioni ed indicatori precalcolati lungo varie dimensioni di analisi.

Il modello multidimensionale

Negli ultimi anni le basi di dati multidimensionali hanno suscitato vasto interesse di ricerca e di mercato essendo alla base di varie applicazioni per il supporto alle decisioni. Il motivo per cui il modello multidimensionale viene adottato nella rappresentazione dei dati nei SID è essenzialmente legato alla sua semplicità ed intuitività anche per utenti non esperti di informatica che sono, però, abituati all'uso di applicazioni di tipo foglio elettronico come strumento di produttività individuale.

Che incassi sono stati registrati l'anno passato per ciascuna regione e ciascuna categoria di prodotto?
Che correlazione esiste tra l'andamento dei titoli azionari dei produttori di PC ed i profitti trimestrali negli ultimi 5 anni?
Quali sono gli ordini che massimizzano gli incassi?
Quale di due nuove terapie risulterà in una diminuzione della durata media di un ricovero?

Il modello multidimensionale

Si basa sui concetti di:

Fatto: oggetto di interesse per l'azienda (per esempio vendite, ordini, spedizioni, ricoveri, interventi chirurgici)

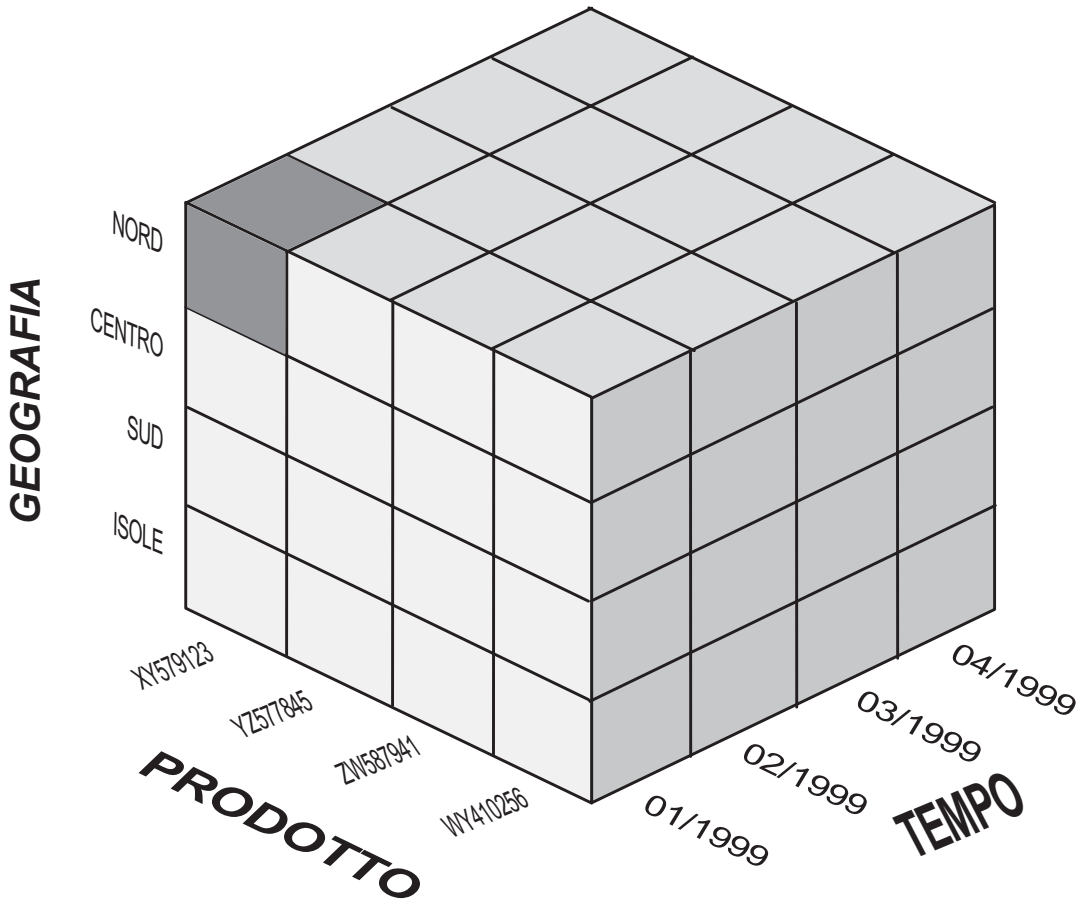
Eventi: occorrenze di un fatto (ogni singola vendita o spedizione effettuata)

Dimensione: prospettiva di analisi degli eventi che definisce lo spazio *n*-dimensionale i cui assi (le dimensioni appunto) rappresentano le diverse prospettive per la loro identificazione (le vendite in una catena di negozi possono essere rappresentate in uno spazio tridimensionale le cui dimensioni sono i prodotti, le date ed i negozi)

Misura: variabile quantitativa che descrive quantitativamente gli eventi (l'incasso di una vendita, la quantità spedita, il costo di un ricovero, la durata di un intervento chirurgico)

E' proprio il concetto di dimensione che ha dato origine alla diffusissima metafora del *cubo* per la rappresentazione dei dati multidimensionali. Gli eventi corrispondono alle celle di un cubo i cui spigoli rappresentano le dimensioni di analisi (se le dimensioni sono più di tre si parlerà di *ipercubo*).

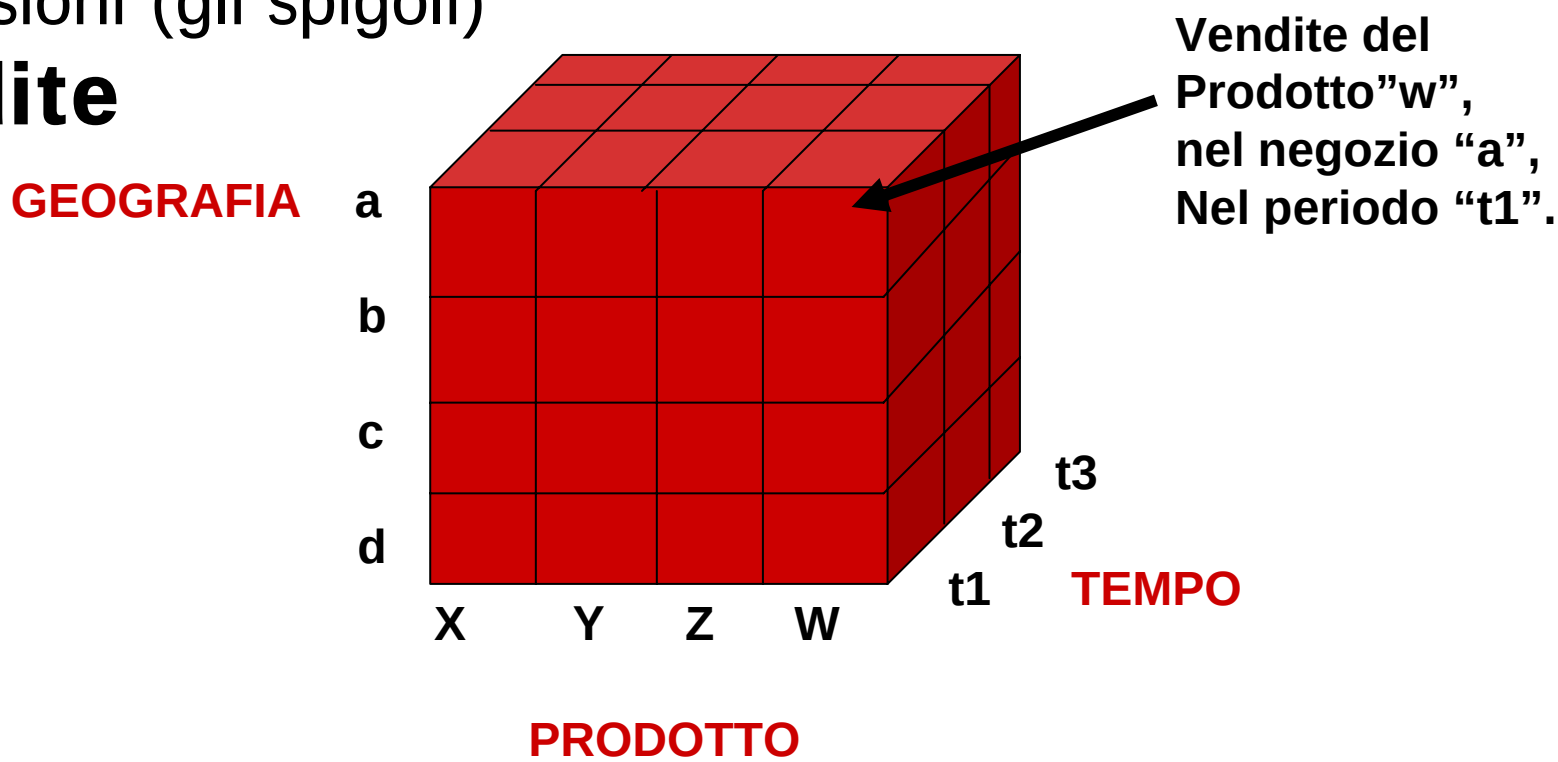
Modello Multidimensionale, cubi e ipercubi



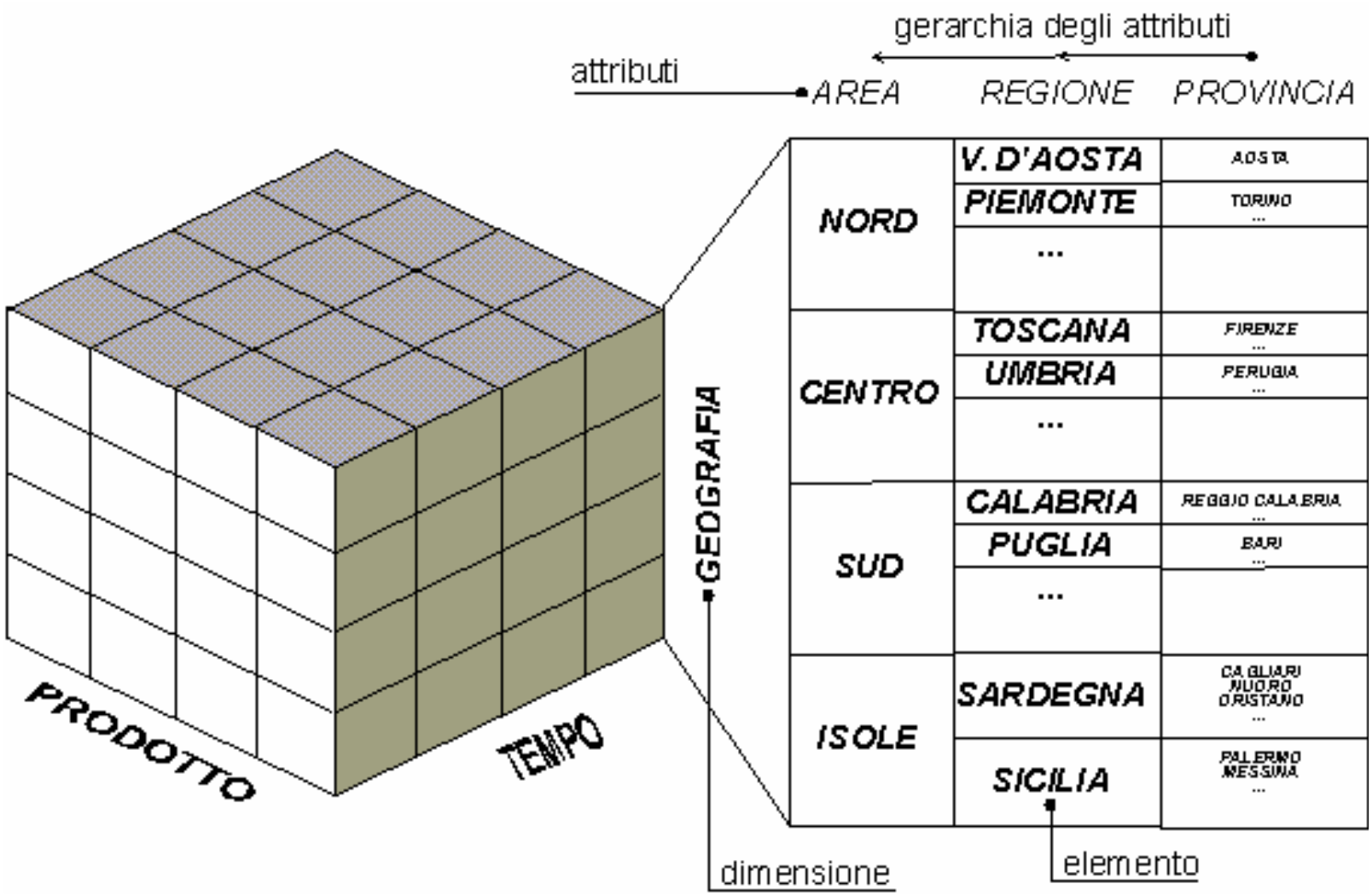
il Dimensional Model

- Fatti (nelle celle)
- Dimensioni (gli spigoli)

Vendite



Attributi della dimensione geografica



Gerarchie nelle Dimensioni

Rappresentazione per livelli delle modalità di una determinata dimensione, raggruppati in classi gerarchiche

- **Granularità**
  - Massimo dettaglio dei dati lungo una determinata dimensione
- **Additività su una dimensione**
  - Misura additiva
  - Misura non additiva
  - Misura non aggregabile

I singoli prodotti di un'azienda possono essere raggruppati in sottogruppi di articoli e questi a loro volta in gruppi di articoli, famiglie, reparti e settori merceologici

Sistemi informativi direzionali – l'architettura

La scelta di come articolare (numero di livelli) l'architettura dati dipende:

- dagli obiettivi che si intendono raggiungere (per esempio, tipologia di informazioni da produrre, volumi dei dati da trattare, livelli di analisi richiesti, numero e tipologia degli utenti, prestazioni desiderate);
- dall'assetto dei sistemi transazionali del SIA (per esempio, complessità ed eterogeneità);
- dal ruolo assunto all'interno dell'azienda dai sistemi di *datawarehouse* e *business intelligence* (ruolo strategico con diffusione a livello aziendale o ruolo dipartimentale o funzionale).

Architettura d): stabilità, affidabilità e buoni livelli di performance

maggiore complessità, costi superiori, minore flessibilità

Architetture a) b) e c): semplicità, veloci da implementare e mantenere

minore sensibilità alla "correttezza concettuale" dell'architettura dei dati

Business Intelligence:

- ricerca intelligente di dati
- produzione e analisi in “tempo reale” di informazioni
- “push”, ma soprattutto “pull”
- per il supporto ad attività e processi di controllo e di decisione di manager e professional (*knowledge information workers*) di qualunque livello aziendale

Sistemi di Business Intelligence:

- sistemi informativi dedicati alla B.I.
- integrano e completano i sistemi di Datawarehousing

Le funzionalità disponibili:

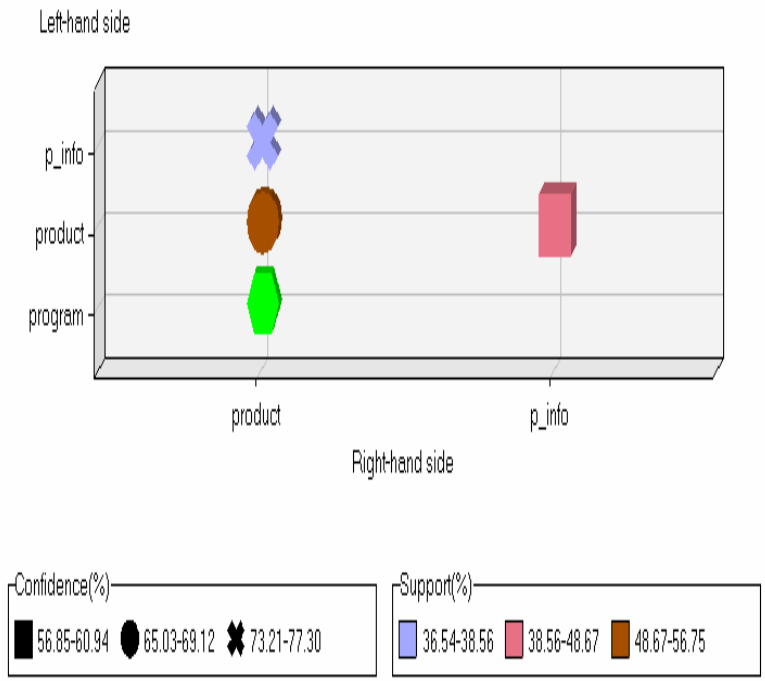
- *Cruscotto aziendale*, *tableau de bord e di scorecard*: presentare in modo statico ed in modalità *push* le informazioni mediante efficaci visualizzazioni grafiche – semafori, tachimetri, *business graphics*, icone e forme grafiche di vario genere.

Collegare gli obiettivi quantitativi alle loro metriche di misurazione
Definire legami di causa-effetto tra le misure

- *Visualizzazione e reporting tabellare e grafico*: visualizzare le informazioni in formato tabellare e grafico (bi-tridimensionale). I *report* prodotti possono essere automaticamente aggiornati ed inviati via e-mail, per fax, su dispositivi portatili per utenti predefiniti. Sono anche disponibili visualizzazioni grafiche che utilizzano forme cartografiche che posizionano i dati georeferenziati su una cartina geografica.

Esempi di report

	Chain Length	Support(%)	Confidence(%)	Transaction Count	Rule
248	2	2.08	4.64	469	catalog ==> regpost
249	2	2.03	10.81	458	cart ==> help
250	2	2.01	10.67	452	cart ==> agb
251	2	2.00	4.12	450	p_info ==> download
252	3	43.93	63.78	9896	program ==> product ==> product
253	3	38.46	55.84	8665	program ==> product ==> p_info
254	3	36.27	75.26	8170	product ==> p_info ==> product
255	3	33.49	84.65	7544	catalog ==> program ==> product
256	3	33.09	59.75	7455	product ==> product ==> product
257	3	30.14	84.64	6789	home ==> program ==> product
258	3	28.80	52.00	6487	product ==> product ==> p_info
259	3	28.74	75.49	6475	program ==> program ==> product
260	3	28.73	74.57	6473	program ==> p_info ==> product



Le funzionalità disponibili:

- *Calcolo di indicatori derivati*: a partire da indicatori disponibili utilizzando funzioni matematiche elementari, funzioni statistiche elementari oppure funzioni di *ranking* dei valori che si stanno analizzando vengono calcolati nuovi indicatori.

Indicatore di fatturato per metro quadro di superficie espositiva del punto vendita

- *Analisi multidimensionale*: permettono di analizzare i dati secondo diverse dimensioni di analisi, anche simultaneamente, sfruttando l'applicazioni di filtri e le funzionalità di *pivoting*, *drill* e *slice & dice*. Tale analisi risulta basata su tecnologie di tipo OLAP (*On-line Analytical Processing*).

OLAP (On-Line Analytical Processing)

Consente a tutti gli utenti le cui necessità di analisi non siano facilmente identificabili a priori di analizzare ed esplorare interattivamente i dati sulla base del modello multidimensionale.

Mentre gli utenti degli strumenti di reportistica svolgono un ruolo essenzialmente passivo, gli utenti OLAP sono in grado di costruire attivamente una sessione di analisi complessa in cui ciascun passo effettuato è conseguenza del passo effettuato al passo precedente.

OLAP (On-Line Analytical Processing)

Una sessione OLAP consiste in pratica in un percorso di navigazione che riflette il procedimento di analisi di uno o più fatti di interesse sotto diversi aspetti e a diversi livelli di dettaglio.

Questo percorso si concretizza in una sequenza di interrogazioni che spesso non vengono formulate direttamente, ma per differenza rispetto all'interrogazione precedente.

OLAP (On-Line Analytical Processing)

Il risultato delle interrogazioni è di tipo multidimensionale; poiché le capacità umane di ragionare in più di tre dimensioni sono molto limitate, gli strumenti OLAP rappresentano tipicamente i dati in modo tabellare evidenziando le diverse dimensioni mediante intestazioni multiple, colore, etc.

OLAP: principali operazioni ed analisi effettuabili

- Drill-Down: disaggrega i dati di un report
  1. Aggiunge una dimensione
  2. Aumenta il livello di dettaglio di una dimensione

     es. da mese a giorno
- Roll-Up: inverso del drill-down
  1. Elimina una dimesione
  2. Diminuisce il livello di dettaglio di una dimensione

     es. da giorno a mese
- Rotation o Pivoting: riorienta il data-cubo
  1. Cambia le modalità di presentazione delle dimensioni analizzate portando in primo piano una differente combinazione di dimensioni

Esempio di Drill-down e Roll-up

down

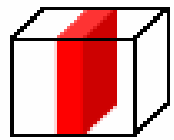
up

Dipartimento	Incassi	Unità vendute
Panificio	Lit. 12100000	5088
Cibo surgelato	Lit. 23000000	15000
...		

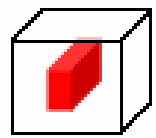
Dipartimento	Marca	Incassi	Unità vendute
Panificio	Barilla	6000000	2600
Panificio	Agnesi	6100000	2488
Cibo surgelato	Findus	15000000	6500
Cibo surgelato	Orogel	8000000	8500
...			

OLAP: principali operazioni ed analisi effettuabili

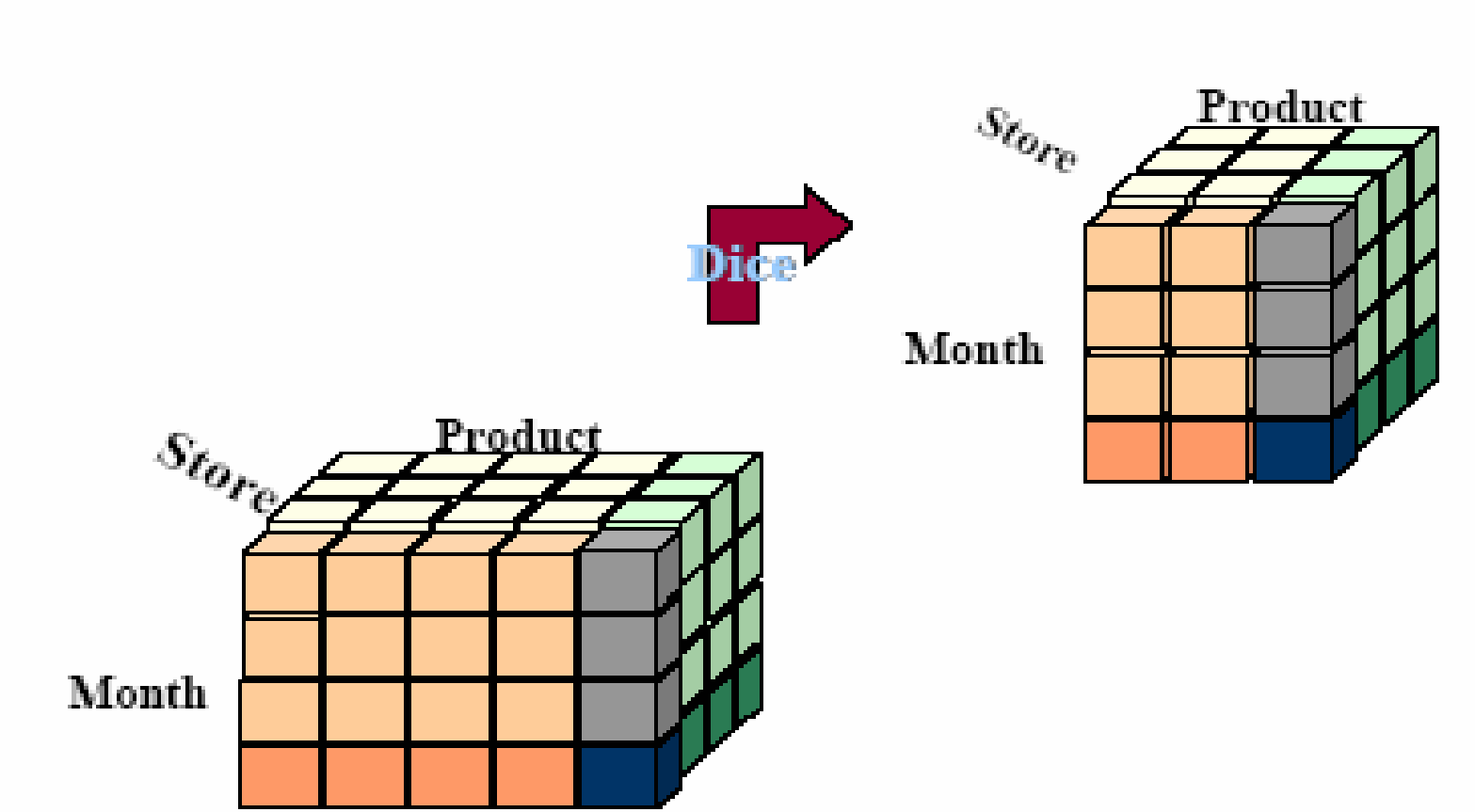
- Slice
  1. Produce “una fetta” dell’ipercubo
  2. Consiste in una selezione con un vincolo di uguaglianza



- Dice
  1. Produce un ipercubo più piccolo estratto da quello corrente
  2. Consiste in una selezione con uno o più vincoli di uguaglianza e di range combinati con OR e/o AND



Esempio di Dicing



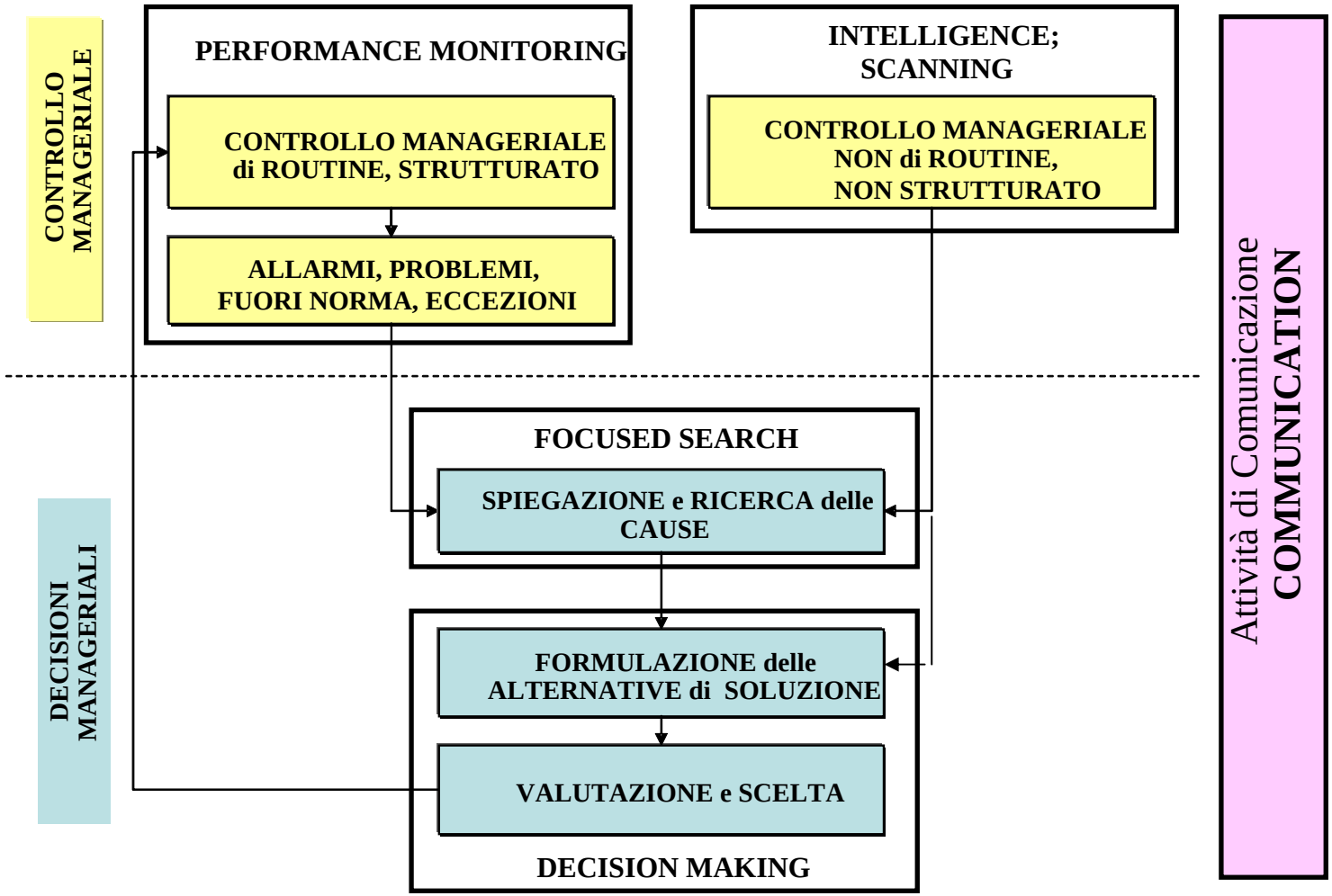
OLAP: principali operazioni ed analisi effettuabili

- Drill-accross
  1. Stabilisce un collegamento tra più cubi correlati (comparazione delle misure)
  2. Consente di calcolare funzioni matematiche e statistiche che coinvolgono le misure prese dai due cubi
- Drill-through
  1. Solo alcuni strumenti OLAP lo prevedono
  2. Consiste nel passaggio dai dati del DW ai dati operazionali presenti nelle sorgenti

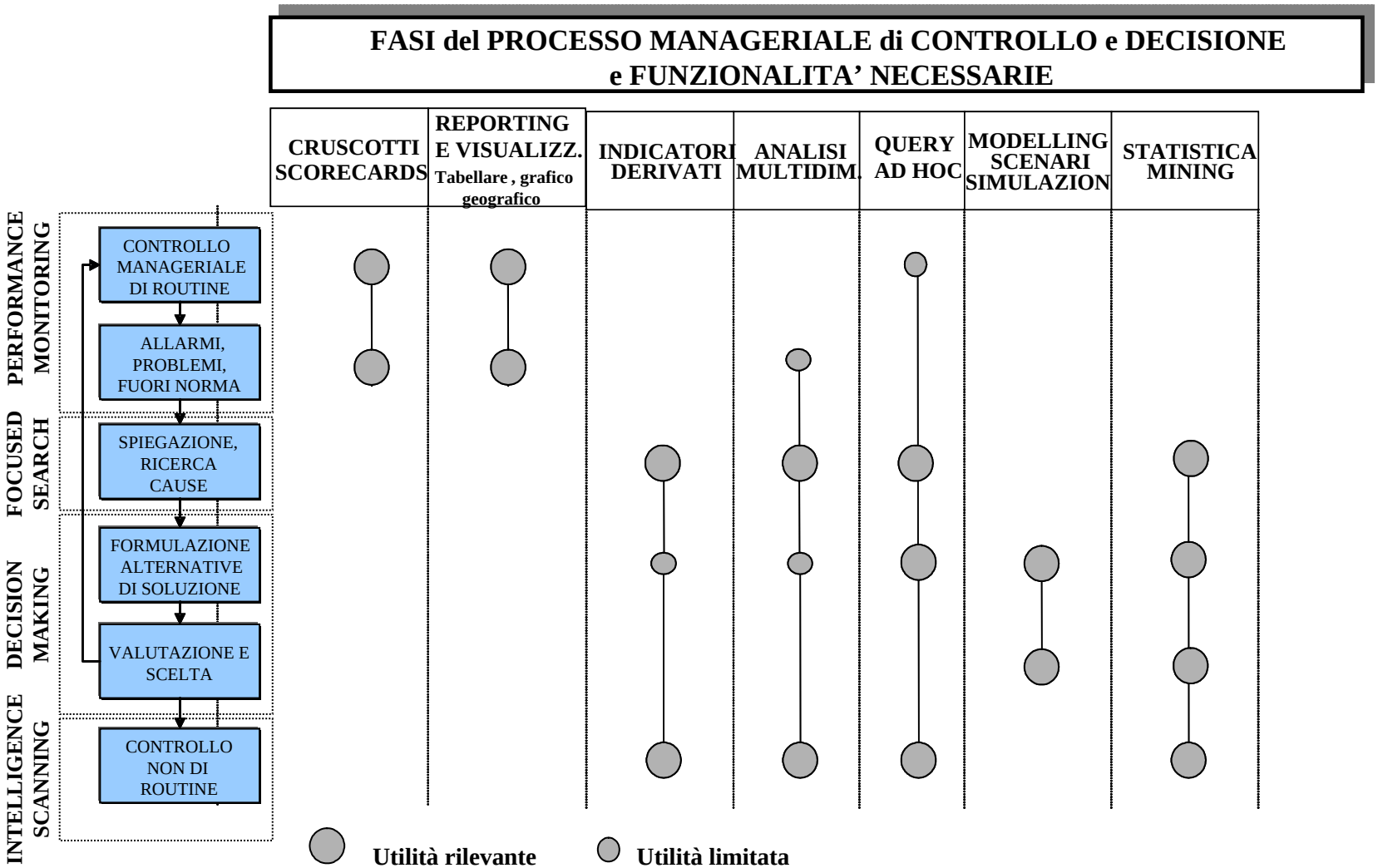
Le funzionalità disponibili:

- *Funzionalità di query*: consentono di interrogare liberamente il *datawarehouse* costruendo interrogazioni (*query*) libere, in maniera *pull* (cioè secondo criteri personali del decisore che ha esigenze informative non definite a priori e di tipo non routinario e strutturato) **senza conoscere** il linguaggio di interrogazione.
- *Modelling, costruzione scenari e simulazione*: definiscono il problema logico-matematico di un problema aziendale da impiegare nella costruzione di scenari e nella simulazione aziendale.
- *Statistica e di mining nei dati*: applicazione ai dati aziendali di sofisticate funzioni statistiche (*clustering*, regressione, ecc.) che permettono di scoprire relazioni logiche tra dati altrimenti non facilmente individuabili.

Il processo di controllo e di decisione manageriale



Sistemi informativi direzionali – i sistemi di BI



Data mining

Per data mining s'intende il processo di **selezione, esplorazione e modellazione** di grandi masse di dati, al fine di scoprire regolarità o relazioni non note a priori in modo automatico o semiautomatico.

E' un approccio multidisciplinare che riunisce un insieme di tecniche quali la statistica, la visualizzazione e i sistemi basati sulla conoscenza ed i sistemi ad autoapprendimento, finalizzato al miglioramento dei processi conoscitivi ed a ridurre l'incertezza legata all'assunzione di decisioni.

Si configura come una delle fasi del complesso processo di **scoperta di conoscenza nei database** (KDD).

Il processo di KDD

KDD - Introduzione

- Crescita notevole degli strumenti e delle tecniche per generare e raccogliere dati (introduzione codici a barre, transazioni economiche tramite carta di credito, dati da satellite o da sensori remoti, servizi on line ...)
- Sviluppo delle tecnologie per l'immagazzinamento dei dati, tecniche di gestione di database e *datawarehouse*, supporti piu' capaci e piu' economici (dischi, CD) hanno consentito l'archiviazione di grosse quantita' di dati

KDD - Introduzione

- Simili volumi di dati superano di molto la capacità di analisi dei metodi manuali tradizionali, come le query ad hoc. Tali metodi possono creare report informativi sui dati ma non riescono ad analizzare il contenuto dei report per focalizzarsi sulla conoscenza utile
- Emerge l'esigenza di utilizzare tecniche e strumenti con la capacità di assistere in modo *intelligente* e *automatico* gli utenti decisionali nell'estrazione di elementi di conoscenza dai dati

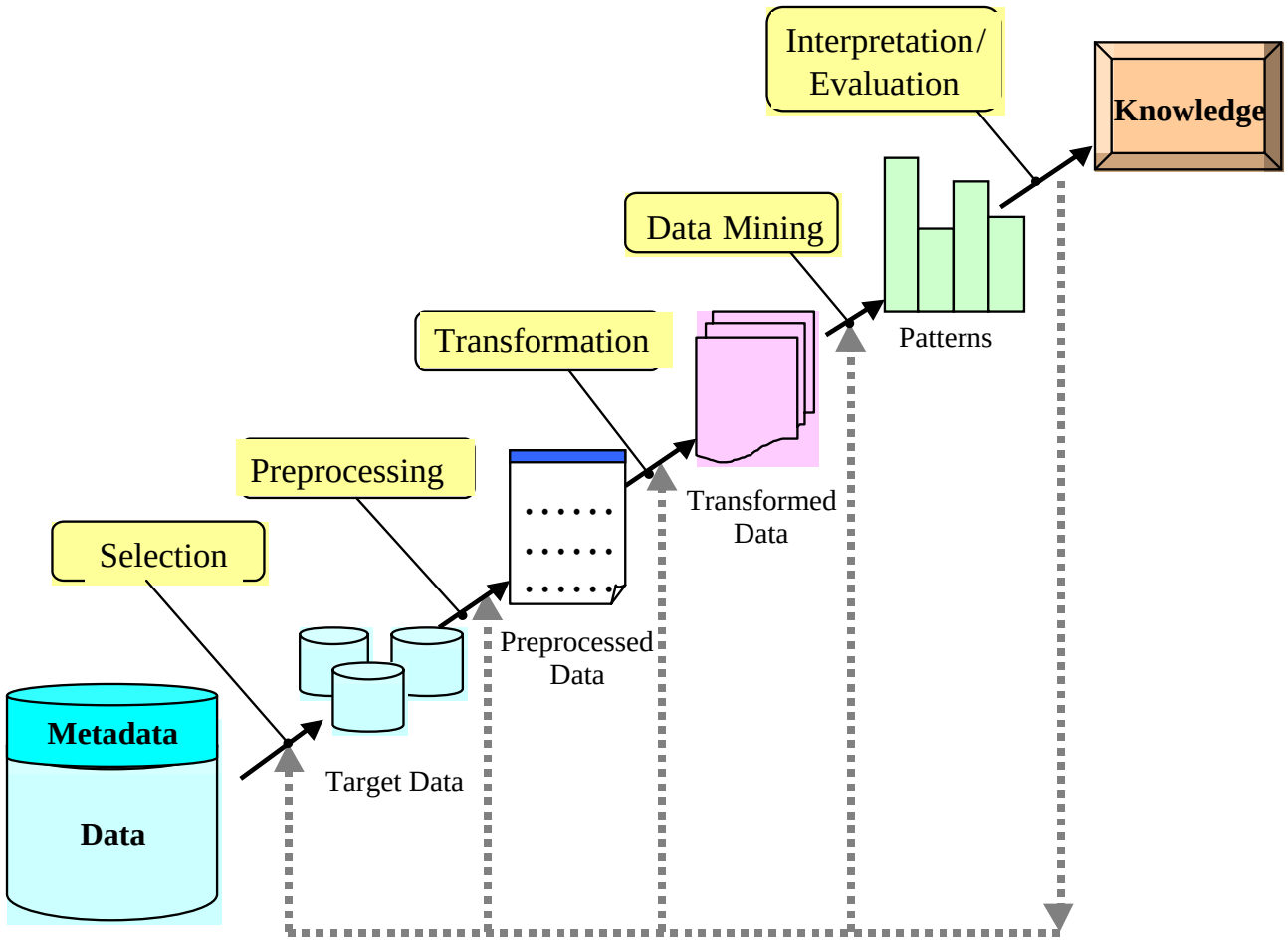
KDD - Introduzione

- Il termine ***knowledge discovery in databases*** indica l'intero processo di ricerca di nuova conoscenza dai dati
- Il termine di ***data mining*** si riferisce all'applicazione di algoritmi per estrarre pattern dai dati senza considerare gli ulteriori passi che caratterizzano il processo di KDD (come, ad esempio, incorporare appropriata conoscenza a priori e fornire una opportuna interpretazione dei risultati)

KDD - Introduzione

- Pertanto l'intero processo, tipicamente interattivo e iterativo, di ricerca, estrazione ed interpretazione di pattern dai dati, che indichiamo come KDD, coinvolge l'applicazione ripetuta di specifici metodi e algoritmi di data mining e l'interpretazione dei pattern generati da tali algoritmi
- Nel seguito forniremo una definizione più dettagliata di KDD e una panoramica sui metodi e gli algoritmi di data mining più usati

Il processo di KDD



Il processo di KDD: le fasi

1. Sviluppo e approfondimento del dominio di applicazione, della conoscenza disponibile a priori e degli obiettivi dell'utente finale.
2. Creazione di un target data set: selezione del data set o focalizzazione su un sottoinsieme di variabili o di campioni di dati oggetto del processo KDD.
3. Cleaning dei dati e preprocessing: operazioni di base come la rimozione del rumore o degli *outliers* se è il caso, raccolta delle informazioni necessarie per modellare o tener conto del rumore, messa a punto di strategie per gestire i dati mancanti e per gestire i dati tempo-varianti.

Il processo di KDD: le fasi

4. Riduzione dei dati e proiezione: rappresentazione dei dati in modo opportuno in relazione agli obiettivi della ricerca. Riduzione delle dimensioni e impiego di metodi di trasformazione per ridurre l'effettivo numero di variabili da sottoporre al processo di ricerca.
5. Scelta del compito del processo di data mining: identificazione dell'obiettivo del KDD, stabilire, cioè se si tratti di una classificazione, di una regressione, di un clustering…
6. Scelta dell'algoritmo o degli algoritmi di data mining: selezione dei metodi da usare per ricercare pattern nei dati. Questa fase comprende la decisione su quali modelli e parametri potrebbero essere appropriati e il matching di un particolare metodo di data mining con i criteri generali del processo KDD (per es. l'utente finale potrebbe essere maggiormente interessato alla comprensione del modello piuttosto che alle sue capacità predittive).

Il processo di KDD: le fasi

7. Data mining: ricerca di pattern di interesse in una particolare forma di rappresentazione o su un set di rappresentazioni diverse (regole di classificazione, alberi decisionali, regressione, clustering…). Il risultato del processo di data mining è considerevolmente influenzato dalla correttezza delle fasi precedenti.
8. Interpretazione dei pattern trovati e possibile ritorno alle fasi iniziali per ulteriori iterazioni.
9. Consolidamento della conoscenza estratta: incorporazione di tale conoscenza nel sistema di performance o, semplicemente, documentazione e reporting alle parti interessate. Questa fase include anche il controllo per la risoluzione di potenziali contraddizioni con la conoscenza precedentemente disponibile.

Gli algoritmi di Data mining

Differenze tra Data Retrieval e Data Mining

Data Retrieval

- Quanti sono i clienti che hanno età tra 30 e 50 anni e comprano Diet Coke
- Quali documenti contengono la parola “Sanità”
- Quanti brevetti ha depositato la società Colgate nel 1999

Data Mining

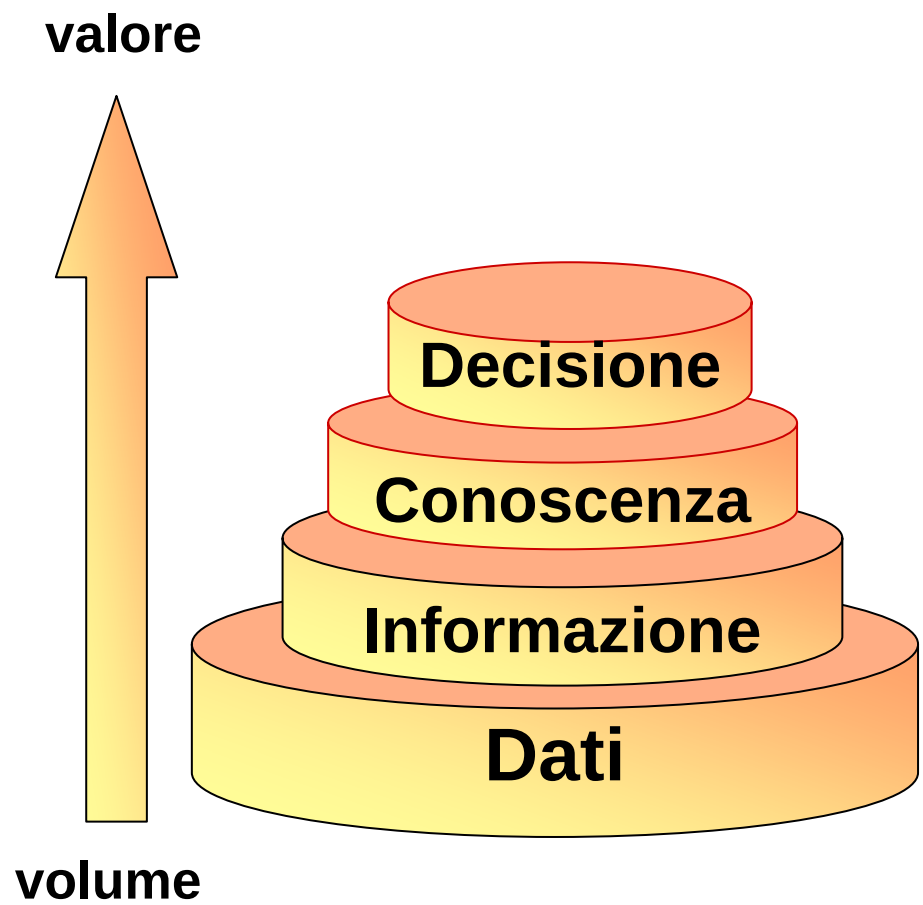
- Quali sono le caratteristiche dei miei clienti
- Quali sono gli argomenti trattati da un insieme di documenti
- Quali sono i miei concorrenti e come evolve la loro attività

I vantaggi del Data Mining

- Trattamento di dati quantitativi, qualitativi, testuali, immagini e suoni
- Non richiede ipotesi a priori da parte del ricercatore
- Non richiede ipotesi sulla forma distributiva delle variabili
- Possibilità di elaborare un numero elevato di osservazioni
- Possibilità di elaborare un numero elevato di variabili
- Algoritmi ottimizzati per minimizzare il tempo di elaborazione
- Semplicità di interpretazione del risultato
- Visualizzazione dei risultati

Perché sono necessari strumenti di Data Mining

- Quantità dei dati
- Natura dei dati
- Rapida evoluzione del mercato
- Inadeguatezza degli strumenti tradizionali



Ambiti di applicazione del Data Mining

- **L'Analisi per gruppi** suddivide una popolazione in sottoinsiemi disgiunti secondo definiti criteri.
- **La Classificazione** cataloga un fenomeno in una certa classe secondo un insieme di regole predeterminate.
- **Le Regole d'associazione** sono legami di casualità validi tra gli attributi delle osservazioni di un *data set*.

Metodi di Data Mining

- **Esplorazione mediante la visualizzazione multidimensionale** (scaling multimensionale, analisi di regressione logistica, stepwise, analisi delle corrispondenze)
- **Associazione e sequenze** (usate nella market basket analysis per misurare l'affinità dei prodotti)
- **Clustering** (segmentazione della clientela in gruppi omogenei)
- **Analisi Fattoriale** (per determinare il numero di fattori da estrarre)
- **Modelli previsivi**
  - di classificazione (**Alberi di Decisione**)
  - **Reti Neurali**
- **Mappe di Kohonen** (Reti Neurali non supervisionate)
- **Algoritmi Genetici**

Vantaggi e limiti dei più diffusi sistemi di data mining

Tecniche	*Vantaggi*	*Limiti*
Visualizzazione	L'utente è in grado di visualizzare grandi moli di dati, di scoprire relazioni e di testarle	Richiede un utente esperto in statistica e in grado di utilizzare altre tecniche di data mining
Reti neurali	Elevata capacità elaborativa con dati in cui si nascondono relazioni non lineari. Lavora con dati incompleti	Incapacità di fornire spiegazioni sui risultati ottenuti sebbene sia possibile utilizzare altri sistemi in grado di fornire un'interpretazione. I dati qualitativi devono essere convertiti in quantitativi
Algoritmi genetici	Buona capacità previsionale utilizzando dati in cui si nascondono relazioni non lineari	"Come sopra"
Fuzzy logic	Può classificare variabili e risultati sulla base di vicinanza alla soluzione desiderata	Numero limitato di fornitori e applicazioni disponibili sul mercato
Decision tree e rule induction	Creano regole e modelli sulla base di dati storici. Le regole ed i modelli sono trasparenti all'utente e facilmente interpretabili	Richiedono un tuning ottimale per evitare la produzione di elevati numeri di regole difficilmente interpretabili e gestibili

Esempi di applicazioni

Esempio	***Tipo di problema***	***Tecnica adottabile***
Quali sono i tre principali motivi che hanno indotto il mio cliente a passare alla concorrenza?	Classificazione	Reti Neurali Decision Tree
Quali sono le fasce di clienti a cui posso offrire nuovi prodotti?	Clustering	Reti Neurali Decision Tree
Quali sono le probabilità che un cliente che ha aperto un c/c acquisterà anche il prodotto x in breve tempo?	Sequencing	Tecniche statistiche Rule induction
Quali sono le probabilità che un cliente acquisti due prodotti completamente differenti?	Associazione	Tecniche statistiche Rule induction
Quale sarà il prezzo del titolo tra un giorno/mese ecc?	Previsione	Reti neurali Tecniche statistiche

Ma non se ne occupava la statistica?

J. Kettenring (ex- presidente dell'ASA) definisce la statistica come "*la scienza di apprendere dai dati*"

Tecniche statistiche orientate alla scoperta di strutture di relazione e di modelli

- **Analisi esplorativa**
- **Analisi esplorativa multivariata**
  - Analisi delle componenti principali
  - Analisi delle corrispondenze
  - Analisi dei cluster
  - Ecc.

Data mining - considerazioni conclusive

Cosa c'e' di nuovo nel Data mining?

• La possibilità di gestire enormi quantità di dati, che rendono obsoleta la definizione classica di grandi campioni (miliardi di record e terabytes di dati non sono inusuali)

•Le recenti tecniche che provengono dal mondo dell'ingegneria informatica (reti neurali, alberi di decisione, regole di inclusione)

•Interessi commerciali nel valorizzare le informazioni esistenti al fine di proporre soluzioni "individuali" per una determinata categoria di clienti

•Disponibilità di nuovi pacchetti, di facile uso, diretti sia a coloro i quali devono assumere le decisioni che agli analisti (ma molto più costosi!)

Data mining - considerazioni conclusive

Il Data mining è una disciplina in grande crescita che si è sviluppata al di fuori della statistica nel mondo dei DBMS, principalmente per motivi commerciali.

Oggi il DM si può considerare come una branca della statistica esplorativa con l'obiettivo di individuare *inattesi* e *utili* modelli e regolarità nei dati mediante l'uso di algoritmi classici e nuovi.

AVVERTENZE ALL'USO

L'espressione *inattesi* non deve essere fuorviante: un ricercatore ha una maggiore possibilità di scoprire qualcosa di interessante se **ha familiarità con i dati.**

L'*utilità* delle regolarità individuate nella struttura dei dati va verificata. **Le associazioni sono solo correlazioni e non implicano relazioni di causa-effetto.**

Non va infine dimenticato che nell'applicazione di questi algoritmi è necessario effettuare valutazioni dell'incertezza e del rischio e pertanto non si può prescindere dall'uso di test per la verifica della validità dei risultati ottenuti (suddividere la base di dati in sotto campioni e verificare se si ottengono gli stessi risultati).

Per concludere

Business Intelligence

In ambito aziendale l'insieme delle applicazioni, dei programmi e delle tecnologie usate per raccogliere, immagazzinare, analizzare e garantire accesso ai dati finalizzate a supportare gli utenti a prendere decisioni di business più efficaci viene indicato con il termine **business intelligence** (BI).

Le applicazioni di BI includono, quindi, le attività di: supporto alle decisioni, interrogazione e reporting, OLAP, analisi statistica e KDD.

Pianificare programmare decidere – Enterprise Resourse Planning

L’ERP gestisce tutta l’azienda

Imprese di ogni dimensione stanno investendo grandi quantità di risorse nell’installazione di software per l’Erp con l’obiettivo di:

- migliorare i processi di business
- sostituire i sistemi informativi esistenti che diventano obsoleti.

Cosa ha determinato la loro evoluzione

La capacità di:

- gestire in modo integrato tutte le risorse che partecipano alla creazione di prodotti/servizi di un’azienda
- coprire totalmente il processo di business.

Pianificare programmare decidere – Business Intelligence

Concetto di base

Le aziende hanno nel DW una fonte di dati estremamente preziosa da cui trarre informazioni sui propri clienti e sul proprio funzionamento.

Questi dati se studiati e valorizzati possono portare a importanti risultati in termini di aumento del business o di efficienza semplicemente evidenziando e sfruttando fenomeni che accadono già.

La tecnologia permette di aumentare la possibilità di governare tutte le fasi di “intelligence” necessarie per arrivare alla decisione.

L’utente assume un ruolo sempre più importante:

il decisore oggi è colui il quale “decide come decidere” e su quali dati decidere!

Pianificare programmare decidere – Business Intelligence

La disponibilità di dati aziendali strutturati e condivisibili garantita dagli ERP, associata a strumenti OLAP, permette al manager che sfrutta le tecnologie di:

- "pescare" i dati che gli servono
- di strutturarli secondo i propri modelli di riferimento
- di simulare scenari possibili in fase di pianificazione
- di ricominciare da capo se il risultato non è soddisfacente.

Tutto ciò avviene in un lasso di tempo esprimibile in ore, non più in settimane, in formati chiari, leggibili, modificabili e non su tabulati interminabili e spesso inutilizzabili perché disponibili solo a decisione presa o a piano approvato.

Pianificare programmare decidere – Business Intelligence

Le tecnologie che rientrano nel BI danno la possibilità di sperimentare modelli e soluzioni diverse, di analizzare secondo segmentazioni o aggregazioni dinamiche.

Si registra un trend di progressiva semplificazione nell'uso; questi strumenti possono essere usati senza presupporre competenze tecniche superiori a quelle richieste per lavorare con i più diffusi strumenti applicativi di uso comune.

Per questo motivo appaiono in misura crescente sui computer dei manager e vengono utilizzati come utili strumenti di lavoro.

Pianificare programmare decidere – Intepretare le informazioni

La capacità di filtrare le informazioni e di interpretare diventa un cardine per evitare l'ingolfamento causato dalla sovrabbondanza di informazioni.

La soluzione non può essere quella di ridurre la quantità assoluto di informazione prodotta per renderla facilmente governabile.

Il valore aggiunto viene conseguito attraverso la qualità dell'organizzazione delle informazioni disponibili.

Queste possono essere strutturate secondo relazioni gerarchiche e criteri significativi che vanno dagli obiettivi strategici dell'azienda ai singoli processi consentendo, quindi, di concentrare l'attenzione su pochi, chiari e fondamentali elementi.

Pianificare programmare decidere – Intepretare le informazioni

I cruscotti direzionali, le balanced score card e tutti quegli strumenti che forniscono una rappresentazione sintetica delle performance critiche – garantendo la possibilità di “scavare” approfondire, dove gli indicatori di sintesi segnalino un problema – sono un prezioso aiuto al manager che è nella posizione da cui si ha una vista di insieme dell’azienda, ma il cui tempo limitato non gli consente di analizzarne tutti i dati.

Attraverso gli strumenti di BI le informazioni necessarie a prendere le decisioni e a valutarne gli effetti non si smarriscono nel rumore di fondo dell’enorme massa di dati chi i sistemi di gestione producono, ma vengono selezionate ed organizzate per essere fruibili in tempo reale.

Pianificare programmare decidere – Intepretare le informazioni

Muovendosi dalla prospettiva di insieme ci si può spostare a una di maggiore dettaglio (*slicing* o *drill down*), oppure cambiare prospettiva (*dicing*) si possono esplorare le informazioni per rispondere a domande come: “Cosa è successo nell’ultima settimana?” “Cosa succederebbe se … ?” “Quali sono i primi 10 clienti?” e a qualsiasi altra domanda che riguarda il proprio business su cui ci siano dati elementari prodotti dalle transazioni.

Il manager dell’era digitale è un figura che si affida ad alcune tecnologie per controllare e gestire la maggior parte delle componenti del processo decisionali.

Ha capacità modellistiche non specificatamente tecnologiche èur essendo attento alle opportunità che i nuovi strumenti gli propongono.

Pianificare programmare decidere – Intepretare le informazioni

Prende il controllo della situazione limitando gli interventi degli specialisti che non partecipano al processo decisionale.

E’ importante rivolgersi agli specialisti, interni ed esterni, solo per quei servizi altamente specifici caratterizzati da bassa ripetitività: uno studio di geomarketing o un’analisi con modelli a reti neurali sono ancora servizi che vanno richiesti all’esterno ma nella quotidianità tutto questo non è necessario.