Capitolo 1

Introduzione alle Basi di Dati

1.1 Sistemi informativi, informazioni e dati

Ogni organizzazione è dotata di un *sistema informativo*, che organizza e gestisce le informazioni necessarie per perseguire gli scopi dell'organizzazione stessa.

Per indicare la porzione automatizzata del sistema informativo di solito viene utilizzato il termina *sistema informatico*. Nei sistemi informatici le informazioni vengono rappresentate per mezzo di *dati*.

Una *base di dati* è una collezione di dati, utilizzati per rappressentare le informazioni di interesse per un sistema informativo.

1.2 Basi di dati e sistemi di gestione di basi di dati

Un *sistema di gestione di basi di dati* (DBMS) è un sistema software in grado di gestire collezioni di dati che siano

- *grandi*: in termini di occupazione di memoria
- *condivise*: applicazioni e utenti diversi devono poter accedervi
- *persistenti*: persistono anche dopo l'esecuzione del programma che le utilizza

assicurando la loro

- *affidabilità*: mantengono intatti i dati
- *privatezza*: mantengono sicuri e privati i dati

ed essendo

- *efficiente*: le operazioni vengono svolte rapidamente
- *efficace*: rendono produttive le attività dei loro utenti

1.3 Modelli dei dati

Un *modello di dati* è un insieme di concetti utilizzati per organizzare i dati di interesse e descriverne la struttura in modo che essa risulti comprensibile a un elaboratore.

Il *modello relazionale* dei dati permette di definire tipi per mezzo del costruttore *relazione*, che consente di organizzare i dati in insiemi di record a struttura fissa.

I *modelli concettuali* vengono utilizzati per descrivere i dati in maniera indipendente dal modello logico. Un tipo di modello concettuale è il modello *Entità-Relazione*.

1.3.1 Schemi e istanze

Nelle basi di dati esiste una parte sostanzialmente invariante nel tempo, detta *schema* della base di dati, costituita dalle caratteristiche dei dati, e una parte variabile nel tempo, detta *istanza* o *stato* della base di dati, costituita dai valori effettivi.

Lo schema di una relazione è costituito dalla sua intestazioen, cioè dal nome della relazione seguito dai nomi dei suoi attributi, ad esempio:

Docenza(Corso,NomeDocente)

L'*istanza di una relazione* è costituita dall'insieme, variante nel tempo, delle sue righe.

Capitolo 2

Il modello relazionale

2.1 Il modello relazionale: strutture

2.1.1 Relazioni e tabelle

Dati due insiemi D_1 e D_2, si chiama *prodotto cartesiano* di D_1 e D_2 l'insieme di coppie ordinate (v_1, v_2) tali che v_1 è un elemento di D_1 e v_2 è un elemento di D_2. Il numero n delle componenti del prodotto cartesiano viene detto *grado* del prodotto cartesiano e della relazione. Il numero degli elementi (n-uple) della relazione viene chiamato *cardinalità* della relazione.

2.1.2 Relazioni con attributi

Nelle basi di dati, ciascuna n-upla contiene dati fra loro collegati. Inoltre, una relazione è un insieme, quindi:

- non è definito alcun ordinamento fra le n-uple
- le n-uple di una relazione sono distinte l'una dall'altra, in quanto tra gli elementi di un insieme non ce ne possono essere presenti due uguali tra loro

Ciascuna n-upla è, al proprio interno, ordinata: l'i-esimo valore di ciascuna proviene dall'i-esimo dominio.

Indichiamo con D l'insieme dei domini e specifichiamo la corrispondeza tra attributi e domini per mezzo della funzione $dom : X \to D$, che associa a ciascun attributo $A \in X$ un dominio $dom(A) \in D$. Diciamo che una *tupla* su un insieme di attributi X è una funzione t che associa a ciascun attributo $A \in R$ un valore del dominio $dom(A)$. Una *relazione* su X è un insieme di tuple su X.

2.1.3 Relazioni e basi di dati

Uno *schema di relazione* è costituito da un simbolo R, detto *nome della relazione*, e da un insieme di *attributi* $X = \{A_1, A_2, \ldots A_n\}$, indicato con $R(X)$. A ciascun attributo è associato un dominio.

Uno *schema di base di dati* è un insieme di schemi di relazione con nomi diversi:

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

I nomi di relazione hanno come scopo principale quello di distinguere le varie relazioni nella base di dati.

Un *istanza di relazione* su uno schema $R(X)$ è un insieme r di tuple su X.

Un *istanza di base di dati* su uno schema $R = \{R_1(X_1), R_2(X_2), \ldots, R_n(X_n)\}$ è un insieme di relazioni dove ogni relazione è una relazione sullo schema $R_i(X_i)$.

2.2 Vincoli d'integrità

Il *vincolo d'integrità* è una proprietà che deve essere soddisfatta dalle istanze che rappresentano informazioni corrette per l'applicazione. Ogni vincolo può essere visto come un *predicato* che associa a ogni istanza il valore *vero* o *falso*. Se il predicato assume il valore vero, allora diciamo che l'istanza *soddisfa* il vincolo. Sono presenti deu categorie di vincoli:

- Un vincolo è *intrarelazionale* se il suo soddisfacimento è definito rispetto a singole relazioni della base di dati
  - un *vincolo di tupla* è un vincolo che può essere valutato su ciascuna tupla indipendentemente dalle altre
  - un vincolo definito con riferimento a singoli valori viene detto *vincolo su valori* o *vincolo di dominio*
- Un vincolo è *interrelazionale* se coinvolge più relazioni

2.2.1 Vincoli di tupla

I vincoli di tupla esprimono condizioni sui valori di ciascuna tupla, indipendentemente dalle altre tuple.

2.2.2 Chiavi

Una chiave è un insieme di attributi utilizzato per identificare univocamente le tuple di una relazione. Formalmente:

- un insieme K di attributi è *superchiave* di una relazione r se r non contiene due tuple distinte t_1 e t_2 con $t_1[K] = t_2[K]$
- K è *chiave* di r se è una superchiave minimale di r, cioè non esiste un'altra superchiave K' di r che sia contenuta in K come sottoinsieme proprio

Ciascuna relazione e ciascuno schema di relazione hanno sempre una chiave. Una relazione è un insieme e quindi è costituita da elementi fra loro diversi; di conseguenza, per ogni relazione $r(X)$, l'insieme X di tutti gli attributi su cui è definita è senz'altro una superchiave per essa. O tale insieme è anche chiave, nel qual caso si conferma l'esistenza della chiave stessa, oppure non è chiave, perchè esiste un'altra superchiave in esso contenuta.

Il fatto che su ciascuno schema di relazione possa essere definita almeno una chiave garantisce l'accessibilità a tutti i valori di una base di dati e la loro univoca identificabilità.

2.2.3 Chiavi e valori nulli

Su una delle chiavi, detta *chiave primaria* si vieta la presenza di valori nulli; sulle altre, i valori nulli sono generalmente ammessi.

2.2.4 Vincoli d'integrità referenziale

Un *vincolo d'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.

Se la chiave di R_2 è unica e composta da un solo attributo B, il vincolo di integrità referenziale fra l'attributo A di R_1 e la relazione R_2 è soddisfatto se, per ogni tupla t_1 in R_1 per cui $t_1[A]$ non è nullo, esiste una tupla t_2 in R_2 tale che $t_1[A_i] = t_2[B]$.

Nel caso generale, bisogna prestare attenzione al fatto che ciscuno degli attributi in X deve corrispondere a un preciso attributo della chiave primaria K di R_2. Allo scopo, è necessario specificare un ordinamento sia nell'insieme X sia in K. Indicando gli attributi in ordine, $X = A_1 A_2 \dots A_p$ e $K = B_1 B_2 \dots B_p$, il vincolo è soddisfatto se per ogni tupla t_1 in R_1 senza nulli su X esiste una tupla t_2 in R_2 con $t_1[A_i] = t_2[B_i]$, per ogni i compreso fra 1 e p.

Capitolo 3

Algebra e calcolo relazionale

3.1 Algebra relazionale

L'algebra relazionale è un linguaggio procedurale, basato su concetti di tipo algebrico. Esso è costituito da un insieme di operatori, definiti su relazioni e che producono ancora relazioni come risultati

3.1.1 Unione, intersezione, differenza

Le relazioni sono insiemi, quindi ha senso definire su di esse gli operatori insiemistici tradizionali di unione, differenza e intersezione:

- l'*unione* di due relazioni r_1 e r_2 definite sullo stesso insieme di attributi X è indicata con $r_1 \cup r_2$ ed è una relazione ancora su X contenente le tuple che appartengono a r_1 oppure a r_2, oppure ad entrambe

- l'*intersezione* di $r_1(X)$ e $r_2(X)$ è indicata con $r_1 \cap r_2$ ed è una relazione su X contenente le tuple che appartengono sia a r_1 sia a r_2

- la *differenza* di $r_1(X)$ e $r_2(X)$ è indicata con $r_1 - r_2$ ed è una relazione su X contenente le tuple che appartengono a r_1 e non appartengono ad r_2

3.1.2 Ridenominazione

L'operatore di *ridenominazione* cambia il nome degli attributi lasciando invariato il contenuto delle relazioni.

Sia r una relazione definita sull'insieme di attributi X e sia Y un altro insieme di attributi con la stessa cardinalità. Inoltre, siano $A_1A_2 \ldots A_k$ e $B_1B_2 \ldots B_k$ rispettivamente un ordinamento per gli attributi in X e un ordinamento per quello in Y. Allora la ridenominazione:

$$\rho_{B_1B_2 \ldots B_k \leftarrow A_1A_2 \ldots A_k}(r)$$

contiene una tupla t' per ciascuna tupla t in r, definita come segue: t' è una tupla su Y e $t'[B_i] = t[A_i]$, per $i = 1, \ldots, k$.

3.1.3 Selezione

La selezione produce un sottoinsieme di tuple su tutti gli attributi ("decomposizione orizzontale").

L'operatore è denotato dal simbolo σ, al pedice del quale viene indicata la "condizione di selezione". Il risultato contiene le tuple dell'operando che soddisfano la condizione di selezione. Le condizioni di selezione possono prevedere confronti fra attributi e confronti fra attributi e costanti, ottenute combinando condizioni semplici con i connettivi logici $\wedge$, $\vee$ e $\neg$.

3.1.4 Proiezione

La proiezione produce un risultato al quale contribuiscono tutte le tuple, ma su un sottoinsieme degli attributi ("decomposizione orizzontale").

3.1.5 Join

Esistono due versioni del join: il join naturale e il theta-join.

Join naturale

Il *join naturale* è un operatore che correla dati in relazioni diverse, sulla base di valori uguali in attributi con lo stesso nome. Il risultato del join è costituito da una relazione sull'unione degli insiemi di attributi degli operandi e le sue tuple sono ottenute combinando le tuple degli operandi con valori uguali sugli attributi comuni.

In generale, un join naturale $r_1 \bowtie r_2$ di $r_1(X_1)$ e $r_2(X_2)$ è una relazione definita su X_1X_2 come segue:

$$r_1 \bowtie r_2 = \{t \to X_1X_2 | t[X_1] \in r_1 \wedge t[X_2] \in r_2\}$$

Join completi ed incompleti

Il join di r_1 e r_2 contiene un numero di tuple compreso fra 0 e $|r_1| \times |r_2|$. Inoltre:

- se il join di r_1 e r_2 è completo, allora contiene almeno un numero di tuple pari al massimo tra $|r_1|$ e $|r_2|$
- se $X_1 \cap X_2$ contiene una chiave per r_2, allora il join di $r_1(X_1)$ e $r_2(X_2)$ contiene al più $|r_1|$ tuple
- se $X_1 \cap X_2$ coincide con una chiave per r_2 e sussiste il vincolo di riferimento fra $X_1 \cap X_2$ in r_1 e la chiave di r_2, allora il join $r_1(X_1)$ e $r_2(X_2)$ contiene esattamente $|r_1|$ tuple

Join esterni

Questa variante del join prevede che tutte le tuple diano un contributo al risultato, eventualmente estese con valori nulli ove non vi siano controparti opportune. Esistono tre varianti dell'operatore: il join esterno *sinistro*, che estende solo le tuple del primo operando, quello *destro*, che estende solo le tuple del secondo operando, e quello *completo* che le estende tutte.

Semijoin

Questo operatore restituisce le tuple di una relazione che partecipano al join naturale di tale relazione con un'altra. L'operatore può essere espresso per mezzo del join e della proiezione.

Join n-ario, intersezione e prodotto cartesiano

Il join naturale è associativo e commutativo. Se $X_1 = X_2$, il join coincide con l'intersezione. Se i set sono disgiunti, il risultato del join sarà in *prodotto cartesiano* delle due relazioni.

Theta-join e equi-join

Il *theta-join* è un prodotto cartesiano seguito da una selezione.

3.1.6 Interrogazioni in algebra relazionale

Dato uno schema R di base di dati, un'interrogazione è una funzione che, per ogni istanza r di R, produce una relazione su un dato insieme di attributi X.

Prendendo come esempio lo schema:

Impiegati(Matr, Nome, Età, Stipendio)
Supervisore(Capo, Impiegato)

si possono formulare le seguenti interrogazioni:

- trovare matricola, nome ed età degli impiegati che guadagnano più di 40 mila euro

$$\pi_{\text{Matr,Nome,Età}}(\sigma_{\text{Stipendio}>40}(\text{Impiegati}))$$

- trovare matricola e nome dei capi i cui impiegati guadagnano più di 40 mila euro

$$\pi_{\text{Matr,Nome}}((\text{Impiegati} \bowtie_{\text{Matr=Capo}}))(\pi_{\text{Capo}}(\text{Supervisione} - \pi_{\text{Capo}}(\text{Supervisione} \bowtie_{\text{Imp=Matr}} \sigma_{\text{Stip}\leq 40}(\text{Impiegati}))))$$

3.1.7 Viste

Possono esistere due tipi di relazioni derivate:

- *viste materializzate*: relazioni derivate effettivamente memorizzate nella base di dati
- *relazioni virtuali*: relazioni definite per mezzo di funzioni, non memorizzate nella base di dati, ma utilizzabili nelle interrogazioni come se lo fossero

3.2 Calcolo relazionale

Con il termine *calcolo relazionale* si fa riferimento a una famiglia di linguaggi d'interrogazione, basati sul calcolo dei predicati del primo ordine, che hanno la caratteristica di essere dichiarativi, cioè di specificare le proprietà del risultato delle interrogazioni.

3.2.1 Calcolo di tuple con dichiarazioni di range

Le espressioni hanno la forma

$$\{T|L|f\}$$

dove:

- T è la *target list*, con elementi del tipo $x.Z$, con x variabile e Z attributo
- L è la *range list*, che elenca le variabili libere della formula f con i relativi range
- f è una formula