

Kapitola 17. Seskupování pod lupou

V úvodní kapitole této části knihy jste si trochu mohli připomenout seskupování pomocí `GROUP BY`. Seskupování však mnohdy není úplně triviální, a tak bude dobré se na něj podívat trochu podrobněji. Přesně to provedeme nyní.

Úvodní úloha

Na začátku potřebujeme vytvořit rámeček, v němž budeme seskupování zkoumat.

ZADÁNÍ

Vypište, kolik je od každého titulu v knihovně exemplářů.

UPŘESNĚNÍ

Zobrazujte vždy název titulu a k němu počet exemplářů.

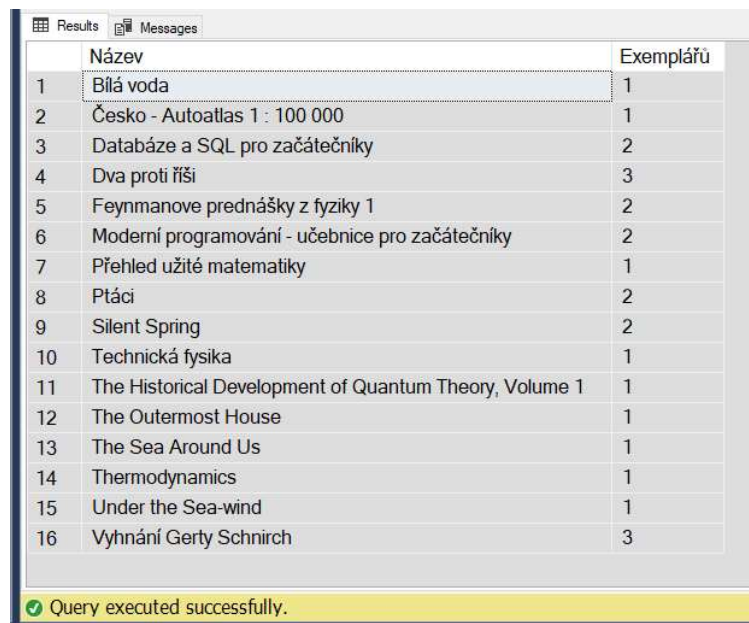
NÁVRH ŘEŠENÍ

Exempláře budeme seskupovat podle jejich příslušnosti k danému titulu. Pro každou skupinu pak vypočteme `COUNT (*)`. V zadání požadovaný název titulu je zase v tabulce titulů. `Select` tedy bude muset pracovat nad spojenými tabulkami titulů a exemplářů:

```
SELECT
    Tit.Název,
    COUNT(*) AS [Exemplářů]
FROM
    [dbo].[Tituly] AS Tit
    JOIN [dbo].[Exempláře] AS Ex ON Ex.IdTitulu = Tit.Id
GROUP BY
    Tit.Název
ORDER BY
    Tit.Název;
```

VÝSLEDKY

Odesláním Selectu na server jsem získal tyto výsledky:



	Název	Exemplářů
1	Bílá voda	1
2	Česko - Autoatlas 1 : 100 000	1
3	Databáze a SQL pro začátečníky	2
4	Dva proti říši	3
5	Feynmanove prednášky z fyziky 1	2
6	Moderní programování - učebnice pro začátečníky	2
7	Přehled užité matematiky	1
8	Ptáci	2
9	Silent Spring	2
10	Technická fyzika	1
11	The Historical Development of Quantum Theory, Volume 1	1
12	The Outermost House	1
13	The Sea Around Us	1
14	Thermodynamics	1
15	Under the Sea-wind	1
16	Vyhnání Gerty Schnirch	3

Query executed successfully.

Namátkově zkontrolujeme:

- *Sea Around Us* má Id titulu 26. V tabulce exemplářů je jeden záznam (Id=9) s cizím klíčem IdTitulu na hodnotě 26. **Souhlasí.**
- *Gerta* má Id titulu 32. V tabulce exemplářů vidím pro IdTitulu=32 celkem tři záznamy (Id 13, 23, 25). **Souhlasí.**

VÝSLEDEK K DISKUSI

Otázkou je ale výsledek pro *Dva proti říši*. V databázi máme první a třetí vydání této knihy, tituly s Id 34 a 36. První vydání je ve dvou exemplářích, třetí vydání v jednom. Celkem tedy tři exempláře, což souhlasí s naším výsledkem, a nejspíš je to přesně to, co knihovník potřeboval vědět - že kniha o největším dramatu na Vychovatelně je v knihovně třikrát.

Na druhou stranu naši databázi knihovny máme koncipovanu tak, že **každé vydání jedné knihy chápeme jako úplně nezávislou knihu**. Nikde v databázi nezaznamenáváme, že ten a ten titul spolu souvisejí, že jde o tutéž knihu. Náš Select získal diskutovaný výsledek **díky shodě názvů**, nikoli díky nějakému propojení v databázi.

Je tedy otázka, zda to je správně. „Stejná kniha“ se totiž nemusí vždy poznat stejným názvem. Na jedné straně dva autoři mohli nezávisle napsat knihu se stejným názvem. Dívám se např. do Souborného katalogu ČR a namátkou vidím dvě různé knihy, které se jmenují přesně *Člověk a vesmír*.

Na druhé straně tatáž kniha v různých vydáních může nést rozdílné názvy. Například u druhého vydání mé knihy *C# Programming for Absolute Beginners* se nakladatel rozhodl rozšířit název na *C# Programming for Absolute Beginners: Learn to Think Like a Programmer and Start Writing Code*.

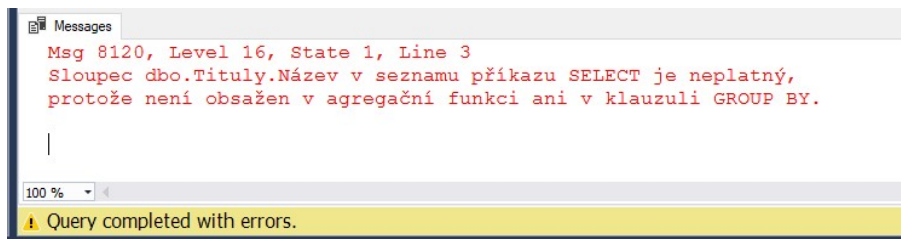
Z hlediska toho, jak jsme naši databázi navrhli, tedy uvedený Select asi úplně správný není. Tímto závěrem se budeme v dalším řídit.

SESKUPOVÁNÍ PODLE ID

Chápeme-li každý titul jako zcela nezávislý na ostatních, musíme seskupovat podle jedinečného Id, nikoli podle potenciálně duplicitního názvu. Změníme tedy GROUP BY (místo Tit.Název dáme Tit.Id):

```
SELECT
    Tit.Název,
    COUNT(*) AS [Exemplářů]
FROM
    [dbo].[Tituly] AS Tit
    JOIN [dbo].[Exempláře] AS Ex ON Ex.IdTitulu = Tit.Id
GROUP BY
    Tit.Id
ORDER BY
    Tit.Název;
```

No, serveru se to nelíbí:



The screenshot shows a 'Messages' window with the following error text:

```
Msg 8120, Level 16, State 1, Line 3
Sloupec dbo.Tituly.Název v seznamu příkazu SELECT je neplatný,
protože není obsažen v agregační funkci ani v klauzuli GROUP BY.
```

At the bottom of the window, a yellow status bar reads: "Query completed with errors."

Jak funguje seskupování

Abychom to vyřešili, přiblížím vám, jak seskupování vlastně funguje.

Uvažujme naše seskupování:

```
...
FROM
  [dbo].[Tituly] AS Tit
  JOIN [dbo].[Exempláře] AS Ex ON Ex.IdTitulu = Tit.Id
GROUP BY
  Tit.Id
...
```

1. SPOJENÍ TABULEK

Databázový server nejprve provede spojení tabulek:

Tit.Id	Tit.Název	Ex.Id	Ex.ČarovýKód	Další atributy
1	Databáze a SQL pro začátečníky	1	0000-0001	...
9	Moderní programování - učebnice pro začátečníky	2	0000-0002	...
15	Přehled užité matematiky	3	0000-0003	...
16	Feynmanove prednášky z fyziky 1	4	0000-0004	...
22	Technická fyzika	5	0000-0005	...
23	The Historical Development of Quantum Theory, Vo...	6	0000-0006	...
24	Thermodynamics	7	0000-0007	...
25	Silent Spring	8	0000-0008	...
26	The Sea Around Us	9	0000-0009	...
28	Under the Sea-wind	10	0000-0010	...
29	The Outermost House	11	0000-0011	...
31	Ptáci	12	0000-0012	...
32	Vyhánění Gerty Schnirch	13	0000-0013	...
33	Bílá voda	14	0000-0014	...
34	Dva proti třetí	15	0000-0015	...
35	Česko - Autoatlas 1 : 100 000	16	0000-0016	...
36	Dva proti třetí	17	0000-0017	...
37	Silent Spring	18	0000-0018	...
1	Databáze a SQL pro začátečníky	19	0000-0019	...
9	Moderní programování - učebnice pro začátečníky	20	0000-0020	...
16	Feynmanove prednášky z fyziky 1	21	0000-0021	...
31	Ptáci	22	0000-0022	...
32	Vyhánění Gerty Schnirch	23	0000-0023	...
34	Dva proti třetí	24	0000-0024	...
32	Vyhánění Gerty Schnirch	25	0000-0025	...

V každém řádku spojení jsou k dispozici všechny atributy obou tabulek, to znáte.

2. SEŘAZENÍ A SESKUPENÍ

Nyní server řádky spojených tabulek seřadí podle `Tit.Id` a za každým `Tit.Id` „udělá tlustou čáru“:

Tit.Id	Tit.Název	Ex.Id	Ex.ČárovýKód	Další atributy
1	Databáze a SQL pro začátečníky	1	0000-0001	...
1	Databáze a SQL pro začátečníky	19	0000-0019	...
9	Moderní programování - učebnice pro začátečníky	2	0000-0002	...
9	Moderní programování - učebnice pro začátečníky	20	0000-0020	...
15	Přehled užití matematiky	3	0000-0003	...
16	Feynmanove prednášky z fyziky 1	4	0000-0004	...
16	Feynmanove prednášky z fyziky 1	21	0000-0021	...
22	Technická fyzika	5	0000-0005	...
23	The Historical Development of Quantum Theory, Vo...	6	0000-0006	...
24	Thermodynamics	7	0000-0007	...
25	Silent Spring	8	0000-0008	...
26	The Sea Around Us	9	0000-0009	...
28	Under the Sea-wind	10	0000-0010	...
29	The Outermost House	11	0000-0011	...
31	Ptáci	12	0000-0012	...
31	Ptáci	22	0000-0022	...
32	Vyhánání Gerty Schnirch	13	0000-0013	...
32	Vyhánání Gerty Schnirch	23	0000-0023	...
32	Vyhánání Gerty Schnirch	25	0000-0025	...
33	Bílá voda	14	0000-0014	...
34	Dva proti říši	15	0000-0015	...
34	Dva proti říši	24	0000-0024	...
35	Česko - Autoatlas 1 : 100 000	16	0000-0016	...
36	Dva proti říši	17	0000-0017	...
37	Silent Spring	18	0000-0018	...

3. VÝSTUP

`GROUP BY` nyní způsobí, že se na výstup dostane **vždy jeden řádek souhrnně za každou skupinu**. Klíčové přitom je, že do výstupu již nemůžete použít jakýkoli atribut spojených tabulek, neboť **jednotlivé záznamy ve skupině se mohou v tom atributu lišit!**

V seznamu výstupních sloupců za slovem `SELECT` (a také ve sloupcích určujících řazení za slovy `ORDER BY`) máte tedy dovoleno použít pouze hodnoty, které mají smysl **pro celou skupinu**, což znamená:

- Charakteristiku skupiny, což v našem případě je `Tit.Id`;
- Nějakou „funkci“ předchozího;
- Hodnotu agregovanou za celou skupinu, typicky výsledek agregačních funkcí `COUNT`, `SUM`, `MIN`, `MAX`, popř. dalších.

Druhou odrážku pomínu, abych neunavoval přílišným detailem, a podíváme se, jak tedy celé řešení správně zapsat.

Dvě správná řešení

Víme, s čím lze pracovat na výstupu. Ještě to zkontrolujeme.

KONTROLA

Když namísto `Tit.Název` všude napíšete `Tit.Id`, server nebude namítat, neboť se odkazujete jen na hodnoty, které mají smysl pro celou skupinu.

```
SELECT
    Tit.Id AS [Id titulu],
    COUNT(*) AS [Exemplářů]
FROM
    [dbo].[Tituly] AS Tit
    JOIN [dbo].[Exempláře] AS Ex ON Ex.IdTitulu = Tit.Id
GROUP BY
    Tit.Id
ORDER BY
    Tit.Id;
```

Samozřejmě, výsledek není, to co chceme. Šlo o zkoušku. K úplnému řešení přistoupíme nyní.

PRVNÍ ŘEŠENÍ

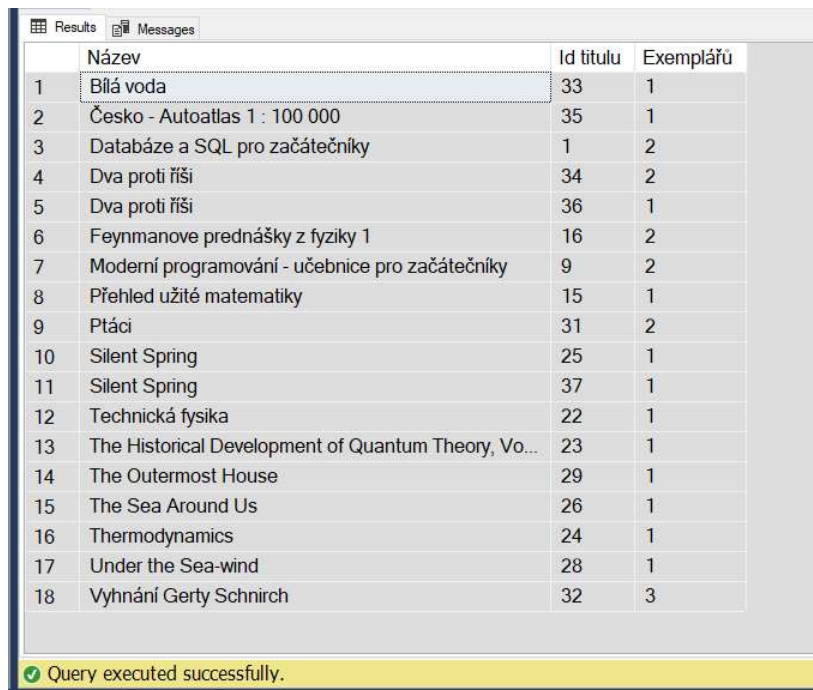
Budeme seskupovat jak podle `Id` titulu, tak podle jeho názvu. Dodatečné seskupovací kritérium podle názvu sice skupiny podle `Id` nezjmenuje - každé `Id` má vždy jeden název - ale vyhovíme tím pravidlu pro použití hodnot, které mají smysl pro celou skupinu.

```
SELECT
    Tit.Název,
    Tit.Id AS [Id titulu],
    COUNT(*) AS [Exemplářů]
FROM
    [dbo].[Tituly] AS Tit
    JOIN [dbo].[Exempláře] AS Ex ON Ex.IdTitulu = Tit.Id
GROUP BY
    Tit.Id,
    Tit.Název
ORDER BY
    Tit.Název,
    Tit.Id;
```

Kromě rozšířeného seskupování jsem provedl ještě jednu změnu - přidal jsem výstupní sloupec `Tit.Id`. Ten jsem nemusel přidávat kvůli fungování `GROUP BY`. Přidal jsem jej, aby výstup dával uživateli větší smysl. Když nyní bude mít *Dva proti říši* ve dvou řádcích, tak ať je dokáže rozlišit.

VÝSLEDKY

Výsledky jsou nyní takové, jak očekáváme. Každý titul na samostatném řádku, ve výpisu název titulu:



	Název	Id titulu	Exemplářů
1	Bílá voda	33	1
2	Česko - Autoatlas 1 : 100 000	35	1
3	Databáze a SQL pro začátečníky	1	2
4	Dva proti říši	34	2
5	Dva proti říši	36	1
6	Feynmanove prednášky z fyziky 1	16	2
7	Moderní programování - učebnice pro začátečníky	9	2
8	Přehled užití matematiky	15	1
9	Ptáci	31	2
10	Silent Spring	25	1
11	Silent Spring	37	1
12	Technická fyzika	22	1
13	The Historical Development of Quantum Theory, Vo...	23	1
14	The Outermost House	29	1
15	The Sea Around Us	26	1
16	Thermodynamics	24	1
17	Under the Sea-wind	28	1
18	Vyhánění Gerty Schnirch	32	3

Query executed successfully.

DRUHÉ ŘEŠENÍ

Úlohu lze řešit také alternativním způsobem, trochu trikem. Namísto (formálního) zjemnění skupin pomocí druhého seskupovacího kritéria použijeme na název titulu agregační funkci `MIN`!

Funkce `MIN` v každé skupině najde „minimální“ - tedy první v abecedě - název. I v tomto případě jde o svým způsobem formální řešení, abychom vyhověli pravidlům. Všechny názvy titulu ve skupině jsou stejné, jde o jeden a tentýž titul, takže `MIN` nevrátí nic jiného než prostě název toho titulu. Server však bude mít jistotu, že jde o hodnotu za celou skupinu.

V následujícím výpisu najdete použití funkce `MIN` jak v seznamu výstupních sloupců za slovem `SELECT`, tak mezi řadicími kritérii za slovy `ORDER BY`:

```
SELECT
  MIN(Tit.Název) AS [Název],
  Tit.Id AS [Id titulu],
  COUNT(*) AS [Exemplářů]
FROM
  [dbo].[Tituly] AS Tit
  JOIN [dbo].[Exempláře] AS Ex ON Ex.IdTitulu = Tit.Id
GROUP BY
  Tit.Id
ORDER BY
  MIN(Tit.Název),
  Tit.Id;
```

Zkontrolujte, že dostanete výsledky shodné s prvním řešením.

Tituly bez exemplářů

V naší úloze na počítání exemplářů je ještě jedna otázka, kterou jsem nepřesnoval v zadání a kterou jsem zatím nediskutoval, abych neodváděl pozornost od průzkumu principu seskupování. Konkrétně jde o to, zda se mají či nemají ve výstupu objevovat knihy bez exemplářů. V našem případě se jedná o titul *Myslete databázově, myslíte v SQL!* s `Id=2`. Zatím jsme uvažovali, že nemají, nyní se podíváme, jak to udělat, kdybychom je zahrnout chtěli.

ZADÁNÍ

Vypište počty exemplářů jednotlivých titulů. Do výstupu zahrňte rovněž tituly, které v databázi nemají žádné přidružené exempláře.

ŘEŠENÍ

Pro řešení pozměněného zadání proveďte tyto úpravy:

- Namísto `JOINU` použijte `LEFT JOIN`, aby se mezi řádky spojených tabulek dostaly také ty, kde záznamu titulu neodpovídá žádný záznam exempláře.
- Namísto `COUNT(*)` použijte `COUNT(Ex.Id)`. Ve skupině nechceme počet všech řádků, ale jen počet řádků napárovaných na exempláře. U titulu bez exempláře budou všechny atributy tabulky `Ex` na hodnotě `NULL` a funkce `COUNT(Ex.Id)` je - správně - nezapočítá.


```

SELECT
    Tit.Název,
    Tit.Id AS [Id titulu],
    COUNT(Ex.Id) AS [Exemplářů]
FROM
    [dbo].[Tituly] AS Tit
LEFT JOIN [dbo].[Exempláře] AS Ex ON Ex.IdTitulu = Tit.Id
GROUP BY
    Tit.Id,
    Tit.Název
ORDER BY
    Tit.Název,
    Tit.Id;

```

VÝSLEDKY

Ověřte, že ve výsledkové sadě přibyl řádek navíc:

	Název	Id titulu	Exemplářů
1	Bílá voda	33	1
2	Česko - Autoatlas 1 : 100 000	35	1
3	Databáze a SQL pro začátečníky	1	2
4	Dva proti říši	34	2
5	Dva proti říši	36	1
6	Feynmanove prednášky z fyziky 1	16	2
7	Moderní programování - učebnice pro začátečníky	9	2
8	Myslete databázově, myslete v SQL!	2	0
9	Přehled užití matematiky	15	1
10	Ptáci	31	2
11	Silent Spring	25	1
12	Silent Spring	37	1
13	Technická fyzika	22	1
14	The Historical Development of Quantum Theory, Vo...	23	1
15	The Outermost House	29	1
16	The Sea Around Us	26	1
17	Thermodynamics	24	1
18	Under the Sea-wind	28	1
19	Vyhánění Gerty Schnirch	32	3

Query executed successfully.