Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Natália Potočeková		
Název práce	Data Lineage Analysis for Databricks platform		
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 systému Manta Flow pro analýzu datových toků o základní podporu platformy Databricks. Hlavní dílčí podúlohy uvedené v zadání byly tyto: (1) vytvořit obecnou kostru pro analýzu datových toků platforem založených na takzvaných "notebooks", jako právě Databricks a Jupyter Notebook, dále (2) implementovat analýzu konkrétně pro "Databricks notebooks", ve kterých se vyskytují buňky (fragmenty skriptů) v jazycích Python a SQL. Záměrem bylo integrovat existující skenery pro tyto jazyky, a doplnit funkčnost potřebnou specificky pro Databricks.

Autorka této práce provedla velmi důkladnou analýzu celé úlohy, relevantních technologií, a možných přístupů ke řešení. Dalším výstupem práce je kompletní a odladěná implementace, která již byla integrována do produktu Manta Flow. V rámci projektu musela autorka vyřešit několik poměrně složitých technických problémů, souvisejících především s výměnou informací mezi různými skenery.

Kvalita zpracování obou částí práce, tedy software a textu, je velmi vysoká. Především textová část je velmi rozsáhlá. Nemám žádné 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 21.8.2023 **Podpis**