ABSTRACT

Along with the rapid development of technology, the need for information has also increased drastically. The increasing need for information is also proportional to the growth of data production and data consumption every day. The production data can become uncontrollable and can produce a lot of noise which can cause dirty data and affect the performance of data processing and business processes within the company. This happened to PT XYZ as one of the largest telecommunications companies in Indonesia. The problem experienced is the duplication and redundancy of contract customer account data. This problem can be solved by performing data cleaning and requires a tool that has high capabilities to perform large data processing. The tools used in this research are Apache Spark with SparkSQL Library. The implementation of the data cleaning design that will be carried out in this study uses the waterfall method which is a development method with a sequential system. This method is used because this case study already has clear needs and problems. In the design carried out, the author divides several stages of data cleaning to solve the problem. These stages are define, design, code, and testing. All designs and tests in this research also follow the company's business rules. In addition, there is an evaluation of Apache Spark's performance with a comparison of the data cleaning process time using Oracle SQL Database which results in Apache Spark being superior by producing 374% faster time.

Keyword: ***data cleansing, apache spark, waterfall, oracle sql database, SparkSQL***