Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Bc. Lukáš Kolek		
Název práce	Aproximativní datové profilování		
Rok odevzdání	2021		
Studijní program	Informatika	**Studijní obor**	Softwarové a datové inženýrství

Autor posudku	RNDr. Michal Kopecký, Ph.D.	**Role**	Vedoucí
Pracoviště	KSI MFF UK		

Text posudku:

Cílem této práce bylo na základě studia možností datového profilování navrhnout a implementovat nástroj, který by datové profilování umožnil. Důraz byl kladen na možnost nabídnout nejen přesné algoritmy pro výpočet charakteristik sloupců, jako jsou histogramy, kvantily a další, ale především algoritmy pro aproximativní výpočty těchto charakteristik za účelem úspory potřebného paměťového prostoru. Dalším požadavkem byla možnost data data vzorkovat za účelem úspory potřebného času.

Autor v rámci řešení nastudoval řadu alternativních řešení pro výpočty sloupcových charakteristik a v práci je přehledně shrnul včetně jejich teoretických prostorových a časových asymptotických složitostí.

Text práce je dobře strukturován a psán formálním a přesto srozumitelným a čtivým jazykem.

Díky modularitě aplikace je možné v rámci jediného běhu aplikace nejen vybrat z většího počtu možných implementací stejného algoritmu, ale počítat stejnou charakteristiku více metodami najednou. To autorovi umožnilo provést poměrně rozsáhlá měření nad reprezentativním vzorkem dat, a získat praktické výsledky a porovnání vhodnosti jednotlivých algoritmů a jejich nastavení pro různé velikosti vstupů a různé vlastnosti vstupních dat.

Práce jako taková se v souladu se zadáním soustřeďuje především na jednosloupcové charakteristiky datových sad, pro které je možné nalézt jejich aproximativní alternativy. Ostatní algoritmy jsou zastoupené pouze v nejnutnější míře, potřebné pro implementaci a získání výsledků měření. V rámci analýzy práce popisuje problematiku v širším kontextu a ukazuje, že má autor přehled i v souvisejících oblastech.

Při návrhu aplikace byly voleny takové prostředky a návrhové vzory, aby bylo docíleno vysoké efektivity s co nejnižší režií. Výsledky ukazují, že předkládané řešení předčí svojí kvalitou, nízkými nároky na server, počítající analýzy, a rychlostí, kterým se data analyzují, dostupná komerční řešení. Použité algoritmy jsou v práci dobře popsány, včetně jejich formulace v pseudo-kódu.

Celkově se domnívám, že práce splňuje všechny požadavky kladené na práce diplomové.

Aplikace je stabilní, a jejímu širšímu využití brání v podstatě jen zatím prototypové rozhraní pro konfiguraci požadovaných analýz a zobrazování výsledků. Pokud by se k aplikaci dodělalo - například prostřednictvím bakalářské práce - odpovídající uživatelské rozhraní, byl

by nástroj použitelný i pro běžné uživatele.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

Pokud práci navrhujete na zvláštní ocenění (cena děkana apod.), prosím uveďte zde stručné zdůvodnění (vzniklé publikace, významnost tématu, inovativnost práce apod.).

Datum 24. ledna 2021 **Podpis**