ABSTRAK

Seiring dengan perkembangan teknologi yang semakin cepat, kebutuhan informasi juga meningkat secara drastis. Peningkatan kebutuhan informasi juga sebanding dengan pertumbuhan produksi data dan konsumsi data setiap harinya. Peningkatan produksi data tersebut dapat menjadi tidak terkendali dan dapat menghasilkan banyak *noise* yang dapat menimbulkan adanya data yang kotor dan berpengaruh terhadap performa pengolahan data maupun proses bisnis dalam perusahaan. Hal tersebut terjadi pada PT XYZ yang menjadi salah satu perusahaan telekomunikasi terbesar di Indonesia. Permasalahan yang dialami adalah adanya duplikasi dan *redundancy* data akun kontrak pelanggan. Permasalahan tersebut dapat diatasi dengan melakukan *data cleansing* dan membutuhkan *tools* yang memiliki kemampuan tinggi untuk melakukan *processing* data berskala besar. *Tools* yang digunakan pada penelitian ini adalah Apache Spark dengan Library SparkSQL. Perancangan implementasi *data cleansing* yang akan dilakukan pada penelitian ini menggunakan metode *waterfall* yang merupakan metode pengembangan dengan sistem berurutan. Metode ini digunakan karena studi kasus ini sudah memiliki kebutuhan dan permasalahan yang jelas. Pada perancangan yang dilakukan, penulis melakukan pembagian beberapa tahap *data cleansing* untuk bisa menyelesaikan masalah tersebut. Tahapan tersebut adalah *define*, *design, code*, dan *testing*. Seluruh rancangan dan pengujian pada penelitian ini juga mengikuti *business rules* yang dimiliki oleh perusahaan. Selain itu, terdapat evaluasi performa Apache Spark dengan komparasi waktu proses *data cleansing* menggunakan Oracle SQL Database yang menghasilkan bahwa Apache Spark lebih unggul dengan menghasilkan waktu proses 374% lebih cepat.

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