Data lineage forms an essential aspect of today's enterprise environment. MANTA Flow is a data lineage analysis platform that works based on extracting and analyzing customers' source files. However, often the customer wants to update the data lineage graph because of a slight change in provided source files. However, all of the input source files are currently reanalyzed, and most of the time is wasted analyzing unchanged files. In the thesis, we presented how the data lineage analyzer can be improved using incremental updates to analyze only a fraction of all input files while still producing the same correct data lineage.

We changed how the whole analysis is done by changing the granularity of the analysis to much smaller pieces. We also improved the merge algorithm to recognize when an unchanged file could generate a different data lineage using new concepts like source segments, node removal, or node creation. The new MANTA client algorithm now analyzes only changed files and a few unchanged files that could generate a different lineage compared to the last analysis. We also implemented a prototype for the MANTA Oracle scanner that contains these new ideas. It was tested for both the correctness and the performance.