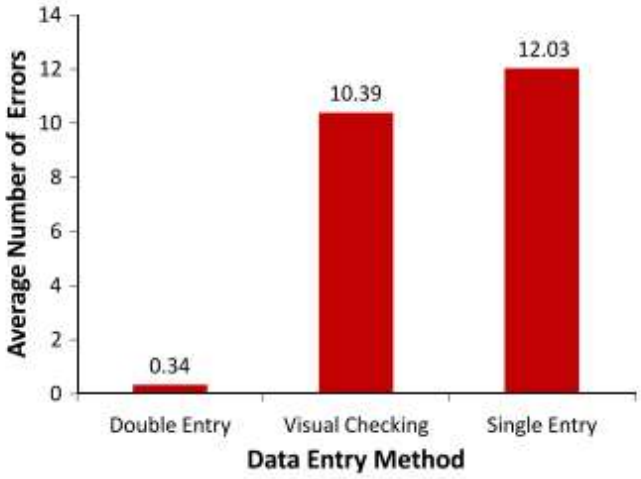
Bab I PENDAHULUAN

I.1 Latar Belakang

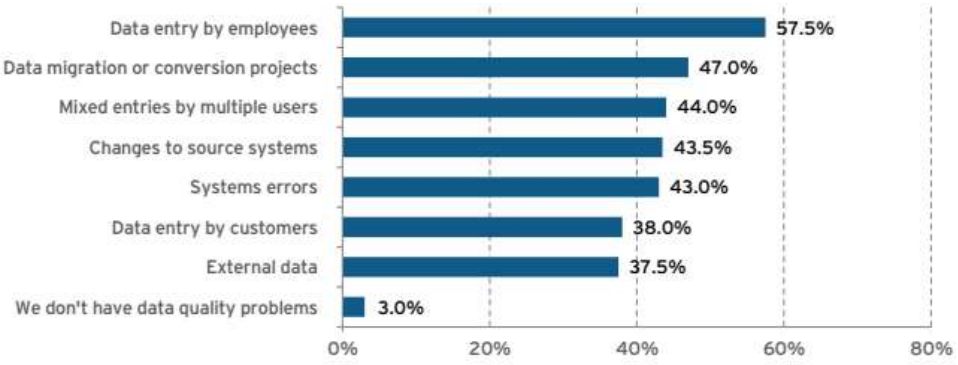
Data merupakan komponen penting di sebuah perusahaan. Banyak perusahaan yang baru menyadari bahwa kualitas data dapat menyebabkan sebuah keuntungan baik di segi waktu dan biaya. Kualitas data yang baik harus akurat, relevan, lengkap dan mudah dipahami. Kurangnya pengelolaan konten data dapat terjadi sebuah kerugian pada perusahaan, maka saat ini banyak perusahan mulai mencari sebuah alat untuk membantu mengoptimalkan konten agar kualitas data dapat sesuai yang diinginkan perusahaan (Olson, 2003).

Hasil analisis oleh Barchard & Pace pada tahun 2011 (Barchard & Pace, 2011) yang dilakukan pada 195 orang secara acak dapat membuktikan bahwa terdapat rata – rata kesalahan pada pemasukan data satu kali tanpa dilakukan pengecekan kembali yaitu berkisar 12.03, sementara ketika melakukan pemasukan data dua kali yang berarti dapat dilakukan pengecekan data yaitu berkisar 0.34.



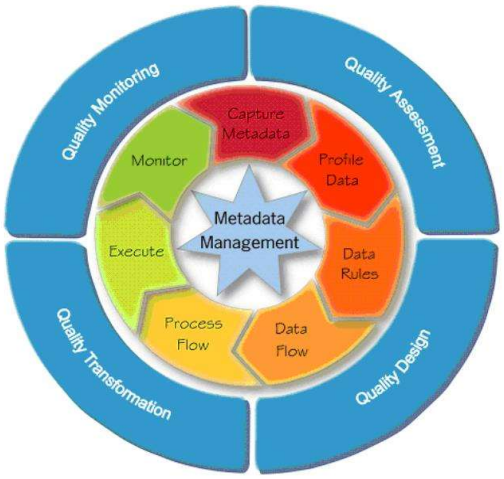
Gambar I-1 Rata - Rata kesalahan pemasukan data oleh manusia (Barchard & Pace, 2011)

Dengan demikian bahwa pemasukan data oleh manusia sangat banyak terjadi kesalahan data yang tidak sesuai. Selain itu perusahaan Blazent meninjau pada tahun 2016 mengenai masalah kualitas data dikarenakan pemasukan data oleh manusia (*human error)*. Pada Gambar I-2 terlihat penyebab kualitas data yang buruk dikarenakan pemasukan data oleh pihak pegawai atau internal dari perusahaan yang dilakukan secara bersamaan (Lehmann, Roy, & Yo, 2016).



Gambar I-2 Penyebab kualitas data yang buruk (Lehmann et al., 2016)

Pengelolaan sebuah kualitas data yang baik mengacu kepada beberapa proses (Corporation, 2009) diantaranya *Quality Assessment, Quality Design, Quality Transformation* dan *Quality Monitoring*, pada Gambar I-3 terdapat proses *Quality Assessment*. *Quality Assessment* pada tahap ini yaitu dengan memuat sumber data, yang bisa disimpan dalam beberapa sumber yang berbeda, setelah dimuatnya sumber data baru dilakukan dengan menggunakan data *profiling.* Data *profiling* merupakan langkah awal bagi sebuah perusahaan untuk meningkatkan kualitas informasi dengan menggunakan proses *ETL* (Corporation, 2009).



Gambar I-3 Tahapan melakukan pengelolaan kualitas data (Corporation, 2009)

Penelitian sebelumnya oleh Tien dan Fitria (Kusumasari, 2016) dengan menggunakan *open source tool* dari Google yaitu OpenRefine pada salah satu kasus

di BPOM dimana data yang diolah berupa Nomer Ijin Edar dan Nama Perusahaan, penelitian ini diterapkannya aturan bisnis yaitu Nomor Ijin Edar tidak boleh kosong, harus unik setiap entitas dan memiliki kesamaan pola alfanumerik. Hasil dari penelitian menunjukan Nomor Ijin Edar memiliki 70 pola pada 5000 baris data. Analisis duplikasi perlu dikombinasikan dengan elemen lain dikarenakan satu produksi dengan satu Nomor Ijin Edar bisa diduplikasi jika lokasi pabrik, volume dan berat paket berbeda (Kusumasari, 2016).

Column Name	Duplicates (%)	Blank(%)	Cluster	Pattern
NIE	46	0	2	70
Company name	79	1	120	-

Gambar I-4 Hasil analisis profiling menggunakan OpenRefine (Kusumasari, 2016)

Dalam penelitian ini, banyaknya perusahaan di Indonesia, khususnya BUMN dan Pemerintah yang memiliki satu aplikasi dengan satu database sehingga ketika diintegrasikan antar aplikasi terjadi duplikasi data baik antar kolom, tabel maupun *database*. Karena setiap aplikasi memiliki *database* sendiri dapat menyebabkan data yang tidak relevan ketika menentukan *business rule* perusahaan karena kesalahan standardisasi perusahaan yang akan berdampak pada data berkualitas buruk. Kualitas data yang buruk akan mempengaruhi *data governance*. *Data governance* adalah perencanaan, pengawasan, dan kontrol atas pengelolaan data dan penggunaan data dan sumber daya yang terkait dengan data (Cupoli, Earley, Henderson, & Deborah Henderson, 2014). *Data governance* melibatkan proses dan kontrol untuk memastikan bahwa informasi pada masing-masing karakter yang dikumpulkan dan di-input oleh organisasi adalah benar dan akurat, dan unik. Ini melibatkan pembersihan data, data yang tidak akurat, atau data asing dan deduplikasi data, untuk menghilangkan data yang berlebihan (Yulfitri, 2016).

Sehubungan dengan masalah ini maka perlu data yang bersih karena *master data management* di mana untuk melakukan *data warehouse* yang diperlukan yang bersih, unik dan memiliki standarisasi yang seragam dalam satu organisasi, dengan jumlah alat yang memberikan solusi, *data profiling* yang diperlukan semakin baik

kualitas datanya. Penelitian ini menggunakan alat open source yang mengacu pada Google OpenRefine. Logika aplikasi yang akan diimplementasikan dalam perangkat open source akan dibandingkan dengan keputusan komparatif dalam menentukan *open source tools*.

I.2 Perumusan Masalah

Berdasarkan uraian masalah yang telah dijelaskan pada latar belakang, maka permasalahan yang akan dikaji pada penelitian ini adalah sebagai berikut:

1. Bagaimana hasil pengujian metode *multi column: outliers* dan *deduplication* menggunakan *open* source *platform* untuk *data profiling*?

2. Bagaimana proses pengujian algoritma yang sesuai dengan *business rule* perusahaan untuk *multi column: outliers* dan *deduplication*?

I.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang ada, tujuan yang ingin dicapai dari penelitian ini dengan mengetahui baik secara metode maupun algoritma untuk *multi column: outliers* dan *deduplication* menggunakan *open source platform* yang nantinya diolah pada *data governance* maupun *data warehouse.*

I.4 Batasan Penelitian

Adapun batasan dalam melakukan penelitian ini adalah sebagai berikut:

1. Dataset menggunakan data Badan Pemerintah Indonesia Tahun 2017 dan Tahun 2018.
2. Metode *outliers* hanya dapat menerima data *integer* atau numerik selain itu proses *outliers* akan tidak dianggap.
3. Metode *outliers* menggunakan model sehingga ketika dataset berubah maka diperlukan konfigurasi ulang model.
4. Setiap metode hanya dapat memilih dua kolom per tabel untuk dilakukannya *profiling*.
5. Implementasi logika analisis menggunakan *tools* Pentaho Data Integration (Kettle) dan Weka Explorer.

I.5 Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini meliputi manfaat secara keilmuan. Manfaat keilmuan yang diharapkan dapat menerapkan analisis *multi column* dan mengimplementasikan *open source platform*. Manfaat lainnya dengan bekontribusi mengoptimalkan kualitas data dengan data *profiling*.

I.6 Sistematika Pelaporan

a) BAB I – PENDAHULUAN, bab ini berisi penjelasan mengenai latar belakang, rumusan masalah, tujuan penelitian, manfaat penelitian dan sistematika laporan.

b) BAB II – TINJAUAN PUSTAKA, berisi penjelasan kajian – kajian literatur pendukung untuk riset dan beberapa *related work* yang pernah dilakukan oleh peneliti sebelumnya.

c) BAB III – METODE PENELITIAN, berisikan penjelasan mengenai konseptual dan sistematika penelitian yang digunakan pada riset yang dilakukan.

d) BAB IV – ANALISIS DAN DESAIN, berisi tentang perhitungan sebuah model analisis yang digunakan untuk pengambilan keputusan.

e) BAB V – IMPLEMENTASI DAN PENGUJIAN, berisi tentang implementasi pembuatan logika, pengujian, menganalisa dari hasil analisis dan evaluasi.

f) BAB VI – KESIMPULAN DAN SARAN, bab ini menyimpulkan hasil dari penelitian yang dilakukan dan saran yang dapat dipertimbangkan untuk penelitian berikutnya.