Posudek bakalářské práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Ondřej Kaštovský
Název práce	Predikce délky trvání datového profilování
Rok odevzdání	2024
Studijní program	Informatika
Specializace	Programování a vývoj software
Autor posudku	Filip Kliber Oponent
Pracoviště	Katedra distribuovaných a spolehlivých systémů

K celé práci

	lepší	OK	horší	nevyhovuje
Obtížnost zadání			X	
Splnění zadání		X		
Rozsah práce *... textová i implementační část, zohlednění náročnosti*		X		

Autor se ve své práci věnoval rozšíření platformy pro správu dat Ataccama ONE, konkrétně rozšíření služby datového profilování. Cílem bylo před vlastní profilování umístit modul, který umožní dobře odhadnout jak dlouho bude plné profilování trvat a díky tomu umožnit lepší a efektivnější plánování jednotlivých běhů. Samotné plánování již součástí práce nebylo. Řešení práce považuji za zdařilé.

Textová část práce

	lepší	OK	horší	nevyhovuje
Formální úprava *... jazyková úroveň, typografická úroveň, citace*	X			
Struktura textu *... kontext, cíle, analýza, návrh, vyhodnocení, úroveň detailu*			X	
Analýza		X		
Vývojová dokumentace		X		
Uživatelská dokumentace		X		

Autor textovou část práce vhodně rozdělil do jednotlivých kapitol. Jazyková úroveň textu je nadstandardní, nebyl jsem však spokojen se strukturou práce. Autor vhodně uvedl čtenáře do problému, nicméně hodně přeskakoval mezi analýzou problému a jeho řešením. Stávalo se, že se autor věnoval řešení dílčího problému, ale čtenáři nebylo jasné proč se tento problém řeší a jaké mají výsledky vliv na celou práci. Například v kapitole 3 se autor věnuje vytváření predikčního modelu a už v sekci 3.1 je příklad kódu v PL/pgSQL pro generování tabulek s náhodnými daty, aniž by bylo zřejmé proč se zrovna do tabulek generují textové řetězce. Myslím si, že by pomohlo rozšířit sekci *Struktura Práce* v úvodu a více popsat jak na sebe jednotlivé kapitoly navazují. Rovněž by čtenáři pomohl příklad dat s jakými se bude autorův software potýkat.

(pokračování na další straně)

V textu práce se píše: „v přiložených grafech ovšem nebudeme z důvodu citlivosti dat úvádět přesné počty záznamů tabulek“. Úplně si nedovedu představit jak by někdo z grafu vyčetl přesné počty záznamů nemluvě o tom, že tento text přímo navazuje na sekci *3.1.1 Generování velkých tabulek v PL/pgSQL* a tedy nevím o jaké citlivosti (náhodně vygenerovaných dat) se mluví.

Autor se věnuje datům, které mají strukturu jedné (SQL-like) tabulky. Změnila by se nějak metodika, kdyby bylo potřeba predikovat dobu trvání nějakého složitějšího spojení (join) více tabulek?

Na kvalitě práce trochu ubírá několik řádků přesahujících šířku stránky (např. v sekcích 5.1 a 5.2).

Implementační část práce

	lepší	OK	horší	nevyhovuje
Kvalita návrhu *... architektura, struktury a algoritmy, použité technologie*		X		
Kvalita zpracování *... jmenné konvence, formátování, komentáře, testování*	X			
Stabilita implementace		X		

Implementační část práce je kvalitně zpracovaná. Autor si prošel celým procesem vývoje SW a výsledné dílo splňuje všechny parametry moderního SW. Z důvodu proprietárnosti nástroje Ataccama ONE se autor nabídl projekt mi ukázat osobně. Tuto nabídku jsem přijal a výsledný produkt vypadá dobře — odhadovaná doba běhů datového profilování a následná skutečná běhu je podobná.

Nástroj Ataccama ONE je psaný v jazyce Java a tím byl autor vázán i při implementaci nové komponenty. Autor si v textu práce stěžoval, že nebylo lehké najít vhodnou knihovnu pro trénování predikčního modelu napsanou v jazyce Java. Proč autor nezvolil pro trénování modelu nějaký jiný jazyk, který by podporoval vhodnější knihovnu a předávání dat do hlavní části aplikace by řešila nějaká forma meziprocesové komunikace?

Zradu na čtenáře autor připravil v ukázkovém výstupu REST API, které vrací odhad doby trvání konkrétního běhu. Povedlo se mu totiž špatně okopírovat výsledný JSON a nesedí v něm součet hodnot. Zprvu jsem nevěděl, jestli se jedná o chybu nebo jestli špatně chápu dokumentaci.

Celkové hodnocení Velmi dobře

Práci navrhuji na zvláštní ocenění Ne

Datum

Podpis