Appunti database modulo 4

Funzioni di Gruppo e Aggregazione in SQL

Ordinamento dei Risultati

L'ordinamento dei risultati rappresenta un'applicazione fondamentale nelle query SQL. Consideriamo un esempio pratico: quando si desidera selezionare nome, cognome e nazionalità dalla tabella RUNNINGCONTEST filtrando per il titolo "MARATHON" e ordinando per tempo di percorrenza, limitando i risultati ai primi tre, si ottengono gli atleti sul podio. In assenza dell'opzione DESC, l'ordinamento avviene dal basso verso l'alto, quindi vengono estratti i tre atleti con il tempo più basso.

Nell'ambito della programmazione Java, si utilizza quasi sempre la forma SELECT con l'asterisco per ricavare tutti i campi, ma si impostano frequentemente dei filtri per caricare dal database solo i dati necessari. Tipicamente si desidera anche specificare un ordine. Sebbene sia possibile ordinare e filtrare anche in Java, risulta meno efficiente rispetto a ricevere i dati già manipolati dal database.

Operazioni Riga per Riga

SQL offre funzioni equivalenti a map e filter, ma fino a questo punto non è ancora stata esplorata l'equivalente di reduce. Una volta selezionate m righe a partire da n, con m minore o uguale a n, è possibile mapparle per ottenere righe diverse dalle originali, ad esempio aggiungendo o togliendo colonne, oppure rinominandole. Tuttavia, non è ancora stato esaminato come ricavare risultati che riguardino più di una riga simultaneamente.

La visibilità in questa fase è limitata alla singola riga. Funzioni come YEAR, NOW e CONCAT, così come i normali operatori aritmetici quali addizione e sottrazione, sono definite scalari. Questo significa che prendono in ingresso i campi di una singola riga, operando appunto riga per riga.

Le Funzioni di Gruppo

Di contro, SQL offre anche funzioni di gruppo che lavorano su insiemi di righe, sintetizzandole attraverso operazioni di riduzione. Nella query per calcolare la somma dei salari delle persone residenti a Raccoon Ville, si sta effettivamente calcolando il prodotto interno lordo della città. Quando si calcola la media dei salari per i ricercatori, si ottiene il salario medio di un ricercatore. Per contare tutte le persone presenti nell'archivio, si utilizza la funzione COUNT con l'asterisco. Infine, per trovare il salario massimo e minimo delle persone nate nel 1980, si applicano le funzioni MAX e MIN filtrate per anno di nascita.

Risultati delle Funzioni di Gruppo

Mentre le funzioni di riga, o scalari, producono un risultato per ogni riga, le funzioni di gruppo producono un risultato per ogni gruppo, intendendo con gruppo un qualunque insieme di righe. Considerando ad esempio la query che conta il numero totale di persone nella tabella PERSON con alias n, questa produrrà un risultato con valore 3 dai dati di prova. Il gruppo su cui è stata applicata la riduzione è l'intera tabella Person. La funzione di gruppo count produce una sola riga di risultato a fronte di un gruppo con tre righe, e avrebbe prodotto una riga di risultato anche in presenza di duemila righe.

Funzioni di Gruppo ed Espressioni

Le funzioni SUM, AVG, COUNT, MAX e MIN sono funzioni di gruppo che operano non su una singola riga ma su gruppi di righe. Entrano insiemi di righe ed escono valori singoli, che rappresenta il funzionamento di base del reduce. È possibile utilizzare queste funzioni anche come parte di una espressione numerica. Ad esempio, per calcolare il divario salariale tra il più pagato e il meno pagato dei nati nel 1980, si sottrae il minimo dal massimo dei salari.

Analisi dell'Esempio Precedente

Partendo dalla query che calcola la differenza tra salario massimo e minimo per i nati nel 1980, si opera una selezione limitata ai nati in quell'anno, e sulle righe selezionate vengono eseguite due funzioni di gruppo. Questo restituisce una sola riga a partire dalle n righe dei nati nel 1980, con valore massimo 2500 e minimo 1200.

Sulla riga ottenuta si esegue successivamente una trasformazione aritmetica scalare, una proiezione sui dati di una sola riga, che produce il delta di 1300. In questo caso si è lavorato su un gruppo composto da una selezione di righe, successivamente compresso in una sola per fornire un unico risultato finale. Si tratta di una operazione di riduzione, o matematicamente parlando una funzione che trasforma un insieme di righe in un valore.

Creazione di Gruppi con GROUP BY

Supponiamo di voler calcolare lo stipendio medio per professione. Una possibilità consiste nell'eseguire query separate per ogni professione, calcolando la media dei salari per i ricercatori, poi per gli impiegati, e così via per ogni professione. Tuttavia, questa non rappresenta una vera soluzione, poiché richiederebbe di conoscere a priori tutte le professioni da tracciare, ed eseguire n queries, ciascuna delle quali restituirebbe una sola riga.

La soluzione corretta consiste nell'ordinare a SQL di raggruppare in base ai valori presenti nella tabella, e di applicare a ogni gruppo una o più funzioni di gruppo. Questo comportamento si ottiene tramite la clausola GROUP BY. Nella query corretta si seleziona il campo JOB insieme alla media dei salari, raggruppando per JOB.

Funzionamento di GROUP BY

L'istruzione GROUP BY per il campo job significa dividere la tabella in gruppi di righe con lo stesso valore per la colonna job. Una definizione alternativa potrebbe essere segmentare l'insieme originale in sottoinsiemi, o il gruppo in sottogruppi. È importante notare che GROUP BY non elimina righe, ma redistribuisce le righe presenti in gruppi separati e disgiunti.

Considerando dati ipotetici con tre ricercatori, due impiegati e un ufficiale nella tabella Person, il funzionamento può essere visualizzato attraverso le varie fasi. Partendo dai dati originali con identificativo, professione e salario, GROUP BY divide questi dati in gruppi in base al valore della colonna job, ottenendo tre gruppi distinti: uno per i ricercatori, uno per gli impiegati e uno per gli ufficiali.

Come precedentemente illustrato, ogni gruppo genera una riga di risultato che contiene il risultato delle funzioni di gruppo calcolate e, potenzialmente, i campi

per cui si è raggruppato, in questo caso job. Il gruppo Officer, pur contando di una sola riga, viene considerato comunque come gruppo e produce quindi un risultato. I risultati finali mostrano la professione e la media salariale per ciascun gruppo.

Raggruppamenti Multipli

È possibile raggruppare o partizionare per diversi campi o espressioni. Considerando una query che seleziona professione, genere e media dei salari raggruppando per entrambi i campi, il totale delle righe rimane sei, ma la partizionamento risulta più fine. I gruppi che prima erano semplicemente raggruppati per professione ora sono stati raggruppati per professione e genere, producendo potenzialmente sei gruppi derivanti dalla combinazione di tre professioni moltiplicate per due generi.

Alcuni gruppi potrebbero mancare dal risultato finale. Avendo un solo ufficiale, questo poteva essere maschio o femmina. Nel caso specifico era femmina, quindi il gruppo Officer maschio risultava vuoto ed è stato omesso dal risultato. I gruppi vuoti vengono infatti omessi dal risultato finale di una query con group by. Avrebbe potuto verificarsi l'assenza di ricercatori maschi, se tutti i ricercatori fossero stati donne. Una volta ottenuti i gruppi superstiti, si producono i risultati applicando a ogni gruppo le funzioni di gruppo desiderate.

Combinazione degli Strumenti

Capita frequentemente di utilizzare insieme raggruppamento, selezione, proiezione e ordinamento. In una query che calcola la media salariale per professione e genere, escludendo chi guadagna meno di mille euro e ordinando per professione, si stanno combinando tutte queste operazioni. L'esclusione dei salari inferiori a mille euro potrebbe creare dei gruppi vuoti. Si sta raggruppando per due campi, ordinando per uno, proiettando due campi e una funzione di gruppo.

L'ordine di esecuzione segue questa sequenza: selezione, partizionamento, riduzione, proiezione e ordinamento. Prima vengono eliminate le righe non desiderate, successivamente si creano i gruppi, si eseguono i calcoli e si proiettano i risultati desiderati. L'ordinamento viene eseguito per ultimo. Si tratta sempre di selezione sulla riga, almeno in questa fase, ma rimuovendo tutte le

righe si rimuove anche il gruppo, poiché un gruppo senza righe non produce risultato.

Proiezione e Funzioni di Gruppo

Quando si utilizzano le funzioni di gruppo, è possibile proiettare solo i campi per cui si è raggruppato, più le funzioni di gruppo o i calcoli su questi. Una query che seleziona due colonne, la media di una terza, la somma di una quarta, la media della somma delle prime due e la differenza tra il minimo della terza e il massimo della seconda, raggruppando per le prime due colonne, è corretta. È possibile prendere il minimo di una colonna anche senza raggruppare per essa.

Invece, aggiungere una terza colonna alla proiezione senza raggrupparla risulta errato. MySQL accetterà questa sintassi, ma altri sistemi di gestione di database, più rigorosi, non lo permettono. La ragione è semplice: non ha senso proiettare un campo che potrebbe cambiare da una riga all'altra dello stesso gruppo. Ogni gruppo genera una riga, e quella riga dovrebbe rappresentare il gruppo nel suo complesso. Non si desidera proiettare un campo che potrebbe variare all'interno dello stesso gruppo.

Selezione sui Gruppi con HAVING

Supponiamo di avere un volume di dati molto maggiore rispetto agli esempi precedenti. Potrebbero essere tutti i dati nazionali italiani, quindi circa sessanta milioni di righe, o forse anche solo trenta milioni non considerando bambini e persone fuori dal mercato del lavoro. Ipotizzando di voler calcolare il salario medio per città, la query risulta abbastanza semplice: si seleziona la città e la media dei salari dalla tabella PERSON, raggruppando per città.