Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

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

Autor posudku	Martin Svoboda	**Role**	Oponent
Pracoviště	KSI		

Text posudku:

Cílem hodnocené diplomové práce bylo seznámení se s existujícími přístupy a nástroji na datové profilování a následný návrh a implementace nástroje, prostřednictvím kterého by jednotlivé vybrané přístupy mohly být experimentálně srovnány. Zadání práce bylo v tomto smyslu splněno.

V souladu s výše uvedeným se autor nejprve věnuje praktickému a uživatelsky motivovanému srovnání dostupných nástrojů, ať už volně šiřitelných nebo komerčních, a to alespoň v rozsahu dostupných informací. Stejně tak se věnuje popisu a vzájemnému srovnání řady konkrétních přístupů a metod, které se k datovému profilování používají. Pozornost je věnována strategiím založeným na přesných výpočtech, aproximativních heuristikách, stejně jako vzorkování. Z hlediska konkrétních analýz byl důraz kladen na frekvenční analýzy, zjišťování kardinalit, vytváření histogramů nebo počítání kvantilů.

Práce je napsána v českém jazyce plynulým a dobře srozumitelným stylem, počet chyb nebo překlepů je minimální. S ohledem na způsob řešení obsahuje práce jako celek všechny očekávané součásti, většina textu ale jen popisuje existující přístupy, a tedy vlastní tvůrčí činnosti nebylo věnováno mnoho prostoru. Z hlediska rozsahu je práce spíše podprůměrná.

Přestože autor sám píše, že hlavním cílem měla být schopnost jednoprůchodového zpracování velkého objemu vstupních dat za využití omezeného množství systémové paměti, skutečné výzvy a příležitosti však v tomto ohledu nebyly adresovány a natož naplněny. Samotné experimenty pochopitelně nemusely být provedeny nad velkými daty, z podstaty kontextu je ale zřejmé, že navržený nástroj bude použitelný jen pro menší desítky GB, přitom objem dat v reálně existujících clusterech velkých společností se pohybuje v řádech přinejmenším desítek PB.

Pozornost však mohla být věnována nejenom již naznačeným distribuovaným a paralelním architekturám a programovacím modelům, ale např. i víceprůchodovým přesným analýzám s odkládáním dat na disk (což je běžná praxe v databázových systémech; navíc i v situaci, kdy časový faktor profilování není z hlediska praktického použití tím úplně klíčovým), netriviálním vícesloupcovým analýzám, jiným logickým modelům než relačnímu (který sice má dominantní postavení, s příchodem rodiny NoSQL systémů se ale běžně používají i jiné modely a ty nejsou přímočaře mapovatelné na tabulkové chápání dat) nebo také provedení skutečně rozsáhlých experimentů, díky kterým by bylo možné vypovídajícím způsobem formulovat případná zjištění a doporučení.

V praktické rovině spočívají vlastní přínosy práce v podstatě jen v implementaci prototypového nástroje, v rámci kterého je možné na základě JSON konfigurace realizovat datové profilování v rámci jednoho osobního počítače se vstupními daty dostupnými přes JDBC nebo umístěnými lokálně v CSV souborech. Přestože škála vybraných analýz je velká a architektura nástroje je navržena modulárně a rozšiřitelně, implementace samotných

existujících přístupů nebyla příliš komplikovaná a nástroj samotný není díky absenci GUI na širší využití zatím dostatečně připravený.

Celková složitost práce je tak na relativně nízké úrovni a bez návrhu vlastních přínosů výzkumného charakteru. Přestože množství citovaných odborných publikací není vysoké, autor prokázal porozumění studované oblasti, stejně jako široké palety konkrétních algoritmů, přístupů a nástrojů. Pozitivní je také diskuze výsledků realizovaného experimentálního srovnání jednotlivých metod. Je však potřeba dodat, že z jedné datové sady, navíc poměrně malé a jednoduché, lze obecně platné závěry usuzovat jen s velkým rizikem.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

Datum 25. ledna 2021 **Podpis**