ABSTRACT

In today's technological developments, data has an important role to support the achievement of goals for the company. The importance of data for companies is to fulfill the quality of supporting the company's business needs. The high quality that data has is of critical value to the company. However, there are many errors in the data that reduce the quality. Low-quality data, i.e. the data is inaccurate, incomplete, or out of date. There is a need for data quality management or Data Quality Management to manage data quality improvements to become consistent, accurate, complete, timely, and unique data. Regulating the quality of the data requires data cleansing. Data cleansing is a method to improve low-quality data by producing high-quality data. Therefore, this study will discuss the analysis and design of the decomposition of process packages for data cleansing to improve the quality of data that does not meet the company's needs. This research design uses the iterative incremental method, where iterative incremental is a development method that consists of activities that are arranged sequentially and are flexible in the changes that will occur in the development process. In the design carried out, the authors provide several solutions based on analysis to meet the needs of the company's data cleansing process flexibly with the decomposition of the process package that will be implemented using open source tools, namely Pentaho Data Integration as the final result of this research. There is a comparative evaluation using OpenRefine which results in data cleansing using Pentaho Data Integration which is superior in the overall data cleaning process.

Keywords— ***data cleansing, pentaho data integration, openrefine, iterative incremental***