

NORMALIZZAZIONE

La Normalizzazione è una attività che serve a verificare la complessità della Base di Dati rispetto a dei criteri per la presenza o meno di anomalie.

Si considera una tabella per codificare tutte le informazioni insieme.

Un database quindi composto da una sola tabella che rappresenta tutte le informazioni.

Ciò comporta delle caratteristiche di:

- ridondanza
- inefficienza
- complessità
- perdita di informazioni

Queste sono caratteristiche legate fra loro.

Si vogliono rappresentare le informazioni su alcuni prodotti: numero d'ordine, data dell'ordine, tipo di prodotto, poi nomi e indirizzi degli acquirenti...

Il tutto in una sola tabella quindi, in maniera assolutamente non elegante, ma possibile.

Questa soluzione è caratterizzata tutt'altro che da vantaggi.

- L'informazione è ridondante.
- Una ricerca può richiedere di scansionare tutte le righe dell'unica tabella.
- L'aggiornamento di un dato, tipo l'indirizzo di un compratore può richiedere la modifica di più righe.
- La cancellazione di un ordine, di una riga, può cancellare anche tutte le altre informazioni, tipo anche quelle del compratore se aveva fatto un solo ordine.

Per risolvere questi problemi è normale suddividere la tabella in più tabelle.

Per rappresentare i diversi livelli di qualità rispetto allo schema logico sono state introdotte delle FORME NORMALI.

Questo processo di verifica attuato tramite queste proprietà dette forme normali è detto appunto NORMALIZZAZIONE.

L'obiettivo tipico è arrivare almeno alla terza forma normale.

Per arrivare ad essa si deve passare la prima e la seconda forma normale.

Oltre la terza vi è la forma normale di Boyce – Codd, ma non tutte le relazioni possono essere convertite in modo da soddisfarla.

PRIMA FORMA NORMALE

Una tabella si dice in prima forma normale se e solo se ciascun attributo è definito su un dominio di valori atomici.

Vale a dire che se non c'è più di una colonna riferita allo stesso attributo allora è verificata la prima forma normale.

Si sostituisce gli attributi aggregati con un unico attributo e si aumentano le righe.

Ad esempio se si hanno 3 autori per un solo libro si fanno 3 righe dove compaiono i dati dello stesso libro con autore diverso, anziché avere una sola riga con 3 attributi ognuno per un autore diverso.

Le successive forme normali prevedono l'aumento delle tabelle, cioè le singole tabelle o la singola vengono decomposte in più tabelle.

Ciò deve essere fatto rispetto ad alcuni vincoli che fanno riferimento a una DIPENDENZA FUNZIONALE.

DIPENDENZA FUNZIONALE

Data una relazione $R(X)$ e due insiemi di attributi A e B non vuoti e sottoinsiemi di X si dice che B è funzionalmente dipendente da A se ogni combinazione di valori di A è associato ad una sola combinazione dei valori di B .

Se B dipende funzionalmente da A cioè $A \rightarrow B$, se due tuple hanno lo stesso valore di A devono avere lo stesso valore di B o gli stessi valori nel caso A o B siano composti da più attributi.

DIPENDENZE NON BANALI

La dipendenza funzionale $A \rightarrow B$ si dice non banale se $A \cap B$ è un insieme vuoto.

Data una dipendenza banale $A \rightarrow B$, se esiste un sottoinsieme di B non vuoto che non è incluso in A .

Per esempio una dipendenza non banale $AC \rightarrow BC$ può essere trasformata in $AC \rightarrow B$ che è non banale se $AC \cap B$ da insieme vuoto.

Sarebbe errore $AC \rightarrow BC \Rightarrow A \rightarrow B$.

Data una relazione $R(X)$ se K ne è una chiave ogni altro insieme di attributi è dipendente funzionalmente da essa e quindi si può dire che: il vincolo di dipendenza funzionale generalizza il vincolo di chiave.

In particolare se $A \rightarrow B$ su $R(X)$ degenera nel vincolo di chiave se $A \cup B = X$ allora A è superchiave per $R(X)$. (Si dimostra per assurdo.)

DECOMPOSIZIONE

La normalizzazione dopo la forma 1 si basa sulla decomposizione di una tabella in più tabelle che verificano le forme normali.

Una decomposizione di $R(X)$ è un insieme di relazioni $R_1 \dots R_n$ ottenute proiettando $R(X)$ su un insieme di attributi $A_1 \dots A_n$ tale che $A_1 \cup A_2 \dots \cup A_n = X$.

Non tutte le decomposizioni di una relazione $R(X)$ sono in grado di rappresentare tutte e sole le informazioni contenute in $R(X)$.

Dipende dalla scelta degli attributi $A_1 \dots A_n$.

Se si decompone una tabella in più tabelle se poi effettivamente con il JOIN si riottiene la tabella di partenza si dice LOSSLESS JOIN.

In caso contrario abbiamo delle perdite e la decomposizione è detta LOSSY-JOIN.

Si deve utilizzare ovviamente quella LOSSLESS.

Condizione sufficiente per non avere perdite è che avendo decomposto una relazione in più relazioni, l'intersezione degli attributi di esse deve dare una superchiave in almeno una delle tabelle.

PROPRIETA' DELLE DIPENDENZE FUNZIONALI

- riflessiva: se X include o è uguale a Y allora $X \rightarrow Y$

- arricchimento: se $X \rightarrow Y \Rightarrow XZ \rightarrow YZ$
- transitiva: se $X \rightarrow Y$ e $Y \rightarrow Z \Rightarrow X \rightarrow Z$
- proiezione: se $X \rightarrow YZ \Rightarrow X \rightarrow Y$
- unione: se $X \rightarrow Y$ e $X \rightarrow Z \Rightarrow X \rightarrow YZ$
- pseudo transitiva: se $X \rightarrow Y$ e $WY \rightarrow Z \Rightarrow WX \rightarrow Z$

Non tutte le decomposizioni rappresentano tutte le informazioni della tabella principale.

Si potrebbe avere tuple in più.

(vedere le slides)

SECONDA FORMA NORMALE

Data una dipendenza funzionale $A \rightarrow B$ si dice che B dipende completamente da A se per ogni sottoinsieme Y di A la dipendenza $(A - Y) \rightarrow B$ è falsa.

Data una dipendenza funzionale $A \rightarrow B$ si dice che B dipende partendo da A.

Un attributo si dice primo se è sottoinsieme di qualche chiave.

Una relazione si dice in seconda forma normale se è in prima forma normale e se ciascun attributo non primo è completamente dipendente dalla chiave primaria.

TERZA FORMA NORMALE

Una relazione R(X) è in terza forma normale se e solo se rispetta già la seconda forma normale e per ogni dipendenza funzionale $X \rightarrow A$, tra due insiemi di attributi A ed X, è vera una delle due seguenti condizioni:

- X è superchiave
- A è prime attribute (attributo primo)

FORMA DI BOYCE-CODD

Una relazione R(X) si dice in forma normale di Boyce-Codd se e solo se per ogni dipendenza funzionale non banale $X \rightarrow A$, X è una superchiave.