Abstract: Manta Flow is a tool for analyzing data flow in enterprise environment. It features Java scanner, a module using static analysis to determine the flows through Java applications. To analyze an application using some framework, the scanner requires a dedicated plugin. Although Java scanner provides plugins for several frameworks, to be usable for real applications, it is essential that the scanner supports as many frameworks as possible, which requires implementation of new plugins. Application using Apache Spark, a framework for cluster computing, are increasingly popular. Therefore we designed and implemented Java scanner plugin that allows the scanner to analyze Spark applications.

As Spark focuses on data processing, this presented several challenges that were not encountered in other frameworks. In particular it was necessary to resolve the data schema in various scenarios and track the schema changes throughout any operations invoked on the data.

Of the multiple APIs Spark provides for data processing, we focused on Spark SQL module, notably on Dataset, omitting the legacy RDD. We also implemented support for data access, covering JDBC and chosen file formats. The implementation has been thoroughly tested and is proven to work correctly as a part of Manta Flow, which features the plugin in the latest official release.