ABSTRAK

ANALISIS DAN PERANCANGAN *DATA CLEANSING: CLUSTERING & DEDUPLICATION* MENGGUNAKAN *OPEN SOURCE TOOLS*

Oleh

DWI CAHYA SETYAWAN

NIM: 1202164140

Saat ini data dapat dikatakan sebagai aset yang sangat dibutuhkan oleh suatu perusahaan. Data sebagai faktor penting dalam pengambilan suatu keputusan, interaksi dengan pelanggan, hingga memprediksi masa depan. Namun seringkali banyaknya data yang terdapat pada suatu perusahaan tidak diimbangi dengan kualitas data yang baik, mulai dari perbedaan format data hingga kesalahaan pada proses input data sehingga pengambilan keputusan tidak membuahkan hasil yang maksimal. Salah satu teknik untuk menjaga kualitas data ialah dengan *data cleansing. Data cleansing* merupakan proses menganalisa kualitas dari suatu data dengan cara mengubah, mengoreksi, atau menghapus data-data yang salah, tidak lengkap, tidak akurat, atau memiliki format yang salah. *Data cleansing* dapat dieksekusi dengan aplikasi gratis maupun berbayar. Pada tugas akhir ini, penulis melakukan analisis dan perancangan arsitektur *data cleansing* dengan menggunakan metode *clustering* dan *deduplication* yang akan diimplementasikan menggunakan *open source tool*.

Kata kunci: *data cleansing, open source, data clustering, data deduplication, data quality management*