ABSTRACT

ANALYSIS AND DESIGN DATA CLEANSING: CLUSTERING & DEDUPLICATION USING OPEN SOURCE TOOLS

By

DWI CAHYA SETYAWAN

NIM: 1202164140

Currently the data can be regarded as an asset that is needed by a company. Data as an important factor in making a decision, interaction with customers, to predict the future. But often the amount of data contained in a company is not balanced with good data quality, ranging from differences in data formats to errors in the data input process so that decision making does not produce maximum results. One technique for maintaining data quality is data cleansing. That is a process starting from analyzing the quality of data by changing, correcting, or deleting wrong data, incomplete, inaccurate, or has the wrong format. Data cleansing can be executed with free and paid applications. In this study, researchers conducted an analysis and design of a data cleansing architecture using clustering and deduplication methods which will be implemented using an open source tool.

Keywords: data cleansing, open source, data clustering, data deduplication, data quality management