ABSTRAK

Data dengan kualitas yang baik merupakan aset berharga bagi suatu perusahaan. Data dapat diproses menjadi informasi untuk membantu perusahaan meningkatkan pengambilan keputusan. Seiring berjalannya waktu, data yang dimiliki oleh perusahaan akan semakin bertambah banyak. Akan tetapi, semakin banyak data dapat meningkatkan kecenderungan untuk timbul permasalahan pada kualitas data. Dengan demikian, manajemen data yang baik penting untuk menjaga kualitas data dalam memenuhi standar perusahaan. Salah satu upaya yang dapat dilakukan adalah melakukan pembersihan data untuk membersihkan data dari kesalahan, ketidakakuratan, duplikasi, perbedaan format, atau anomali lainnya. Penelitian ini akan membahas tentang penerapan *data cleansing* menggunakan metode Analytics Canvas terhadap data akun pelanggan di perusahaan telekomunikasi. *Data cleansing* akan diterapkan pada *dataset* akun pelanggan dan *dataset* tagihan pembayaran dengan jumlah puluhan juta baris menggunakan Apache Spark modul SparkSQL untuk memperoleh performa pemrosesan *query* yang baik. Terdapat tiga tahapan dalam penelitian ini, yakni *preprocessing stage*, *processing stage*, dan *validation stage*. Selain itu, penelitian ini juga mengulas performa Apache Spark dalam memproses *query*. Dalam penelitian ini, kinerja Spark dan Oracle akan dibandingkan berdasarkan waktu pemrosesan *query*. Keduanya akan diuji pada *query* pembersihan data dan *query* tahapan validasi hasil *data cleansing*. Setelah penerapan *data cleansing* selesai, diperoleh hasil berupa *dataset* berkualitas yang berisi akun pelanggan dengan jumlah 30% dari total *dataset* awal. Hasil penelitian lainnya adalah adanya perbedaan dalam waktu pemrosesan *query* pada kedua alat. Apache Spark dinilai lebih baik karena memiliki waktu pemrosesan *query* yang relatif lebih cepat daripada Oracle Database. Dapat disimpulkan bahwa Oracle lebih dapat diandalkan dalam hal menyimpan model data yang kompleks daripada dalam melakukan analisis data besar. Untuk penelitian di masa depan, penelitian ini dapat digunakan sebagai dasar untuk kebutuhan optimalisasi *query* sehingga *query* yang paling efektif dapat diperoleh dengan waktu pemrosesan tercepat.

Kata kunci—***akun pelanggan, query, data cleansing, Apache Spark, komparasi***