Analytické dotazy a materializované pohledy

Cíl

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

Popis problematiky

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

V PosgreSQL 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 hlášeních, 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 (<u>podtržené atributy</u> jsou primární klíče daného schéma)

conn_log(<u>log_key</u>, sim_imsi, time, car_key, pda_imei, gsmnet_id, method, program_ver)

service_log(<u>service_key</u>, 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 PosgreSQL
- 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 výš
- *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 a analytických dotazů
- 29.3 – 8.4 se provede analýza testovacích dat, návrh analytických dotazů a začne se implementace analytických dotazů
- 8.4 – 15.4 implementace dotazů a jejich validace
- 16.4 – 25.4 testování výkonnosti dotazů a vyhodnocení testů
- 26.4 – 5.5 příprava finálního výstupu projektu
- 10.5 odevzdání projektu