Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Andrej Jurčo		
Název práce	Data Lineage Analysis for PySpark and Python ORM Libraries		
Rok odevzdání	2023		
Studijní program	Informatika	**Studijní obor**	Softwarové a datové inženýrství

Autor posudku	Pavel Parízek	**Role**	Vedoucí
Pracoviště	Katedra distribuovaných a spolehlivých systémů		

Text posudku:

Cílem této práce bylo rozšíření platformy Manta Flow pro analýzu datových toků, a především komponenty určené pro analýzu skriptů napsaných v jazyce Python, o podporu knihovny PySpark a základní kostru analýzy knihoven pro objektově-relační mapování (například SQLAlchemy).

Autor navrhl a implementoval plugin, který ošetřuje volání knihovny PySpark a modeluje vliv jednotlivých operací na datové toky. Důležitým výsledkem je také rozšíření jádra komponenty pro analýzu skriptů v jazyce Python o další funkčnost, konkrétně například o podporu takzvaných "flow variables". Hlavní myšlenka je implicitní propagace datových toků z určitých objektů, jako třeba instance třídy PySpark DataFrame, na speciální atributy "read" a "write" používané jako rozhraní ke čtení a zápisu dat. Během vývoje toho pluginu musel autor vyřešit také různé další složité dílčí úlohy, především správné a dostatečně přesné modelování sloupců tabulek (zdrojů dat) a vyhodnocení dotazů (přístupů) na jednotlivé sloupce, pro které vznikají příslušné atributy objektů dynamicky.

Dále pak autor zpracoval důkladnou analýzu a design podpory objektově-relačního mapování, a všechno popsal v textu diplomové práce. Nicméně vývoj pluginu pro ošetření knihovny PySpark (včetně implementace, testování a odladění) zabral mnohem delší čas než jsme očekávali, takže proto není dokončena implementace podpory objektově-relačního mapování (knihovny SQLAlchemy). Podrobnosti jsou zdokumentované v textu diplomky.

Úroveň zpracování obou částí práce, tedy software a textu, je vysoká. Nemám zásadní připomínky.

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 25.5.2023 **Podpis**