

# Analytické dotazy a materializované pohledy

## Cíl

Cílem tohoto projektu je představit podporu materializovaných pohledů a analytických dotazů (OLAP) v PostgreSQL. Na základě obdržených testovacích dat se pak navrhne a implementují různé analytické dotazy, a pak se porovná jejich výkonnost.

## Popis problematiky

Analytické dotazy jsou dotazy, které provádějí komplexní analýzu dat. Provádění analytických dotazů může být výpočetně náročné při počítání takových dotazů nad daty (v tomto projektu i nad tabulkami obsahujícími miliony záznamů), které nebyly předtím předzpracovány (před agregovány). OLAP využívá před počítání hodnot v různých dimenzích a vytváření OLAP kostek a pivot tabulek pro zrychlení takových dotazů.

V PostgreSQL 9.5 byla přidány nové operace *GROUPING SETS*, *CUBE* a *ROLLUP*. Tyto operace velmi usnadňují vytváření pivot tabulek tím, že za nás efektivně řeší komplexní seskupovací a agregační operace. Pivot tabulky následně můžeme realizovat pomocí materializovaných pohledů, které budeme periodicky obnovovat a analytické dotazy budeme počítat na základě těchto materializovaných pohledů.

## Testovací data

První část dat se týká spojení, které vznikají mezi aplikací běžící na PDA a serverem. PDA je jednoznačně identifikováno svým *pda\_imei*. Každé PDA je umístěno v nějakém autě a dané auto je identifikováno svým *car\_key*. Každé PDA obsahuje sim kartu. Sim karta je identifikována *sim\_imsi*. GSM síť je identifikována *gsmnet\_id*. U každého vytvořeného spojení se také ukládá verze aplikace použité pro připojení (atribut *program\_ver*). Součástí záznamů je i použitý protokol pro vytvoření spojení (atribut *method*). Každý záznam obsahuje položku *time* identifikující čas vytvoření záznamu v tabulce. Tyto data jsou uložena v tabulce *conn\_log*. Pro každé nové spojení se vloží nový záznam do tabulky *conn\_log*.

Druhá část dat se týká stavových hlášení, které aplikace periodicky posílají serveru. Každé hlášení obsahuje identifikátor auta (*car\_key*) ze kterého bylo hlášení podáno a informace o čase běhu aplikace (*app\_run\_time*) a zařízení (*pda\_run\_time*) v hodinách. Také součástí hlášení je typ zařízení (*device*). Každý záznam obsahuje položku *time* identifikující čas vytvoření záznamu v tabulce. Tyto data jsou uložena v tabulce *service\_log*. Každé nové stavové hlášení se vloží jako nová položka do tabulky *service\_log*.

**Relační schémata** (podtržené atributy jsou primární klíče daného schéma)

*conn\_log*(log\_key, *sim\_imsi*, *time*, *car\_key*, *pda\_imei*, *gsmnet\_id*, *method*, *program\_ver*)

*service\_log*(service\_key, *car\_key*, *time*, *app\_run\_time*, *pda\_run\_time*, *device*)

## Objemy dat

Testovací vzorek obsahuje v tabulce *conn\_log* 3710626 záznamů a v tabulce *service\_log* 2323534 záznamů.

Skutečný objem dat se pak pohybuje u *conn\_log* kolem 53559217 záznamů a u *service\_log* kolem 30068431 záznamů.

## **Postup řešení**

- Nastudování problematiky materializovaných pohledů a analytických dotazů v PostgreSQL
- Analýza testovacích dat
- Návrh analytických dotazů (Slovní popis vhodných analytických dotazů)
- Implementace analytických dotazů (implementace v SQL)
  - s vytvořením pivot tabulek, s použitím operací ROLLUP, CUBE,...
  - bez pivot tabulek (analytické dotazy vyhodnocovány nad neagregovanými záznamy)
- Validace dotazů (zda implementované dotazy dělají co opravdu mají)
- Testování výkonnosti dotazů
- Vyhodnocení výkonnosti implementovaných dotazů
  - příprava srozumitelné prezentace výsledků = grafy, tabulky...
- Závěrečné zhodnocení projektu a příprava finálního výstupu

## **Technické prostředky**

### **SW:**

- PostgreSQL server verze 9.5.0 a vyš
- *psql* -- PostgreSQL interactive terminal

### **HW:**

benchmarky budou prováděny na těchto strojích

- CPU Intel Core i5 2500K (3,3GHz, 6MB)  
RAM 8 GB DDR3 1600MHZ  
SSD Samsung 850 EVO 250GB
- CPU Intel Core i5 (3,4 Ghz, 6MB)  
RAM 16 GB DDR3 1600 MHZ  
SSD 256 GB onboard SSD (Macbook Pro)

## **Příklady analytických dotazů**

- Pro každou verzi programu zjistit počty různých zařízení
- Pro každou verzi programu zjistit počet restartů jeho programu (*app\_run\_time* ~ 0)
- Pro každé zařízení zjistit počet restartů
- Pro každé zařízení zjistit, kdy bylo uvedeno do provozu
- Pro každé auto zjistit počet různých zařízení
- Pro každý typ zařízení zjistit počet restartů aplikace
- Kolik bylo vytvořeno spojení v určitých časových dimenzích (rok, měsíc, den, ...)
- Kolik unikátních zařízení se připojilo v určitých časových dimenzích
- Pro každé zařízení zjistit, kolik hodin bylo používáno

## Návrh výkonnostních testů

Testy musí být prováděny na verzi PostgreSQL 9.5.0 a výš, protože nižší verze neobsahují příkazy *GROUPING SETS*, *CUBE*, *ROLLUP*. Měření výkonnosti se bude provádět pomocí příkazu *EXPLAIN ANALYSE*, který obsahuje podrobný exekuční plán provádění daného dotazu. Dotazy budou testovány na strojích výše zmíněných. Všechny analytické dotazy by se měli porovnávat s jejich neoptimalizovanými variantami (varianty dotazů nepoužívající pivot tabulky).

## Rozdělení práce

Členové týmu jsou Ondřej Benkovský, Tom Bartoň, Vojtěch Frnoch a Andrea Navrátilová.

- Ondřej Benkovský jako vedoucí týmu se bude starat o řízení projektu a provede analýzu testovacích dat. Také se bude podílet na implementaci analytických dotazů a postará se o přípravu finálního výstupu projektu.
- Tom Bartoň se postará o validaci všech implementovaných analytických dotazů a bude se podílet na implementaci analytických dotazů.
- Vojtěch Frnoch provede testování výkonnosti dotazů a vyhodnocení těchto testů.
- Andrea Navrátilová navrhne a implementuje analytické dotazy.

## Časový plán

- Do 29.3 se nastuduje problematika materializovaných pohledů, analytických dotazů a provede se analýza testovacích dat
- 29.3 – 5.4 návrh analytických dotazů
- 6.4 – 18.4 implementace dotazů a jejich validace
- 19.4 – 25.4 testování výkonnosti dotazů a vyhodnocení testů
- 26.4 – 5.5 příprava finálního výstupu projektu
- 6.5 odevzdání projektu