PA036: Projekt z DB systémů

Vlastislav Dohnal

Diskuze v Discord, server PA036-2023

Cíle předmětu

- Praktické používání databázového systému
  - Využití rozšířených vlastností
- Vytvoření týmového projektu a jeho prezentace
  - Tým = vždy 4 studenti
- Témata a průběh projektů: „Nové“ možnosti DB systémů
  - Představení problematiky
  - Otestování výkonnosti
  - Porovnání s jinými možnostmi
  - Co tým, tak jedno téma

Průběh předmětu

1. fáze
   - Sestavení týmu
   - Volba tématu projektu, volba náhradního tématu projektu; určení kompetencí v týmu
2. fáze
   - Bližší specifikace projektu, konzultace s vyučujícím, stanovení finálního rozsahu
   - Včetně Gantt diagramu prací (co, kdy, kdo) včetně odhadů pracnosti (v člověkohodinách) s týdenní granularitou termínů
3. fáze
   - Odevzdání specifikace projektu
   - Vypracování posudku na projekt týmu „s náhradním tématem“
4. fáze
   - Prezentace projektu ostatním, představení posudku oponenty, diskuze
5. fáze
   - Vlastní realizace (doma 🙂, konzultace s vyučujícím v rámci hodin semináře, týmové schůzky a týdenní reporty)
6. fáze
   - Odevzdání hotového díla, včetně původního Gantt diagramu a upraveného Gantt diagramu podle skutečného průběhu (tj. plán vs. skutečnost)
7. fáze
   - Finální prezentace výsledků projektu

Přesné časování a termíny sledujte v interaktivní osnově v ISu.

Fáze 5. se musí překrývat s fázemi 3. a 4.

Osnova

- Databázové systémy a krátké povídání k NoSQL
- Průběh řešení projektu
- Témata projektů

Databázové systémy a přístup k nim

- Relační – PostgreSQL
- NoSQL – dokumentové, klíč-hodnota, sloupcové, grafové
- Přístup k DBMS
  - Aplikační frameworky (Nette, Laravel, …)
  - REST API (DBCore)
  - JDBC/ODBC

Dokumentové NoSQL databázové systémy

- MongoDB
  - Datový model - JSON dokument
  - Dynamické schéma
  - Primární přístup k dokumentu pomocí ID
  - Sekundární přístup – index nad vybraným atributem dokumentu
- CouchDB
  - Analogický k MongoDB

Klíč-hodnota NoSQL databázové systémy

- Redis
  - Datový model – dvojice klíč-hodnota
  - Primitivní, vysoce výkonné operace nad seznamy, množinami a asoc. poli
  - Často jako in-memory cache – lze nastavit expiraci záznamů
  - Zpracování GEO souřadnic – včetně rozsahových dotazů
- Riak
  - Analogický k Redis, navíc zvládá JSON dokumenty jako hodnoty, umí sekundární indexy
    - Indexy se musí udržovat ručně

Sloupcové / grafové NoSQL DB systémy

- Cassandra
  - Relační datový model
  - Variabilní schéma – skupiny sloupců
  - CQL jazyk podobný SQL
- Neo4j
  - Datový model – orientovaný multigraf

Měření výkonnosti

- jednotlivé dotazy (INS/UPD/DEL) - transactions per second (TPS)
- dávky dotazů (emulace aplikačního využití) - transactions per second (TPS)
- nástroje
  - pgBench (vlastní scénáře; implementuje TPC-B benchmark)
  - jMeter
- plán vyhodnocení dotazu
  - příkaz EXPLAIN
  - vypisuje jak odhadovaný čas, tak i čas vyhodnocení

Měření výkonnosti

- Každý test musí být proveden opakovaně
  - se zachycením požadovaných metrik
- Prezentovat průměrnou hodnotu metriky, její odchylku
- HW
  - vlastní
  - PC v učebně
  - Stratus@FI
  - MetaCentrum

Testovací data

- buď součástí zadání, nebo vlastní
- pgBench umí generovat data (analogie k TPC-B)
  - jiný TPC benchmark
- Kaggle
- generátory dat v Python
- pro každý projekt je minimální objem 1 milion záznamů
  - pokud nemá vlastní specifická data

Průběh řešení projektu - začátek až první prezentace

- Zvolení tématu a specifikace cílů projektu
  - nahlášení emailem primární téma (preferované), sekundární a terciální (náhradní)
  - přidělení tématu Vám zpětně potvrdím
- Vytvoření projektu v gitLab.fi.muni.cz (podle instrukcí na Wiki)
- První prezentace:
  - podrobný plán a cíle projektu, způsob jak naplnit cíle
  - metodika testování a vyhodnocení (jaká data, jak velká, co/jak měřit, kolik opakování, ...)
  - definice kompetencí členů týmu, časový harmonogram (Gantt) s dělením na členy týmu
- Každý tým obdrží oponentský posudek od jiného týmu

Průběh řešení projektu - během semestru

- commity ve vašich projektech
  - týdenní doplňování aktuálního Gantt diagramu o čas strávený na projektu (každého člena týmu)
  - stačí jméno, počet hodin a několik slov -> identifikace nepracujících a jejich hodnocení „X“
- konzultace s vyučujícím v učebně, popř. na Discordu

Průběh řešení projektu - druhá prezentace

- Projekt je hotový
- Vložená finální prezentace do repozitáře
- Odevzdání přiřazením merge-requestu vyučujícímu (viz Wiki)
- Finální prezentace:
  - zopakování (stručně) zadání a cílů projektu
  - přístup k řešení, např. popis technologií, dat
  - výsledky experimentů
  - zhodnocení časového plánu (plánovaný Gantt vs. průběžně aktualizovaný Gantt)
- Každý tým obdrží oponentský posudek od jiného týmu

Časové nároky předmětu

- dotace 2 kredity -> cca 52 člověkohodin na semestr a studenta
  - realizace během výuky semestru - spíše tedy 3-4 týdně!
- Gantt
  - započítat povinné přítomnosti na semináři (první a druhé prezentace, plus úvodní hodina)
    - plánujte 6h pro první prezentace a 6h pro druhé prezentace
  - nástřel průběhu projektu
    - rozdělení do několika samostatných etap
    - odhad jejich náročnosti v hodinách (přes členy týmu)

Témata projektů pro 2023

1. Cizí datové zdroje
2. Řízení přístupu k záznamům relace
3. Rozšířené statistiky a výkon DB
4. Monitorování databázového systému
5. Aplikační framework a kešování DB
6. Notifikace DB serveru a aplikační framework
7. High Availability in PostgreSQL
8. Optimalizace úložiště DB
9. Horizontální dělení relací
10. Výkonnost zpracování JSON dokumentů
11. Výkonnost zpracování XML dokumentů
12. Full-text vyhledávání
13. Zpracování časových řad
14. Archivace dat a obnovení
15. Analytické dotazy a materializované pohledy A
16. Analytické dotazy a materializované pohledy B
17. Prohledávání grafů
18. Šifrování dat
19. Volba typu primárního klíče
20. Vyhledání odpovědí v knowledge base
21. Klasifikace emailové žádosti
22. Rozlišení pojmenovaných entit
23. Rozpoznání objektu ve fotografii