ABSTRAK

Banyak perusahaan yang masih belum mengetahui pentingnya kualitas data untuk kemajuan perusahaan. Banyaknya perusahaan di Indonesia terutama perusahaan BUMN dan Pemerintah memiliki satu aplikasi dengan satu database, maka terjadi masalah ketika akan diintegrasikan satu aplikasi dengan aplikasi lain terkait duplikasinya data dan penyebaran data antar kolom, tabel dan aplikasi. Permasalahan ini dapat ditangani dengan dilakukannya data *preprocess*, salah satu metode data *preprocess* yaitu data *profiling.* Data *profiling* merupakan proses pengumpulan informasi yang dapat ditentukan sesuai proses atau logika. Proses data *profiling* dapat dilakukan dengan berbagai *tools* baik yang berbayar maupun *open source tools,* masing-masing mempunyai keunggulan baik secara performa maupun dalam pengolahan data sesuai studi kasus yang diinginkan. Dalam penelitian ini, fokus utama pada analisis data dengan melakukan data *profiling* menggunakan metode *deduplication* dan *outliers*. Hasil dari *profiling* akan diimplementasikan dalam bentuk logika pada aplikasi *open source* dan akan dilakukannya komparasi antar aplikasi *open source*.

Kata kunci: data *profiling, open source,* penyebaran data, duplikasi data