Sử dụng GROUPING SETS trong thống kê dữ liệu

Trong SQL, ta sử dụng mệnh đề GROUP BY kết hợp với các hàm thống kê như SUM, AVG, MIN, MAX, COUNT để thực hiện các phép thống kê trên từng nhóm dữ liệu. Mỗi một truy vấn xác định cụ thể một nhóm dữ liệu cần thống kê. Trong nhiều trường hợp, chúng ta cần thực hiện truy vấn thống kê dựa trên nhiều nhóm dữ liệu khác nhau. Để làm điều này, chúng ta có thể sử dụng mệnh đề GROUPING SETS bên trong mệnh đề GROUP BY.

Để hình dung được mục đích và cách sử dụng mệnh để GROUPING SETS, chúng ta sẽ bắt đầu bởi các ví dụ dựa trên bảng NhanVien.

Tạo bảng NhanVien có cấu trúc như sau:

```
CREATE TABLE NhanVien
(
	MaNV nvarchar(10) primary key,
	HoTen nvarchar(50) not null,
	ChucVu nvarchar(50) not null,
	TienLuong money not null,
	DonVi nvarchar(50) not null
)
GO
```

Bổ sung dữ liệu cho bảng NhanVien:

```
INSERT INTO NhanVien(MaNV, HoTen, ChucVu, TienLuong, DonVi)
VALUES
('NV01', N'Lê Thanh An', N'Trưởng phòng', 3000, N'Nhân sự'),
('NV02', N'Mai Thị Hạnh', N'Nhân viên', 1000, N'Nhân sự'),
('NV03', N'Trần Văn Hữu', N'Phó phòng', 1200, N'Nhân sự'),
('NV04', N'Nguyễn Thị Hoa', N'Nhân viên', 800, N'Nhân sự'),
('NV05', N'Trần Chí Hiếu', N'Trưởng phòng', 2500, N'Kinh doanh'),
('NV06', N'Vũ Thanh Lê', N'Nhân viên', 500, N'Kinh doanh'),
```

```
('NV07', N'Hoàng Mai Hương', N'Nhân viên', 800, N'Kinh doanh'),
('NV08', N'Nguyễn Ngọc Tùng', N'Nhân viên', 1200, N'Kinh doanh'),
('NV09', N'Trần Thanh Toàn', N'Nhân viên', 1000, N'Kế toán'),
('NV10', N'Nguyễn Văn Bắc', N'Nhân viên', 800, N'Nhân sự'),
('NV11', N'Trần Thanh Phong', N'Nhân viên', 1000, N'Kinh doanh'),
('NV12', N'Vũ Thị Hoa', N'Nhân viên', 500, N'Nhân sự'),
('NV13', N'Lương Chí Hữu', N'Trưởng phòng', 2000, N'Kế toán'),
('NV14', N'Nguyễn Hữu Cảnh', N'Nhân viên', 800, N'Kế toán'),
('NV15', N'Nguyễn Quang Linh', N'Nhân viên', 850, N'Kinh doanh'),
('NV16', N'Nguyễn Hoàng Vũ', N'Phó phòng', 1500, N'Nhân sự'),
('NV17', N'Lương Thị Hải', N'Phó phòng', 1800, N'Kế toán'),
('NV18', N'Vũ Văn Vương', N'Phó phòng', 1600, N'Nhân sự'),
('NV19', N'Bạch Hải Châu', N'Phó phòng', 1700, N'Kế toán'),
('NV20', N'Nguyễn Văn Hùng', N'Phó phòng', 1400, N'Kinh doanh');
GO
```

Dữ liệu trong bảng trên như sau:

```
SELECT * FROM NhanVien
```

MaNV	HoTen	ChucVu	TienLuong	DonVi
NV01	Lê Thanh An	Trưởng phòng	3000.00	Nhân sự
NV02	Mai Thị Hạnh	Nhân viên	1000.00	Nhân sự
NV03	Trần Văn Hữu	Phó phòng	1200.00	Nhân sự
NV04	Nguyễn Thị Hoa	Nhân viên	800.00	Nhân sự
NV05	Trần Chí Hiếu	Trưởng phòng	2500.00	Kinh doanh
NV06	Vũ Thanh Lê	Nhân viên	500.00	Kinh doanh
NV07	Hoàng Mai Hương	Nhân viên	800.00	Kinh doanh
NV08	Nguyễn Ngọc Tùng	Nhân viên	1200.00	Kinh doanh
NV09	Trần Thanh Toàn	Nhân viên	1000.00	Kế toán
NV10	Nguyễn Văn Bắc	Nhân viên	800.00	Nhân sự
NV11	Trần Thanh Phong	Nhân viên	1000.00	Kinh doanh
NV12	Vũ Thị Hoa	Nhân viên	500.00	Nhân sự
NV13	Lương Chí Hữu	Trưởng phòng	2000.00	Kế toán
NV14	Nguyễn Hữu Cảnh	Nhân viên	800.00	Kế toán
NV15	Nguyễn Quang Linh	Nhân viên	850.00	Kinh doanh
NV16	Nguyễn Hoàng Vũ	Phó phòng	1500.00	Nhân sự
NV17	Lương Thị Hải	Phó phòng	1800.00	Kế toán
NV18	Vũ Văn Vương	Phó phòng	1600.00	Nhân sự
NV19	Bạch Hải Châu	Phó phòng	1700.00	Kế toán
NV20	Nguyễn Văn Hùng	Phó phòng	1400.00	Kinh doanh

Giả sử, chúng ta cần thống kê mức lương trung bình của các nhân viên theo các nhóm (tiêu chí) khác nhau như: theo đơn vị và chức vụ, theo đơn vị, theo chức vụ và trên toàn bộ nhân viên. Chúng ta có thể sử dụng các truy vấn thống kê bằng GROUP BY cho mỗi yêu cầu trên như sau:

- Thống kê lương trung bình của nhân viên theo đơn vị và chức vụ:

```
SELECT    DonVi, ChucVu, AVG(TienLuong) AS LuongTB
FROM    NhanVien
GROUP BY DonVi, ChucVu;
```

DonVi	ChucVu	LuongTB
Kế toán	Nhân viên	900.00
Kinh doanh	Nhân viên	870.00
Nhân sự	Nhân viên	775.00
Kế toán	Phó phòng	1750.00
Kinh doanh	Phó phòng	1400.00
Nhân sự	Phó phòng	1433.3333
Kế toán	Trưởng phòng	2000.00
Kinh doanh	Trưởng phòng	2500.00
Nhân sự	Trưởng phòng	3000.00

- Thống kê lương trung bình của nhân viên theo từng đơn vị:

```
SELECT  DonVi, AVG(TienLuong) AS LuongTB
FROM    NhanVien
GROUP BY DonVi;
```

DonVi	LuongTB
Kế toán	1460.00
Kinh doanh	1178.5714
Nhân sự	1300.00

- Thống kê lương trung bình của nhân viên theo từng chức vụ:

```
SELECT  ChucVu, AVG(TienLuong) AS LuongTB
FROM    NhanVien
GROUP BY ChucVu;
```

ChucVu	LuongTB
Nhân viên	840.909
Phó phòng	1533.3333
Trưởng phòng	2500.00

- Thống kê lương trung bình của tất cả nhân viên:

```
SELECT  AVG(TienLuong) AS LuongTB
```

```
FROM    NhanVien;
```

LuongTB
1297.50

Các kết quả của các câu lệnh trên là các kết quả rời rạc. Trong trường hợp chúng ta cần tổng hợp các kết quả trên vào một kết quả duy nhất, chúng ta có thể sử dụng lệnh UNION ALL để thực hiện phép hợp cho các câu lệnh ở trên. Khi đó, câu lệnh mà chúng ta sử dụng sẽ như sau:

```
SELECT  DonVi, ChucVu, AVG(TienLuong) AS LuongTB
FROM NhanVien GROUP BY DonVi, ChucVu
UNION ALL
SELECT  DonVi, NULL, AVG(TienLuong) AS LuongTB
FROM NhanVien GROUP BY DonVi
UNION ALL
SELECT  NULL, ChucVu, AVG(TienLuong) AS LuongTB
FROM NhanVien GROUP BY ChucVu
UNION ALL
SELECT  NULL, NULL, AVG(TienLuong) AS LuongTB
FROM NhanVien
ORDER BY DonVi, ChucVu;
```

Câu lệnh trên sẽ trả về một kết quả duy nhất cho tất cả các yêu cầu thống kê theo các nhóm tiêu chí khác nhau:

DonVi	ChucVu	LuongTB
NULL	NULL	1297.50
NULL	Nhân viên	840.909
NULL	Phó phòng	1533.3333
NULL	Trưởng phòng	2500.00
Kế toán	NULL	1460.00
Kế toán	Nhân viên	900.00
Kế toán	Phó phòng	1750.00
Kế toán	Trưởng phòng	2000.00
Kinh doanh	NULL	1178.5714
Kinh doanh	Nhân viên	870.00
Kinh doanh	Phó phòng	1400.00
Kinh doanh	Trưởng phòng	2500.00
Nhân sự	NULL	1300.00
Nhân sự	Nhân viên	775.00
Nhân sự	Phó phòng	1433.3333
Nhân sự	Trưởng phòng	3000.00

Có thể thấy, cách viết như ở trên tồn tại một số nhược điểm như sau:

- Câu lệnh phải viết khá dài.
- Tốc độ thực thi câu lệnh sẽ chậm vì phải thực thi nhiều câu lệnh (truy vấn con) và hợp các kết quả của các câu lệnh thành một kết quả.

Để khắc phục nhược điểm này, chúng ta có thể sử dụng mệnh đề GROUPING SETS trong mệnh đề GROUP BY. Khi đó, câu lệnh trên sẽ viết lại như sau:

```
SELECT  DonVi, ChucVu, AVG(TienLuong) AS LuongTB
FROM NhanVien
GROUP BY GROUPING SETS
(
    (DonVi, ChucVu),
    (DonVi),
    (ChucVu),
    ()
)
ORDER BY DonVi, ChucVu
```

Trong câu lệnh trên, chúng ta sử dụng mệnh đề GROUPING SETS để định nghĩa bốn nhóm dữ liệu cần thống kê là:

- `(DonVi, ChucVu)` : Nhóm theo đơn vị và chức vụ
- `(DonVi)` : Nhóm theo đơn vị
- `(ChucVu)` : Nhóm theo đơn vị
- `()` : Thống kê trên toàn dữ liệu

Kết quả của câu lệnh ở trên tương tự như cách chúng ta sử dụng phép hợp trên các câu lệnh rời rạc. Tuy nhiên, câu lệnh được viết theo cách này ngắn gọn, dễ đọc và hiệu quả hơn so với việc phải sử dụng phép hợp.

Hàm GROUPING

Hàm GROUPING trả về giá trị 0 hoặc 1, được sử dụng để xác định xem một cột xuất hiện trong mệnh đề GROUP BY có phải là cột được dùng để nhóm (tổng hợp) dữ liệu hay không.

Ví dụ, câu lệnh dưới đây:

```
SELECT GROUPING(DonVi) AS Group_DonVi,
       GROUPING(ChucVu) AS Group_ChucVu,
       DonVi, ChucVu, AVG(TienLuong) AS LuongTB
FROM NhanVien
GROUP BY GROUPING SETS
```

```
    (
        (DonVi, ChucVu),
        (DonVi),
        (ChucVu),
        ()
    )
```

Có kết quả như sau:

Group_DonVi	Group_ChucVu	DonVi	ChucVu	LuongTB
1	1	NULL	NULL	1297.50
1	0	NULL	Nhân viên	840.909
1	0	NULL	Phó phòng	1533.3333
1	0	NULL	Trưởng phòng	2500.00
0	1	Kế toán	NULL	1460.00
0	0	Kế toán	Nhân viên	900.00
0	0	Kế toán	Phó phòng	1750.00
0	0	Kế toán	Trưởng phòng	2000.00
0	1	Kinh doanh	NULL	1178.5714
0	0	Kinh doanh	Nhân viên	870.00
0	0	Kinh doanh	Phó phòng	1400.00
0	0	Kinh doanh	Trưởng phòng	2500.00
0	1	Nhân sự	NULL	1300.00
0	0	Nhân sự	Nhân viên	775.00
0	0	Nhân sự	Phó phòng	1433.3333
0	0	Nhân sự	Trưởng phòng	3000.00

Như chúng ta thấy được ở kết quả ở trên, kết quả của hàm GROUPING sẽ giúp cho chúng ta hiểu rõ hơn việc tính toán giá trị thống kê được thực hiện trên các nhóm dữ liệu như thế nào.

Sử dụng WITH ROLLUP và WITH CUBE

Như đã đề cập đến ở trên, sử dụng mệnh đề GROUPING SETS cho phép ta tạo ra những truy vấn thống kê đa chiều, với các chiều cần thống kê được chỉ định một cách tường minh bên trong mệnh đề này. Trong một số trường hợp, thay vì phải mô tả tường minh các chiều dữ liệu cần thống kê, SQL cho phép chúng ta sử dụng hai tùy chọn WITH ROLLUP và WITH CUBE trong mệnh đề GROUP BY nhằm đơn giản hóa cách thể hiện các chiều cần thống kê trong câu lệnh.

WITH ROLLUP

WITH ROLLUP được sử dụng nhằm tạo ra các chiều thống kê dữ liệu thông qua việc tạo ra các nhóm con dựa trên các cột được chỉ định trong mệnh đề GROUP BY theo cú pháp:

```
GROUP BY cột_1, cột_2,…, cột_N-1, cột_N WITH ROLLUP
```

Các chiều thống kê được tạo ra bởi WITH ROLLUP bắt đầu từ nhóm bao gồm tất cả các cột được chỉ định sau GROUP BY, các nhóm tiếp theo có được bằng cách loại bỏ dần từng cột theo thứ tự từ phải qua trái cho đến khi không còn cột nào. Hiểu một cách đơn giản, cú pháp ở trên tương đương với cách viết sử dụng GROUPING SETS như sau:

```
GROUPING SETS
(
        (cột_1, cột_2,…, cột_N-1, cột_N),
        (cột_1, cột_2,…, cột_N-1),
        …
        (cột_1, cột_2),
        (cột_1),
        ()
)
```

WITH CUBE

Tương tự như WITH ROLLUP, tùy chọn WITH CUBE cũng được sử dụng để tạo ra các chiều thống kê dữ liệu. Tuy nhiên, các chiều thống kê được tạo ra bởi WITH CUBE là tất cả các tổ hợp có thể có của các nhóm từ các cột được chỉ định trong mệnh đề GROUP BY, bao gồm cả các nhóm con và các nhóm không liên quan trực tiếp đến nhau. Thống kê đa chiều với WITH CUBE cung cấp cho ta một cái nhìn tổng quan toàn diện hơn về dữ liệu, bao gồm số liệu thống kê của mọi tổ hợp có thể có.

Lựa chọn WITH ROLLUP và WITH CUBE khi nào?

WITH ROLLUP và WITH CUBE là hai tùy chọn mở rộng cho mệnh đề GROUP BY trong SQL, được sử dụng để tạo ra các báo cáo tổng hợp với nhiều cấp độ phân nhóm và các hàm tổng hợp. Cả hai đều được sử dụng để tạo ra các kết quả tổng hợp từ dữ liệu được nhóm theo một hoặc nhiều cột, hay còn gọi là thống kê đa chiều. Tuy nhiên, tùy vào mục đích thống kê mà sử dụng chúng cho phù hợp:

- Sử dụng WITH ROLLUP khi ta muốn tập trung vào một phân cấp cụ thể của dữ liệu và xem số liệu tổng hợp theo từng cấp độ.
- Sử dụng WITH CUBE khi ta muốn có một cái nhìn tổng quan toàn diện về dữ liệu và xem số liệu tổng hợp cho mọi tổ hợp có thể có của các cột được dùng để nhóm.