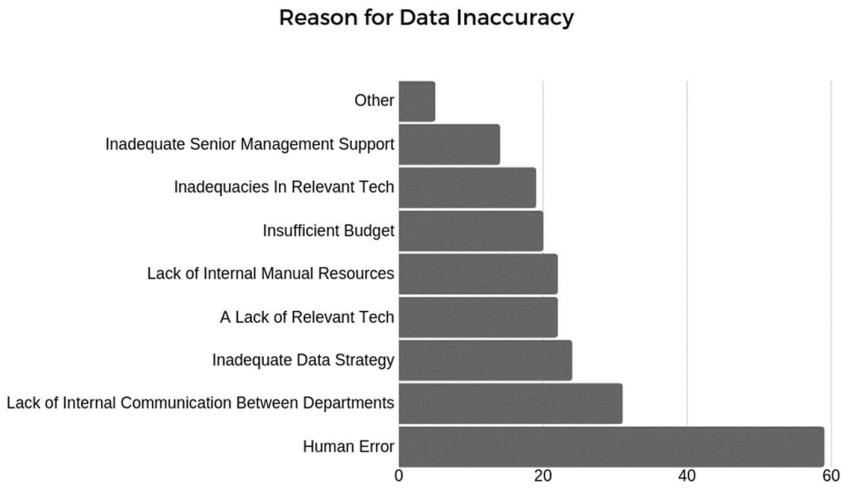
Bab I Pendahuluan

I.1 Latar Belakang

Salah satu komponen penting dalam perkembangan teknologi informasi saat ini ialah data. Data dapat mempengaruhi suatu organisasi dalam mengambil suatu keputusan. Pada dasarnya sebuah data dapat menjelaskan fakta dan angka – angka yang diolah oleh suatu perusahaan setiap hari. Semakin berkembangnya informasi akan menimbulkan semakin banyaknya format dan bentuk data yang dihasilkan, untuk menjaga kualitas data diperlukan kemampuan untuk menganalisis dan mengelola data agar dapat menciptakan sebuah pengetahuan/*knowledge* yang dapat berguna dalam meningkatkan nilai bisnis suatu perusahaan (Hellerstein, 2008).

Namun dengan banyaknya data mengakibatkan beberapa masalah muncul seperti ketidakakuratan data. Penyebab utama dari ketidakakuratan data paling banyak disebabkan oleh kesalahan pada manusia, kesalahan manusia ini dapat menyebabkan efek domino ketika data telah diterima dan diakses oleh departemen lain dalam suatu perusahaan. Hal ini juga menjadi salah satu penyebab utama dari ketidakuratan data yang terjadi (Desai, 2014).



Gambar I-I Grafik penyebab data tidak akurat (Experian Information Solution, Inc, 2013)

Menurut penelitian yang dilakukan oleh Thomas Schutz, seorang General Manager pada perusahaan North American Operations Experian Data Quality menunjukkan bahwa 91 persen dari perusahaan menderita kesalahan data umum. Kesalahan data yang paling umum adalah data yang tidak lengkap atau hilang. Data yang tidak akurat akan sangat mempengaruhi perusahaan dalam pengambilan keputusan dan proses pengolahan data selanjutnya.

Pada tahun 2019, Presiden Joko Widodo mengeluarkan Peraturan Presiden Republik Indonesia No. 39 Tahun 2019 tentang Satu Data Indonesia. Satu Data Indonesia merupakan sebuah inisiatif dari pemerintah dalam rangka memperbaiki dan meningkatkan kualiatas tata kelola data dalam lingkungan pemerintahan, karena pemerintah sadar pemanfaatan data saat ini dan kedepannya akan sangat penting selain untuk pengambilan kebijakan juga sebagai bentuk pemenuhan kebutuhan data bagi publik serta meningkatkan transparansi dan akuntabilitas pemerintah.

Pada tugas akhir ini, studi kasus yang diangkat adalah sebuah badan pemerintah (BPOM) yang bertanggung jawab dalam pengecekan keaslian sebuah makanan, minuman, obat, dan kosmetik di Indonesia. Masalah yang terjadi yaitu banyak terjadinya kesalahan *input* data yang menyebabkan data yang tersimpan dalam *database* memiliki format yang berbeda-beda, tidak terisi / *null,* hingga duplikasi data. Selain itu perubahan aturan bisnis yang kerap kali terjadi juga mempengaruhi format pada data. Sehingga menjadikan kualitas data kurang baik dan terjadilah ketidakakuratan data. Kesalahan input seperti penamaan produk yang tidak seragam akan menimbulkan masalah salah satunya saat pengeksekusian *query.* Serta akan menimbulkan masalah berkelanjutan karena data yang masuk akan selalu meningkat dari waktu ke waktu. Jadi, diperlukan pengembangan *data quality management* untuk menjaga dan meningkatkan kualitas data agar lebih baik.

Data Cleansing merupakan solusi yang dapat digunakan untuk mengatasi permasalahan data yang umumnya terjadi di badan pemerintahan dan perusahaan

berskala *enterprise*. *Data cleansing* berfokus untuk identifikasi data yang melibatkan serangkaian aktivitas pada domain tertentu yang berguna untuk membersihkan ketidaksesuaian dalam basis data. Dalam hal bisnis untuk melakukan pembersihan pada data sangat memakan waktu, karena setiap perusahaan memiliki aturan bisnis yang berbeda, untuk mengatasi hal tersebut harus dilakukan analisis dan pemodelan mengenai aturan-aturan dalam proses penginputan data untuk menjaga kualitas data (Boselli, Cesarini, Mercorio, & Mezzanzanica, 2014).

I.2 Rumusan Masalah

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

1. Bagaimana pemodelan Algoritma *Clustering* dan *Deduplication* dalam pemrosesan *Data Cleansing* untuk *Data Quality Management?*
2. Bagaimana hasil komparasi dari implementasi algoritma dengan *open source tools* lain?

I.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dilakukan, adapun tujuan yang ingin dicapai dari penelitian ini adalah sebagai berikut:

1. Pemodelan Algoritma pemrosesan *Data Cleansing* menggunakan *open source tools*.
2. Implementasi algoritma untuk *Data Cleansing* dengan metode *data clustering* dan *data deduplication* menggunakan *open source tools*.

I.4 Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini terdapat 2 bagian yaitu manfaat secara keilmuan dan manfaat untuk Industri.

1. Manfaat Keilmuan

   Mampu mengimplementasikan arsitektur *data cleansing* yang diintegrasikan dengan konsep *data quality management* berbasis *open source*.

2. Manfaat Untuk Industri

   Mampu melakukan *data cleansing* agar dapat memperbaiki serta menjaga kualitas data yang ada pada perusahaan.

I.5 Ruang Lingkup

Adapun ruang lingkup dalam melakukan penelitian ini, adalah sebagai berikut:

1. Studi kasus yang digunakan untuk penelitian dilakukan pada BPOM.
2. Tugas akhir ini hanya dilakukan pada bidang *Data Cleansing.*
3. Implementasi penerapan *Data Cleansing* menggunakan *open source tool* Pentaho Data Integration.
4. Algoritma pada tugas akhir ini berfokus pada *data cleansing* dengan konsep *data clustering* dan *data deduplication.*
5. *Cleansing data deduplication* mengacu pada kolom nama pada tabel ereg_m_pabrik dan tabel ereg_m_trader.
6. *Cleansing* berfokus pada kolom nama.
7. Uji coba aplikasi hanya dilakukan pada *windows platform*.

I.6 Sistematika Penulisan

Sistematika penulisan ini terbagi menjadi beberapa bab dari pokok pembahasan, secara umum dapat dijabarkan sebagai berikut:

a. BAB I – PENDAHULUAN, bab ini berisi penjelasan mengenai latar belakang, rumusan masalah, tujuan penelitian, manfaat penelitian dan sistematika laporan.
b. BAB II – LANDASAN TEORI, berisi penjelasan kajian – kajian literatur pendukung untuk riset dan beberapa *related work* yang pernah dilakukan oleh peneliti sebelumnya.
c. BAB III – METODOLOGI PENELITIAN, berisikan penjelasan mengenai konseptual dan sistematika penelitian yang digunakan pada riset yang dilakukan.
d. BAB IV – ANALISIS SISTEM, berisi tentang model dari sistem atau penelitian yang akan dilakukan.

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.