Vo svete ETL nástrojov a spracovania dát je Python jedným z najčastejšie používaných jazykov. Skripty napísané v jazyku Python, ktoré definujú manipuláciu s dátami, zvyčajne používajú rovnakú knižnicu, PySpark, čo je Python API pre framework Spark, spoločne s databázovými knižnicami, využívajúc ich ORM funkcionalitu. Táto funkcionalita zvyčajne funguje podobným spôsobom vo väčšine relevantných knižníc. Nedávno bol MANTA Flow, vysoko automatizovaný nástroj na analýzu data lineage, rozšírený o skener jazyka Python a teraz je vo fáze rozširovania o podporu bežných frameworkov.
V tejto práci sme analyzovali knižnicu PySpark a technológiu SQLAlchemy ORM s cieľom rozšíriť Python skener firmy MANTA o podporu týchto dvoch často používaných nástrojov. V prípade knižnice PySpark sme navrhli a implementovali jadro pluginu pre skener jazyka Python, ktorý podporuje elementárnu funkcionalitu. Plugin je schopný analyzovať rôzne vstupné a výstupné možnosti DataFramov dostupné v PySparku pre súborové aj databázové dátové zdroje a je schopný propagácie dátových tokov počas transformácií s primeranou úrovňou overaproximácie, ako sme v práci demonštrovali. V prípade SQLAlchemy ORM sme navrhli riešenie, ktoré umožní skeneru analyzovať zdrojový kód využívajúci funkctionalitu ORM a jeho jadro by bolo možné použiť aj pre podporu iných knižníc poskytujúcich funkcionalitu ORM.