

Aggregácia v SQL

SQL a agregácia

- niekedy chceme namiesto vypísania zoznamu riadkov radšej zistiť ich počet / súčet a pod.
 - na to slúžia tzv. agregáčné funkcie:
 - Sum, Min, Max, Avg, Stdev, Count, ...
 - <https://www.postgresql.org/docs/current/static/functions-aggregate.html>
- niekedy chceme riadky zoskupiť podľa nejakého atribútu
 - napr. chceme zoskupiť zamestnancov podľa ich oddelenia a pod.
 - alebo chceme počet zamestnancov na jednotlivých oddeleniach

GROUP BY

- **SELECT** <zoznam atribútov>
FROM <zoznam relácií>
WHERE <podmienka>
GROUP BY <zoznam atribútov>
HAVING <podmienka>
- GROUP BY zoskupí riadky s rovnakou hodnotou v uvedených atribútoch
- pre každú skupinu bude vo výstupe 1 riadok
- zoznam atribútov za SELECT môže obsahovať len atribúty uvedené za GROUP BY a agreg. funkcie
 - Toto nie je celkom pravda pre všetky databázové systémy (napr. MySQL takúto reštrikciu nemá)
- podmienka v HAVING môže obsahovať agregáčnne funkcie, zatiaľ čo za WHERE nemôže

GROUP BY príklad

- **SELECT** deptno, COUNT(*) as c
FROM emp
WHERE sal > 3000
GROUP BY deptno
HAVING COUNT(*)>=2
- všimnite si, že COUNT(*) musíme napísať dvakrát

GROUP BY príklad

name	deptno	sal
John	10	1000
Thomas	10	3100
George	10	3200
Lucas	20	3100
Bob	20	2050
Joe	30	1000
Francis	30	3050
Hugo	40	1000
Mike	40	5000
Robert	40	2900
Anna	50	8000

WHERE

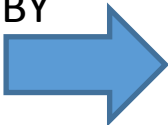


name	deptno	sal
George	10	3200
Thomas	10	3100
Lucas	20	3100
Francis	30	3050
Mike	40	5000
Anna	50	8000

```
SELECT deptno, COUNT(*)  
FROM emp  
WHERE sal > 3000  
GROUP BY deptno  
HAVING COUNT(*) >= 2
```

GROUP BY príklad

name	deptno	sal
George	10	3200
Thomas	10	3100
Lucas	20	3100
Francis	30	3050
Mike	40	5000
Anna	50	8000

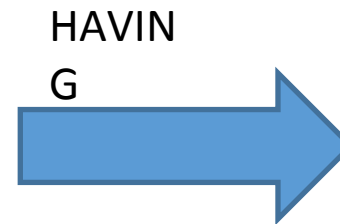
GROUP
BY


deptno	COUNT(*)	Name	deptno	sal
10	2			
		George	10	3200
		Thomas	10	3100
20	1			
		Lucas	20	3100
30	1			
		Francis	30	3050
40	1			
		Mike	40	5000
50	1			
		Anna	50	8000

```
SELECT deptno, COUNT(*)  
FROM emp  
WHERE sal > 3000  
GROUP BY deptno  
HAVING COUNT(*)>=2
```

GROUP BY príklad

deptno	count(*)	Name	deptn	sal
10	2			
		George	10	3200
		Thomas	10	3100
20	1			
		Lucas	20	3100
30	1			
		Francis	30	3050
40	1			
		Mike	40	5000
50	1			
		Anna	50	8000



deptno	count(*)
10	2

```
SELECT deptno, COUNT(*)  
FROM emp  
WHERE sal > 3000  
GROUP BY deptno  
HAVING COUNT(*)>=2
```

Ďalšie aspekty

- zoznam atribútov za SELECT môže obsahovať len atribúty uvedené za GROUP BY a agregáčn  funkcie
- toto je pre program torov trochu otrava:
 - student(StudentID, Meno, Priezvisko, TriedaID)
trieda(TriedaID, Nazov)
 - **SELECT s.triedaid, t.nazov, COUNT(*)**
 - **FROM student as s, trieda as t**
 - **WHERE s.triedaid = t.triedaid GROUP BY s.triedaid, t.nazov**
 - TriedaID jednozna ne ur uje n zov triedy, no program tor to mus  zbyto ne zap sať 2x

Ďalšie aspekty

- MySQL:

V SELECT časti môžem použiť akýkoľvek atribút. Ak je z množiny atribútov, ktoré nie sú v GROUP BY, vyberie sa náhodný prvok zo skupiny

- PostgreSQL:

When GROUP BY is present, or any aggregate functions are present, it is not valid for the SELECT list expressions to refer to ungrouped columns except within aggregate functions or when the ungrouped column is functionally dependent on the grouped columns, since there would otherwise be more than one possible value to return for an ungrouped column. A functional dependency exists if the grouped columns (or a subset thereof) are the primary key of the table containing the ungrouped column.

Ďalšie aspekty

- Ak SELECT obsahuje agregáčnú funkciu, ale bez GROUP BY
 - potom všetky riadky akoby boli zaradené do jednej skupiny – výstupom je riadok, napr.
 - SELECT COUNT(*) FROM emp;
 - SELECT MAX(sal) FROM emp;
 - podobne ak uvedieme HAVING bez GROUP BY
 - (tomuto sa vyhnite, je to mätúce a na rozdiel od predošlého neužitočné)

Ďalšie aspekty

- pozor na NULL pri agregáčnych funkciách
- pri niektorých agregáčnych funkciách sa vynechávajú riadky s NULL
- napr. SUM pre $1 + \text{NULL}$ je 1, pritom inokedy výraz $1 + \text{NULL}$ má hodnotu NULL
- navyše ak všetky riadky sú NULL, tak výsledok je NULL

Ďalšie aspekty

- ak potrebujeme nájsť hodnoty, pre ktoré sa dosahuje napr. maximum (arg max), nedá sa to zapísať v SQL jedným dotazom, treba vnorený dotaz:

SELECT name

FROM emp

WHERE salary = (SELECT MAX(salary) FROM emp);

- databázový systém automaticky konvertuje výsledok vnoreného selektu (reláciu s 1 stĺpcom a 1 riadkom) na číslo
- vyskúšajte, čo sa stane, ak relácia emp neobsahuje žiadne záznamy