

Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce Michal Jurčo
Název práce Data Lineage Analysis Service for Embedded Code
Rok odevzdání 2023
Studijní program Informatika **Studijní obor** Softwarové a datové inženýrství

Autor posudku David Bednárek **Role** Oponent
Pracoviště KSI MFF UK

Text posudku:

Softwarová část posuzované práce reprezentuje nový modul do komerčního produktu Manta Flow, rozšiřující existující mechanismy data lineage analysis na fragmenty kódu v jazyce Python, vložené uvnitř skriptů mechanismu AWS Glue.

Přestože z hlediska algoritmů data lineage analysis práce nepřináší nic nového, není možné říci, že by obsahem práce byla pouhá implementace známých algoritmů. Ve zdrojových kódech tohoto typu se často vyskytují obraty (jako například programem skládaná jména otevíraných souborů nebo dokonce volaných procedur), které vyžadují částečnou interpretaci daného kódu, aby bylo vůbec možno kompletní lineage analysis provést. Tento problém je navíc zálužný tím, že skutečný rozsah obrátů, pro které je tato částečná interpretace nutná, vyjde najevo až při aplikaci na značně objemné zdrojové kódy z reálného života, což znamená, že analyzátor je potřeba implementovat inkrementálně zároveň s jeho testováním na reálných vstupech.

Druhou obtížnou částí problému byla integrace s existujícími analytickými nástroji, které sice od začátku počítají s integrací produktů analýzy z několika vstupních jazyků, evidentně však nebyly stavěny na masový výskyt krátkých fragmentů jednoho jazyka uvnitř jiného. To vedlo k nutnosti upravit i existující rozhraní a nástroje.

Celkově nejde o nijak zvlášť rozsáhlý software (cca 5000 řádek v Javě plus 1000 řádek úprav existujících kódů v několika jazycích), jde však o netriviální téma, komplikované i zapojením do existujícího a několik let používaného komerčního produktu. Výsledný software je funkční (ikdyž rychlost analýzy není nijak závratná), včetně zapojení do netriviálního soukolí, jehož součástí je mj. i připojení ke cloudu AWS.

Textová část obsahuje vysvětlení cílů a mechanismů analýzy, druhá polovina práce se pak věnuje technickým detailům integrace s existujícím software a komplikované architektuře celého systému, která zjevně pro krátké analyzované fragmenty není z výkonového hlediska ideální, což se autor práce pokusil zachránit přidáním dalšího komplikovaného mechanismu pro cachování mezivýsledků.

Z hlediska formy a kvality prezentace je text v kontextu softwarově orientovaných diplomových prací nadprůměrný, překvapivě v něm ovšem chybí jakýkoliv formalismus: Interní datové struktury i algoritmy jsou prezentovány pomocí obrázků (připomínajících spíše UML než matematickou strukturu), slovního popisu a případně kódem v jazyce Python. Výhodou tohoto přístupu je srozumitelnost i pro matematicky nevzdělané čtenáře,

nevýhodou je však nedostatečná přesnost, někdy i zbytečně dlouhý popis, který by se dal zkrátit odkazem na vhodnou matematickou strukturu.

V textu práce nejsou odkazy na žádnou literaturu z oblasti překladačů, statické analýzy kódu, v seznamu literatury dokonce není ani žádná práce z oblasti data lineage analysis (všech 17 prvků seznamu literatury jsou manuály k software). Tato absence je částečně pochopitelná vzhledem k tomu, že práce je rozšířením existujícího analyzátoru a základní algoritmus analýzy byl tak určen v rámci zadání. V akademickém kontextu je však absence řešerše související vědecké literatury nevhodná a pravděpodobně souvisí i s absencí formálního modelu problému v práci.

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 25. August 2023

Podpis