In the world of ETL tools and data processing, Python is one of the main languages used in practice. Python scripts that define data manipulations usually use the same Python framework, PySpark, which is the Python API for the Spark framework, alongside database libraries, using their ORM features. These ORM features usually work in a similar way in most of the relevant libraries. Recently, MANTA Flow, a highly automated data lineage analysis tool, was extended with a Python language scanner and now it is in the phase of being extended to support more commonly used frameworks.

In this work, we analyzed the PySpark library and the SQLAlchemy ORM technology in order to extend the MANTA's Python scanner with the support for these two frequently used tools. In case of the PySpark library, we designed and implemented a core of the plugin to the Python scanner which supports elementary functionality. The plugin is capable of analyzing various DataFrame input and output options available in PySpark for both file and database data sources, and it is able to propagate data flows during transformations with reasonable level of overapproximation, as demonstrated in the work. In case of the SQLAlchemy ORM, we designed a solution that would allow the scanner to analyze the ORM source code and its core could be used to support other libraries with ORM functionality as well.