Posudek bakalářské práce

Autor: Michal Stanke
Název: Fulltext search in HBase database
Posudek vypracoval oponent práce: Ing. Marek Šmíd

Předložená bakalářská práce si klade za cíl vyzkoušet a porovnat možnosti fulltextového vyhledávání v Apache HBase databázi, a to jak na vlastní implementaci, tak u již existujících řešení. Jako pokusná data, na kterých bylo vyhledávání zkoušeno, byly adresy URL.

Text práce je napsán v anglickém jazyce, čehož si vážím, i když jeho úroveň má několik málo nedostatků. Určitě by pomohla korektura. Např. mi nepřijde obvyklé označovat fyzický server (host) pro VPS jako *maternal server*, viz str. 37, 1. odst. K jazyku obecně patří, že v několika případech se vyskytují zkrácené formy (např. *it's* v 1. odst. závěru), občas chybí členy, a někdy je uplatněn český větný pořádek. Překlepů je velmi málo (*immerse* vs. *immense*, str. 39., první odst.).

Po technické stránce je text vhodně formátován, obrázky jsou dobře čitelné. Bibliografie je rozsáhlá, a i přesto, že logicky obsahuje většinu zdrojů elektronických, jsou záznamy korektně citovány. Veškeré ostatní náležitosti jsou splněny. Ke grafům mám několik připomínek a jedno vylepšení. Chybou u grafů v obr. 30 (str. 46), v obr. 32 (str. 47) a v obr. 33 (str. 48) je, pokud jsem dobře pochopil slovní shrnutí výsledků v jejich okolí, že osa *y* má jednotku v milisekundách, nikoliv v sekundách, jak je v nich uvedeno. U grafu v obr. 25 (str. 44) chybí popis osy *y* úplně, u grafu v obr. 33 (str. 48) chybí legenda. Možné vylepšení by bylo všechny naměřené body v grafech opatřit tzv. error bar, tedy sloupečkem, znázorňujícím rozptyl – dalo by to možnost na první pohled vidět (oproti poznámkám v textu), z jak rozptýlených hodnot průměr pochází.

Vlastní obsah práce má příjemný a přirozený tok, text je ucelený a dobře uzavřený. Rozdělení do kapitol je úměrné rozsahu textu, jen bych navrhl dvě změny: V kapitole 2 Motivation jsou mj. představeny existující systémy, což by zasloužilo jinak pojmenovanou kapitolu/podkapitolu (např. Related systems). Kapitola 3 State of the art obsahuje na konci představení API a konfigurace několika systémů – to by také mohlo být v podkapitole věnované implementaci.

Není zcela detailně popsáno, jak probíhalo opakované měření (jen že se měřilo třikrát, jak je uvedeno v kapitole 5.b na str. 36), bylo by zajímavé vědět, jestli se mezi měřeními nějakým způsobem vyprazdňovala např. disková cache, a jestli byla tři stejná měření vždy hned po sobě (možná by bylo vhodné je více časově rozložit a střídat různá měření, aby se více potlačilo proměnlivé zatížení poskytovatele VPS).

Případné dotazy pro autora:

- Jaký vliv na výsledky by mělo rozložení stejných dat na více uzlů místo pouze na jeden? Jak moc by se snížil čas např. vyhledávání?
- Na stránce 30 v kapitole 4.c, ve 4. odst. píšete, že bychom rádi při vyhledávání měli složitost nezávislou na velikosti uložených dat. Předpokládám, že tedy myslíte konstantní složitost *O(1)*. Je toho možné dosáhnout? Jaká je typická složitost vyhledávání v indexu, a na jakou složitost odhadujete Vaše naměřené výsledky?

Celkově práce zcela splnila zadání a má i přes drobné technické nedostatky dobrou úroveň.

Předloženou bakalářskou práci hodnotím známkou

A – výborně.

V Praze dne 15. června 2015

Ing. Marek Šmíd