GESTIONE DELLE BASI DI DATI

Guida all'approfondimento pratico

Dott. Roberto Puccetti
Dipartimento di Informatica
Università di Pisa

Indice

GESTIONE DELLE BASI DI DATI

Ogni entità strutturale, associazione, raggruppamento di interessi, ufficio, comuni, biblioteche, etc) ha esigenza di avere, trattare e recuperare un insieme di informazioni. Un *sistema informativo* è l'insieme dei dati e delle procedure che servono a gestire tali dati.

Queste procedure possono ovviamente modificare i dati su cui operano. Un sistema informativo è quindi qualcosa che prescinda dal calcolatore. Oggi si tende a parlare di *sistema informativo automatizzato* nel senso che le procedure di acquisizione e gestione dei dati sono automatizzate dal calcolatore. In genere non tutto il sistema informativo è automatizzato; chiameremo *sistema informatico* quella parte del sistema informativo che viene automatizzato. Naturalmente si cerca di automatizzare il più possibile. Mentre una volta il sistema informatico veniva organizzato in più archivi disgiunti sui quali agivano procedure diverse, ora si tende ad avere un'integrazione dei dati, un unico archivio senza ripetizioni. Tale archivio forma una BASE SI DATI e tutte le procedure operano su di essa.

Come si costruisce un sistema informatico.

Il sistema informatico nasce come specializzazione di un sistema informativo di cui si conoscono esattamente le esigenze degli utenti.

Analisi dei requisiti

Il primo passo è quello di rendere le esigenze degli utenti il più razionale possibile in modo da poter descrivere i requisiti che deve avere il sistema informativo. Questa è la fase più delicata ed importante nella costruzione del sistema informatico e serve ad

individuare le necessità di memorizzare certi dati e di avere determinate funzioni su di essi.

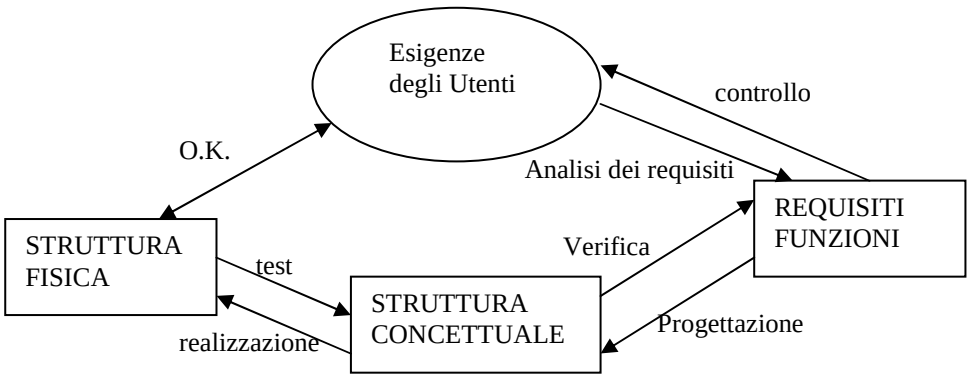
Progettazione

Il passo successivo è quello di dare una strutturazione logica a questi dati in modo che su di essi possono effettivamente venire eseguite le procedure richieste.

Implementazione

Successivamente si passa alla strutturazione fisica dei dati. E' qui che interviene il calcolatore.

Una volta costituito il sistema informatico occorre ripercorrere i passi eseguiti per controllare che non si siano introdotte incoerenze fra le fasi.



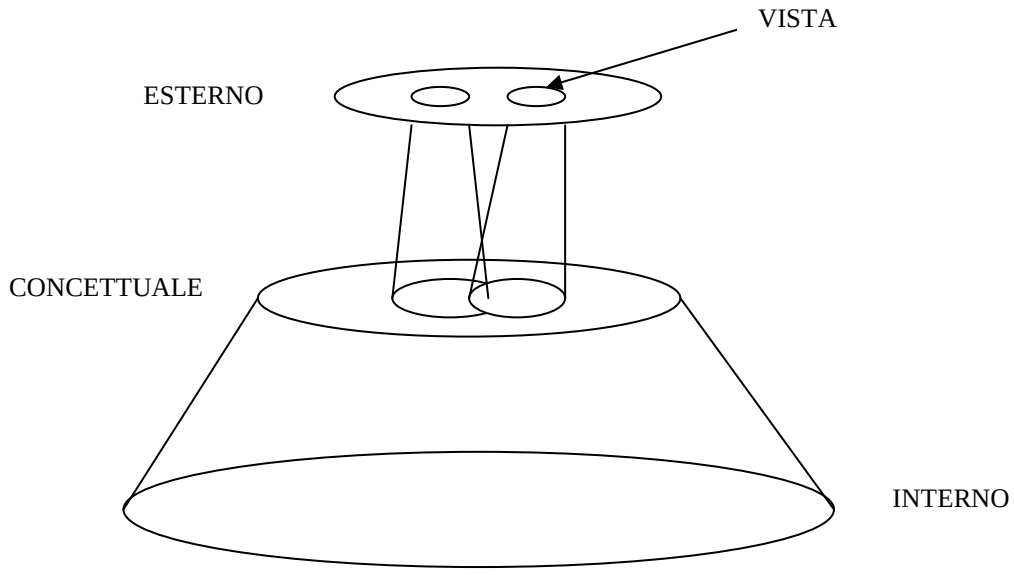
Quando si deve studiare una certa Base di Dati si devono realizzare, nei requisiti le esigenze dell'utente e contemporaneamente si deve avere una forma concettuale astratta dell'organizzazione dei dati. Lo *schema concettuale* è lo schema generale astratto dei dati che costituiscono il

sistema informatico. Tale schema è assolutamente indipendente dalla realizzazione fisica sul calcolatore. Ciò facilita la trasportabilità della Base di Dati (astratta) da un calcolatore all'altro: basterà cambiare lo *schema interno* dei dati.

Il passaggio da schema concettuale a schema interno viene chiamato *mapping logico-fisico*. Questa fase dipende dai mezzi a nostra disposizione; infatti l'HW condiziona il SW che condiziona il Mapping.

Lo *schema esterno* corrisponde ai requisiti, che possono essere immaginati come le esigenze di un ente e più precisamente come la somma delle esigenze dei vari uffici che gestiscono l'ente.

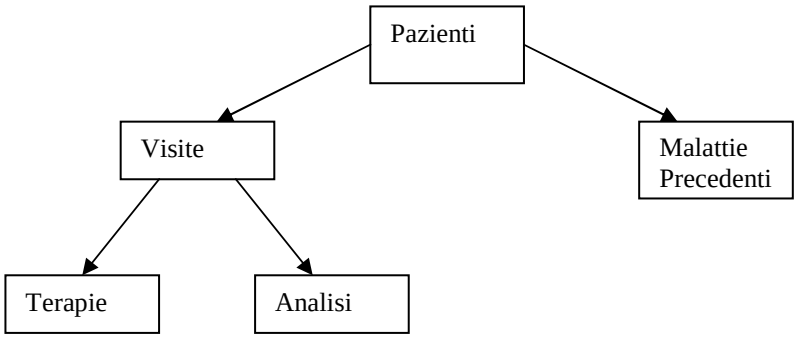
In generale ogni ufficio non sarà interessato alla globalità dei dati ma solo ad una parte di essi. Ogni ufficio avrà quindi una sua particolare *vista* dei dati.



Un esempio: Il sistema ospedaliero per la gestione dei pazienti.

Quando il paziente entra in ospedale gli verrà chiesta la sua anamnesi (malattie precedenti). Ad ogni paziente sarà associato un certo numero di visite. In conseguenza ad una visita i medici assegneranno al paziente delle analisi ed una terapia.

Schematizzando:



Questo tipo di organizzazione come struttura le informazioni ?

Immaginiamo 5 archivi (P,V,MP,T,A). L'archivio P conterrà i dati: nome, indirizzo, data di nascita, N. USL, luogo di nascita, ecc.. tutti i dati cioè che individuano univocamente il paziente e quelli che possono servire per indagini statistiche. Lo stesso vale per gli altri quattro archivi (immaginare quali dati) Quindi per creare un archivio (o FILE di DATI) bisogna focalizzare prima:

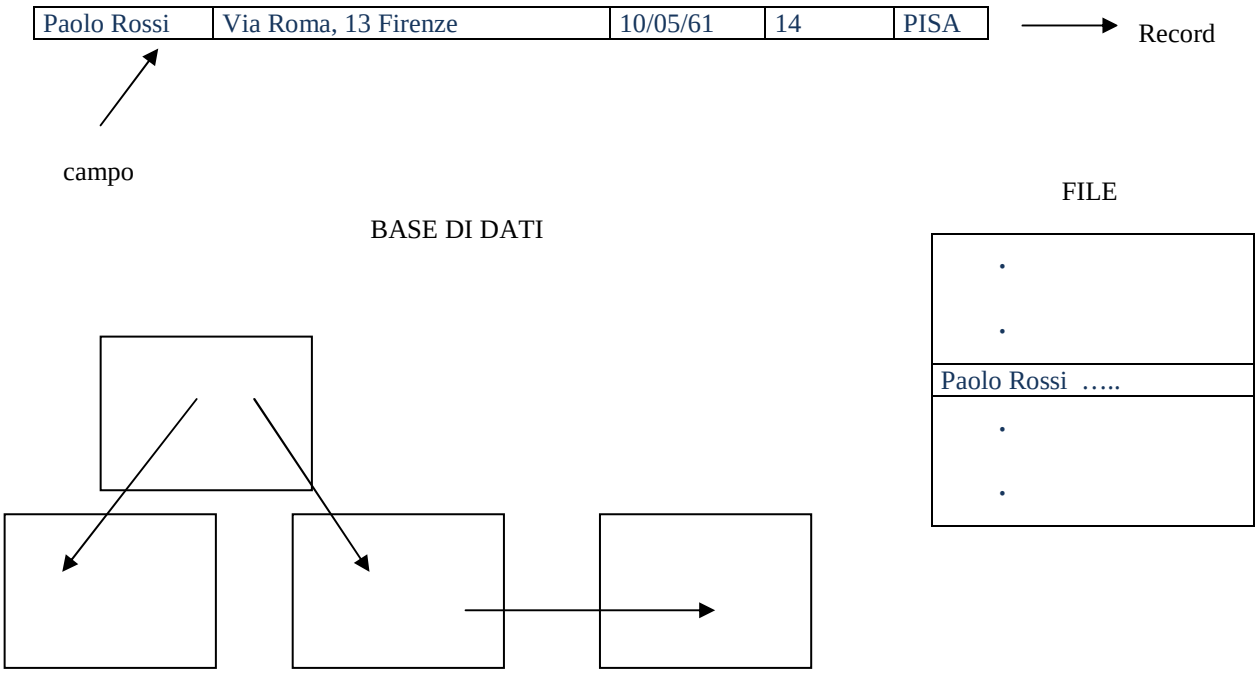
- quali sono le informazioni che si vogliono memorizzare
- in che forma si vogliono memorizzare
- quali delle informazioni sono significative per ritrovare i dati

Tornando al file P:

- le informazioni sono quelle elencate precedentemente

- il nome sarà composto da caratteri e fissato ad una lunghezza massima (es. 35 caratteri)
- lo stesso sarà per l'indirizzo
- la data di nascita sarà di tipo data (8)
- USL sarà un numero da i a 999
- luogo di nascita: caratteri (20)

Questi vincoli individuano il *dominio dei campi*. Per campo si intende ogni singola informazione componente il *record*. Per record si intende l'insieme delle informazioni associate ad un singolo elemento del *file*. Il file è la globalità delle informazioni.



Così come al file ed alla Base di Dati, anche ai campi viene associato un nome (*attributo*) mentre il record viene individuato dalla sua posizione all'interno del file e quindi da un numero.

Connessioni

Disegnando lo schema di fig. 3 abbiamo tracciato anche delle *connessioni* fra i file. Vediamo di analizzarne il tipo:

- Un paziente può essere visitato più volte in generale N volte. La connessione fra P e V viene chiamata quindi di tipo 1 a N.
- A sua volta una visita può dar luogo a più terapie ed analisi e quindi anche in questo caso si tratta di connessioni di tipo 1 a N.
- Oltre al tipo 1 a N esistono altri due tipi di connessioni:

a) 1 a 1:

si ha nel caso in cui si vogliano dividere i dati relativi ad una persona in due archivi separati uno contenente l'anagrafica l'altro il resto dei dati (per esempio per motivi di riservatezza)

b) N a M

ad esempio P e M è una connessione N a M perché ogni paziente può aver avuto più malattie ed ogni malattia può essere stata contratta da più pazienti

Come vengono realizzate queste connessioni all'interno dei file ?

Per rispondere introduciamo il concetto di *chiave primaria*:

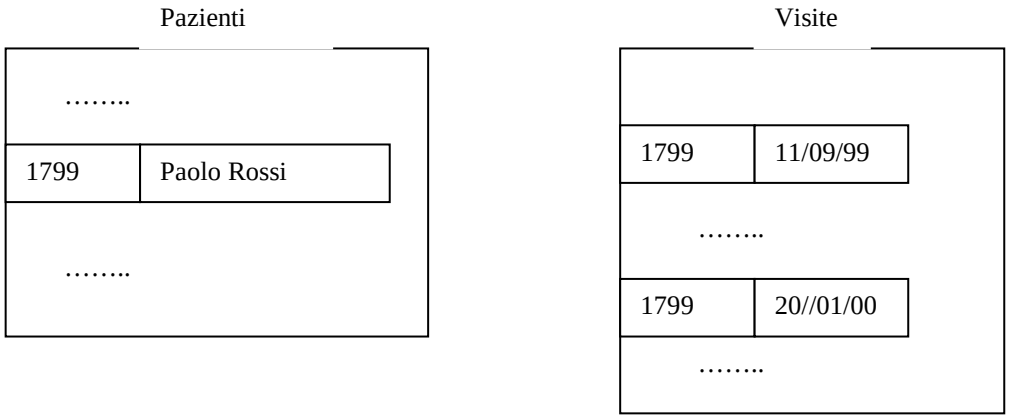
uno o più attributi che identificano univocamente le entità dell'archivio (un record)

Per esempio: il NOME (che comprendeva nome e cognome) non è una chiave primaria perché possono esistere benissimo più persone con lo stesso nome; NOME + DATA VISITA va già meglio ma può non bastare ad identificare univocamente un paziente.

In molti casi si introduce una *chiave primaria artificiale*. (Nell'esempio si potrebbe pensare al progressivo di entrata in ospedale). La chiave primaria artificiale ha inoltre il vantaggio di essere corta e quindi agevolmente maneggiata e memorizzata, ma ha l'inconveniente di non essere significativa per il record ad essa associato.

Connessioni 1 a N

Un modo naturale per connettere due archivi è quello di abbinare al secondo la chiave primaria del primo; ad esempio per connettere i P con le V si può introdurre in quest'ultimo archivio il campo NOME + DATA NASCITA. Risulta evidente quindi la convenienza di introdurre chiavi primarie artificiali le quali permettono di risparmiare spazio nella Base di Dati.

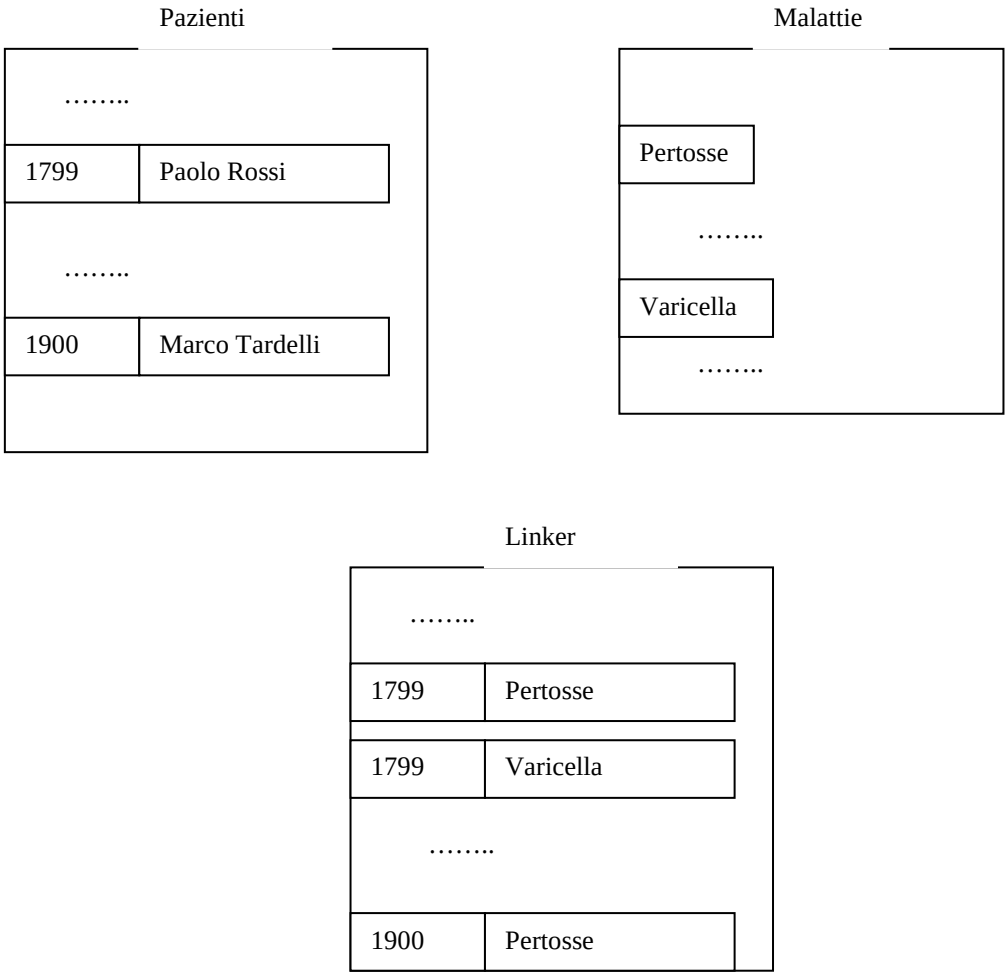


- Abbiamo creato così un collegamento *logico* tra i due archivi.
- Questo tipo è il più usato a livello di Sistemi di Gestione di B. di D. (applicativi) ma non è l'unica maniera di risolvere le connessioni fra archivi. Infatti esiste anche il collegamento *fisico* realizzato attraverso puntatori a catena: per il paziente viene ricordato il

numero del record contenente la prima visita, in questa il numero del record contenente la seconda e così via fino all'ultima visita. Da notare che in questo caso lo spazio occupato è sempre lo stesso qualunque sia la lunghezza della chiave primaria del primo archivio. Questo metodo è molto usato nelle realizzazioni a "basso livello" di B. di D.

Connessioni N a M

Il metodo più usato è quello di creare un altro archivio (*linker*), i record del quale sono composti dalle chiavi primarie dei record associati nei due archivi da collegare.



Il linker permette anche di memorizzare le informazioni relative alla connessione non ai singoli archivi: ad esempio se vogliamo memorizzare la data nella quale il paziente ha contratto una certa malattia, questa è una informazione associata al linker e quindi un campo dei suoi records. Da notare che anche questo è un collegamento logico: se voglio elencare tutte le malattie di un certo paziente, devo scorrere tutto l'archivio linker per cercare in quali record compare la chiave primaria del paziente. Anche per questo tipo di connessione esiste una soluzione attraverso il collegamento fisico dei records (provare ad immaginare tale soluzione), che complica l'organizzazione degli archivi ma ottimizza il tempo di ricerca.

Modelli di B.di D.

I sistemi di B. di D. si dividono in tre grossi modelli a seconda del tipo di connessioni presenti:

- GERARCHICO: il più semplice e gestisce le connessioni del tipo 1a1 e 1aN. Per questo modello non esiste nessun tipo di standardizzazione, nel senso che una realizzazione gerarchica non può essere trasportata su un sistema diverso da quello sul quale è stata creata, anche se ancora gerarchico: bisogna ridefinire tutti i file le procedure su di essi (utilizzando lo stesso sistema concettuale).
- RETICOLARE: generalizzazione di 1 e permette di gestire anche connessioni N a M: Per questo modello è stata tentata una stesura di specifiche (CODASYL) ma a causa degli interessi dei costruttori, la trasportabilità dei sistemi esistenti non è completa

- RELAZIONALE: generale quanto 2 ma di impostazione totalmente astratta. Trattandosi di modello astratto e teorico, il problema non si pone in quanto la teoria di base è la stessa e da essa discendono le realizzazioni.

Operazioni sugli archivi

Ricordiamo che per definire una B. di D. bisogna definire:

- Struttura degli archivi
    - Raggruppamento delle informazioni
    - Definizioni delle chiavi primarie

- Struttura delle connessioni
    - individuare quali archivi sono connessi fra loro
    - il tipo delle connessioni e relative realizzazioni con la eventuale aggiunta di archivi linker o campi di connessione

- Struttura dei singoli record
    - Denominazione dei campi
    - Definizione del dominio

- Operazioni sugli archivi

  Possiamo raggruppare tali operazioni in due insiemi:

    - Operazioni standard
    - Operazioni particolari

Le prime sono quelle fondamentali e che tutti i sistemi di gestione di B. D. mettono a disposizione:

Introduzione dati

Questa operazione deve consentire un controllo sul dominio dei dati. Infatti proteggere l'operatore da eventuali errori di introduzione dati è una prima maniera per garantire l'efficienza del sistema.

Ricerca

Il ritrovamento dei dati inseriti è una fase fondamentale nella gestione della B. di D. Questo avviene attraverso l'uso di CONDIZIONI di RICERCA, parametri cioè ai quali deve sottostare il dato desiderato. La complessità delle condizioni di ricerca dipende dalla possibilità del sistema di gestione della B. di D.

Modifica

Questa operazione viene fatta una volta posizionati sull'informazione trovata grazie alla ricerca e vale quanto detto per l'inserimento. Può essere conveniente "coprire" alcuni dati e permettere la modifica solo degli altri; per esempio cambiare il valore di una chiave primaria di un record del file P, implica la correzione di tutti i record di V che fanno riferimento ad esso. E' quindi conveniente non permettere tale operazione

Cancellazione

Anche per questa bisogna prima selezionare i dati attraverso la ricerca. Le connessioni risolte attraverso la soluzione logica

complicano molti meno problemi che non quelle risolte attraverso la soluzione fisica. Infatti questa ultima implica una gestione di puntatori che devono essere continuamente aggiornati sia in inserimento che in cancellazione e modifica.

Selezione

Formalmente assomiglia all'operazione di ricerca, ma mentre quest'ultima per mette di posizionarsi sul primo dato trovato, la selezione recupera tutti i dati che soddisfano la condizione.

Tutte queste operazioni hanno una particolare importanza per l'efficienza del sistema. Esse infatti svolgono le funzioni basilari sui dati e sono quindi le più usate: velocizzare le operazioni standard vuol dire velocizzare la gestione della B. di D. Poiché la realizzazione di queste è fatta all'interno del sistema di gestione di B. di D., l'unico mezzo che abbiamo a disposizione per cercare di ottimizzare le prestazioni della nostra B. di D. è organizzare in maniera razionale i dati. Supponiamo infatti di avere un elenco di persone:

Giuseppe

Irene

Fabio

Carlo

Lola

Elena

Fabio

Se dobbiamo ricercare un nome, la procedura da fare è:

1. leggi il primo nome della lista
2. se nome trovato fermati
3. leggi prossimo nome torna a 2

Nel caso migliore leggo 1 volta. Nel caso peggiore leggo 7 volte.

Mediamente leggo $\frac{1+7}{2} = 4$ volte

Se il dato non è presente si casca sempre nel caso peggiore. Supponiamo ora di organizzare l'elenco precedente in ordine alfabetico:

Carlo

Elena

Fabio

Fabio

Giuseppe

Irene

Lola

Il problema di prima si può risolvere con la stessa procedura che porta allo stesso numero di letture. Cambia però il caso del dato non presente che rientra nelle letture medie. Se poi cambio procedura:

1 leggi nome centrale
2. se nome trovato fermati
3. se nome > cercato considera la metà superiore del file
4. se nome < cercato considera la metà inferiore del file
5. vai a 1

Nel caso migliore leggo 1 volta. Nel caso peggiore leggo 3 volte.

Mediamente leggo $\frac{1+3}{2} = 2$ volte

Il vantaggio risulta evidente anche nel caso di una selezione delle persone che hanno lo stesso nome.

Il caso si presenta più complicato quando aumenta il numero dei campi e il numero delle ricerche possibili.

1)	Giuseppe	Rossi
2)	Irene	Bianchi
3)	Fabio	Verdi
4)	Carlo	Bianchi
5)	Lola	Neri
6)	Elena	Bianchi
7)	Fabio	Rossi

Se si vuol ricercare sia per nome che per cognome non si può dare un'organizzazione unica ai dati e non si può neanche riordinarli prima di ciascuna ricerca.

Si usano quindi degli *indici* che sono archivi a parte nei quali si ha il riepilogo dei dati organizzati e la maniera di recuperare l'intero record nell'archivio principale che rimane inalterato:

Nome		**Cognome**	
Carlo	4	Bianchi	2, 4, 6
Elena	6	Neri	5
Fabio	3, 7	Rossi	1, 7
Giuseppe	1	Verdi	3
Irene	2		
Lola	5		

Per alcuni sistemi di gestione di B. di D. basta definire gli indici, per altri bisogna pensare anche all'aggiornamento.

Le operazioni particolari, quelle cioè che non rientrano nell'elenco precedente, dipendono dalle specifiche del problema. Il sistema di gestione di B. di D. deve poter permettere la definizione di tali operazioni. Un esempio può essere:

- ricavare il valore di un campo in funzione di altri campi

(Totale = imponibile + IVA oppure prezzo di vendita = prezzo di costo + ricarico)

Una volta strutturato il progetto (struttura concettuale) della B. di D. questa va realizzata fisicamente sul calcolatore; ciò avviene tramite il sistema di gestione di B. di D. (database) i limiti del quale possono portare a modifiche anche sostanziali del prospetto stesso. Il lavoro fin qui svolto però non deve andare perduto, in quanto sarà la base per una nuova realizzazione qualora migliorino i mezzi messi a disposizione.