ABSTRACT

Many companies still don‘t know the importance of data quality for the company‘s improvement. The number of companies in Indonesia, especially BUMN and Government companies have one application with one database, then there is a problem when it will be integrated one application with other applications related to duplication of data and the spread of data between columns, tables and applications. This problem can be handled by doing the data preprocess, one of the data preprocess method is data profiling. Data profiling is the process of gathering information that can be determined by process or logic. The process of profiling data can be done with various tools both paid and open source tools, each has advantages both in performance and in data processing according to the desired case study. In this study, the main focus on data analysis by conducting data profiling using deduplication and outliers methods. The results of the profiling will be implemented in logical form in open source application and will do comparisons between open source applications.

Keywords: data profiling, open source, data outliers, data duplication