

Posudek oponenta diplomové práce

Student: Tutko Jakub, Bc.

Téma: Distribuované zpracování rozsáhlých dat na platformě Java (id 20647)

Oponent: Rychlý Marek, RNDr., Ph.D., UIFS FIT VUT

- 1. Náročnost zadání** **průměrně obtížné zadání**
Diplomová práce porovnává výkonnost distribuovaných databází Halyard a HGraphDB při využití vlastní Spark aplikace pro získávání a ukládání příspěvků ze sítě Twitter a Facebook a jejich pozdější analýzu. Jedná se o průměrně obtížné zadání.
- 2. Splnění požadavků zadání** **zadání splněno**
Zadání je splněno bez výhrad.
- 3. Rozsah technické zprávy** **je v obvyklém rozmezí**
Rozsahem je technická zpráva v obvyklém rozmezí, od úvodu po závěr obsahuje 56 vysázených stran. Text je celkově informačně bohatý (zejména v druhé polovině zprávy) a také vložené obrázky, tabulky aj. jsou přínosné a pro doprovodný text nezbytné.
- 4. Prezentací úroveň předložené práce** **85 b. (B)**
Technická zpráva má velmi dobrou prezentací úroveň. Zpráva má logickou strukturu vycházející ze struktury zadání, kapitoly jsou přiměřeného rozsahu a dobře na sebe navazují. Přestože první polovina práce spíše shrnuje obecné informace o použitých technologiích, možná více podrobně, než by bylo potřeba (např. popis HDFS v kap. 2.1.2 není dále plně využit), druhá polovina je velmi dobře čitelná a pochopitelná (oceňuji zde časté použití ilustrativních příkladů).
- 5. Formální úprava technické zprávy** **70 b. (C)**
Z hlediska formální úpravy má práce drobné nedostatky. Vyskytují se zde občasné gramatické (zejména interpunkci) či typografické chyby (nepoužívejte spojovník místo pomlčky; API je zkratka, takže má být verzálkami), které však významně nesnižují srozumitelnost obsahu.
- 6. Práce s literaturou** **75 b. (C)**
Seznam literatury čítá 20 položek, z nichž asi polovina je online dokumentace použitých technologií a druhá polovina pak odborné publikace. Všechny zdroje ze seznamu jsou v textu zprávy řádně citovány a je patrný způsob a rozsah jejich použití. Vytknout lze nesprávné uvedení jmen autorů v případech 12. a 17. položky seznamu literatury.
- 7. Realizační výstup** **85 b. (B)**
Realizační výstupem je programové řešení v jazyce Java pro inicializaci databází Halyard a HGraphDB, Spark úloha pro získávání Twitter a Facebook dat a jejich ukládání do databází, a aplikace pro dotazování těchto dat z databází. Části lze volat samostatně, přičemž z výstupu lze vyčíst způsob a rychlost provedení (výsledky ukázkových úloh autor popsal v textu zprávy). V autorem prezentované konfiguraci, která byla použita pro změření v textu prezentovaných výsledků, je využit výchozí "standalone" plánovač Spark úloh, což je škoda, protože použitím alternativního YARN plánovače by bylo nejspíš dosaženo lepších výsledků, vzhledem ke znalosti rozložení datových bloků vstupních dat na HDFS. I při autorem použité konfiguraci je však výsledné řešení plně funkční a měření poskytuje validní výsledky.
- 8. Využitelnost výsledků**
Výsledky porovnání výkonu databází Halyard a HGraphDB lze využít pro volbu vhodné technologie v navazujících projektech.
- 9. Otázky k obhajobě**
 - Jaký je vliv plánovače Spark úloh (angl. "job scheduler") na výkon prezentovaného řešení?
- 10. Souhrnné hodnocení** **85 b. velmi dobře (B)**
Výsledkem diplomové práce je kvalitní a přínosná technická zpráva a dobře provedené programové řešení. Celkově považuji výsledek za mírně nadprůměrný a navrhuji hodnotit práci stupně **velmi dobře (B)**.

Prohlášení: Uděluji VUT v Brně souhlas ke zveřejnění tohoto posudku v listinné i elektronické formě.

.....

podpis