

Technológie založené na notebookoch, ako sú Databricks a Jupyter Notebooks, si v posledných rokoch získali popularitu vďaka svojej prispôsobivosti a jednoduchému používaniu. Notebook je interaktívne výpočtové prostredie, ktoré umožňuje používateľom vytvárať dokumenty obsahujúce kód, vizualizácie a vysvetľujúci text na jednom mieste. Notebooky poskytujú priestor na prieskum dát, analýzu a dokumentáciu, čo používateľom umožňuje jednoducho rozvíjať a prezentovať svoju prácu. Schopnosť kombinovať vykonávanie kódu s vysvetleniami a vizualizáciami v rámci jedného dokumentu podporuje reprodukovateľnosť, zlepšuje spoluprácu medzi členmi tímu a motivuje vedcov k efektívnej práci s dátami. V tejto práci sme analyzovali technológiu Databricks, aby sme rozšírili platformu Manta Flow, vysoko automatizovaný nástroj na analýzu datových tokov, na podporu tejto technológie. Navrhli sme a implementovali nový skener, ktorý poskytuje základnú podporu pre analýzu notebookov Databricks napísaných v jazykoch Python a Databricks SQL. Poskytujeme tiež implementáciu takzvaného zdieľaného kontextu, ktorý možno použiť na prenos informácií medzi rôznymi skenermi na platforme Manta Flow. Na vizualizáciu interakcií medzi jazykmi a skenermi sme rozšírili Manta graf o nový typ uzla, ktorý predstavuje zdieľaný kontext. Okrem toho sme implementovali takzvaný jazykový kontext, ktorý umožňuje skenerom ukladať informácie užitočné na analýzu ďalších buniek napísaných v rovnakom jazyku v danom notebooku. Na záver demonštrujeme funkčnosť skenera a výsledné grafy, ktoré vytvára, na príkladoch Databricks notebookov.