

High Performance Analytics

Předkládaná diplomová práce se zabývá moderní problematikou zpracování rozsáhlých dat (tzv. Big Data). V první, teoretické části práce autor popisuje tuto oblast z obecného hlediska, vysvětluje související pojmy a diskutuje klíčové problémy a možné přístupy k jejich řešení, včetně zmínění existujících implementací a nástrojů. Ve druhé, praktické části autor demonstruje problematiku zpracování rozsáhlých dat na konkrétních datech (z oblasti automobilového průmyslu a z oblasti výroby hraček) pomocí konkrétního nástroje (*SAS Visual Analytics* firmy *SAS Institute*), který poskytl vedoucí práce. Jedná se tedy především o přehledovou práci, jejím hlavním přínosem by měl být ucelený popis a zhodnocení rozsáhlejší a dosud ne dobře popsané oblasti, doplněný o praktickou ukázkou.

Hlavní klady práce:

- Aktuální problematika, k níž zatím neexistuje mnoho snadno dosažitelných a především úplných zdrojů informací a se kterou se autor práce pravděpodobně dříve nesetkal.
- Autor evidentně získal dobrý přehled o zvoleném tématu a dokázal v textu (teoretické části) postihnout většinu souvisejících témat a problémů.
- Rozšíření přehledové práce o reálný příklad (praktická část) je určitě zajímavý a v existující literatuře ne zcela typický prvek.

Připomínky:

- Z obecného hlediska je diplomová práce především odborný text, ve kterém očekáváme pouze fakticky přesné informace, bez okras. Nejsou tedy na místě ležérní až hovorové obraty, pokusy o vtipné poznámky, citoslovce, kvanta otázek k zamyšlení, pobízení čtenáře ke čtení dalšího textu apod. Všechny tyto aspekty bohužel kazí dobrý dojem a snižují úroveň diplomové práce z odborné publikace na úroveň popularizačního článku.
- U poměrně velkého množství zmíněných technologií chybí reference (viz např. tabulka 2). V současné době je zřejmé, co má autor na mysli, ale za pár měsíců už nebude jasné např. o jakém systému, resp. o jaké jeho verzi byla řeč. Přínos přehledového zpracování dané oblasti se tedy opět zbytečně snižuje.
- Zatímco teoretická část je poměrně dobře zpracovaná (pomineme-li výše uvedené připomínky), praktická část je slabá. Kromě popisu použitých datových množin, způsobu jejich předzpracování a operací, které daný SW nabízí, najdeme pouze dva potenciálně zajímavé výstupy, které mají značné nedostatky:
 - Kapitola 7.4.4 obsahuje několik screen shotů z použitého SW, aniž by autor diskutoval k čemu jsou dobré a proč mají být zahrnuty do práce. Další množství screen shotů je v příloze C, resp. přes 40 je na přiloženém CD. Ale opět bez jakéhokoli komentáře. Smysl tohoto výstupu je tedy nejasný, působí spíš jako snaha o zaplnění prostoru.
 - Kapitola 8 (něco málo přes 2 stránky textu) hodnotí použitý SW především z hlediska rychlosti zpracování vybraných analytických operací nad danými datovými množinami. Dozvídáme se ovšem především to, že vybrané operace použitý SW provede v rozumném čase. Diskuse nedostatků nebo problémů je velmi krátká a typicky končí slovem „future work“.

Obecně bych v praktické části očekávala jeden ze dvou scénářů: (1) Rozsáhlou diskusi použitého SW v rámci mnoha různých případů užití a jejich odlišných nároků, diskusi výstupů v kontextu daných případů užití, analýzu nevýhod a problémů použitého SW, diskusi možných řešení apod. (2) Srovnání použitého SW s několika jinými podobnými a diskusi jejich výhod a nevýhod pro různé aplikace. Jinak řečeno, v tomto typu práce bych čekala objektivnější a podstatně propracovanější zhodnocení.

Drobné připomínky:

- V českém abstraktu je velké množství chyb.
- Obrázky často přesahují okraje, některé (např. obrázek 8) nejsou celé.
- V odborném textu se nepoužívají stažené tvary (don't, can't apod.).
- Při odkazování je třeba uvést také na co se odkazujeme – např. „see 6.3.3“ (je to obrázek, tabulka, příklad nebo kapitola?).
- Míchání jazyků je poměrně zvláštní – práce i použitý SW je v angličtině, ale např. názvy sloupců pro datové množiny (viz např. obrázek 13 a 14, resp. tabulka 14) jsou v češtině, resp. česko-anglicky (např. „název kraje dealera“).
- Systematické chyby: „s“ pro třetí osobu není totéž, co „s“ v množném čísle (např. „paragraphs contains“); Czech Republic => the Czech Republic; last, first, previous, following, next apod. velmi často postrádají určitý člen;

Přes uvedené připomínky práce splňuje původní zadání i cíl. Student pronikl do zvolené problematiky a prokázal schopnost řešit zadaný problém. Práce Andreje Kalického podle mého názoru splňuje podmínky na diplomovou práci kladené, a proto ji **doporučuji** k obhajobě.

V Praze, 24.8. 2013

RNDr. Irena Holubová, Ph.D.
KSI MFF UK