Data integration tools often use embedded code for data manipulation tasks. Popular examples of such tools include AWS Glue data integration service, Databricks platform, Snowflake data cloud or SQL Server Integration Services (SSIS). Embedded code is typically written in programming languages such as Python, Java, C# or JavaScript. Manta Flow is an automated platform that can analyze data lineage in database models, data pipelines of data integration tools, and in application source code, but it lacks the ability to analyze embedded code.

In this work, we discussed potential ways to extend the capabilities of Manta Flow with the ability to analyze data lineage in embedded code. We created a general design of a reusable Embedded Code Service that leverages the existing potential of data flow analysis of source code, and uses it to analyze embedded code. We implemented a specialization of this service for the Python programming language, and to demonstrate its usefulness, we designed and implemented a prototype of data lineage scanner for AWS Glue data integration service. This scanner extensively uses the service to analyze data lineage in embedded Python scripts, which we demonstrated on a realistic example.