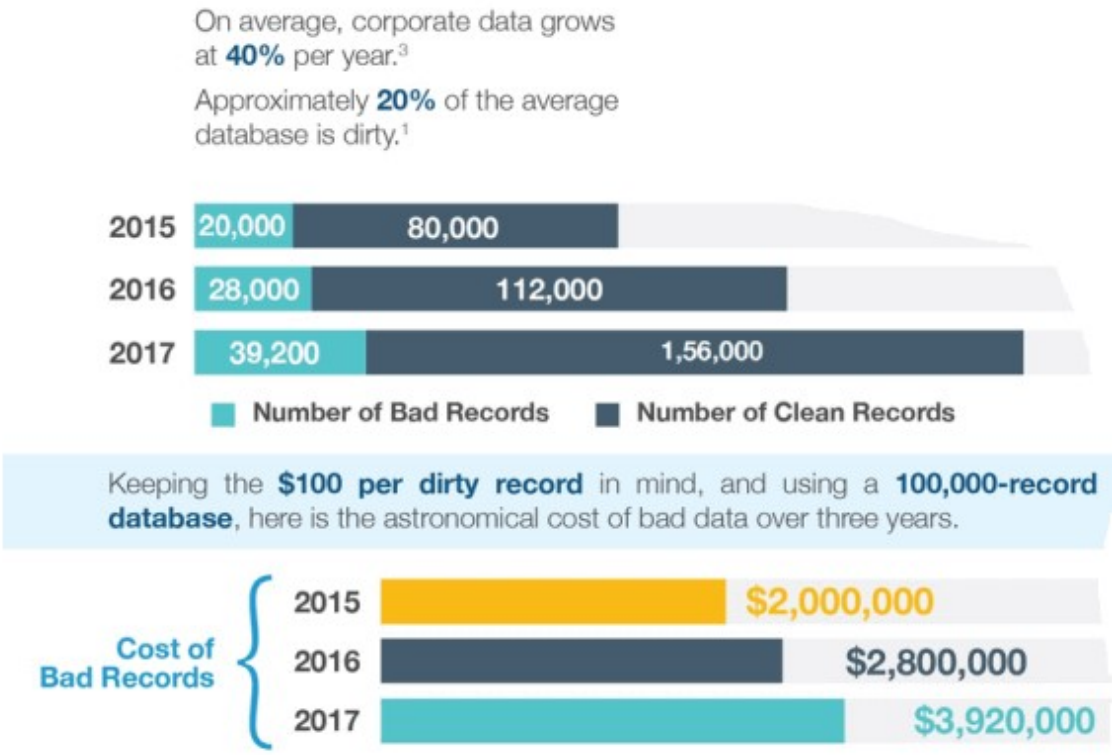
Bab I PENDAHULUAN

I.1 Latar Belakang

Data adalah fakta mentah yang belum diolah, yang sering kali tidak langsung dapat dimengerti oleh penerima data tersebut, maka dari itu data tersebut harus diolah terlebih dahulu sebelum akhirnya disajikan sebagai informasi untuk dapat diterima dan dimengerti dan dimanfaatkan oleh perusahaan untuk mendukung proses bisnisnya. Dengan demikian, data harus memiliki kualitas yang baik, sehingga dapat memberikan nilai terhadap perusahaan.

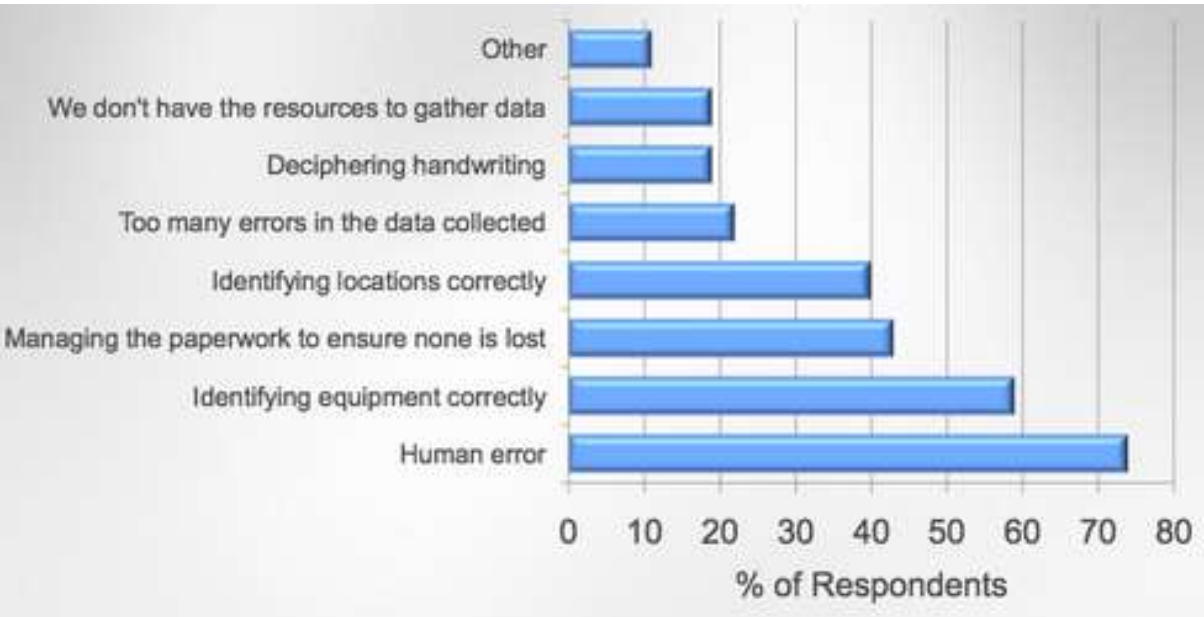
Kualitas data terus menjadi tantangan bagi banyak perusahaan saat perusahaan mencari cara untuk meningkatkan efisiensi dan interaksi pelanggan. 91% dari perusahaan mengalami data error, dimana yang paling umum adalah data yang hilang ataupun yang tidak lengkap dan ketidakakuratan data.



Gambar I-1 Grafik konsekuensi dari pengelolaan kualitas data yang buruk (RingLead Inc, 2015)

Buruknya kualitas data dapat mempengaruhi banyak aspek dari sebuah organisasi, antara lain seperti seberapa akurat perusahaan dapat memahami pelanggannya, seberapa akurat perusahaan dapat membuat keputusan dalam bisnis, berapa banyak biaya yang dikeluarkan untuk pemasaran serta seberapa efektif pemasaran tersebut (Lebied, 2017). Dalam Gambar I-

1, menjelaskan bahwa jumlah rata-rata data perusahaan naik hingga 40% per tahun, dan dari 2015 hingga 2017 jumlah data yang buruk naik cukup tinggi. Jika diibaratkan satu data yang memiliki kualitas buruk memiliki nilai setara seratus dollar, maka jumlah kerugian yang dialami perusahaan-perusahaan dalam tahun 2017 sebesar tiga juta sembilan ratus dua puluh ribu dollar (Ringlead, 2015).



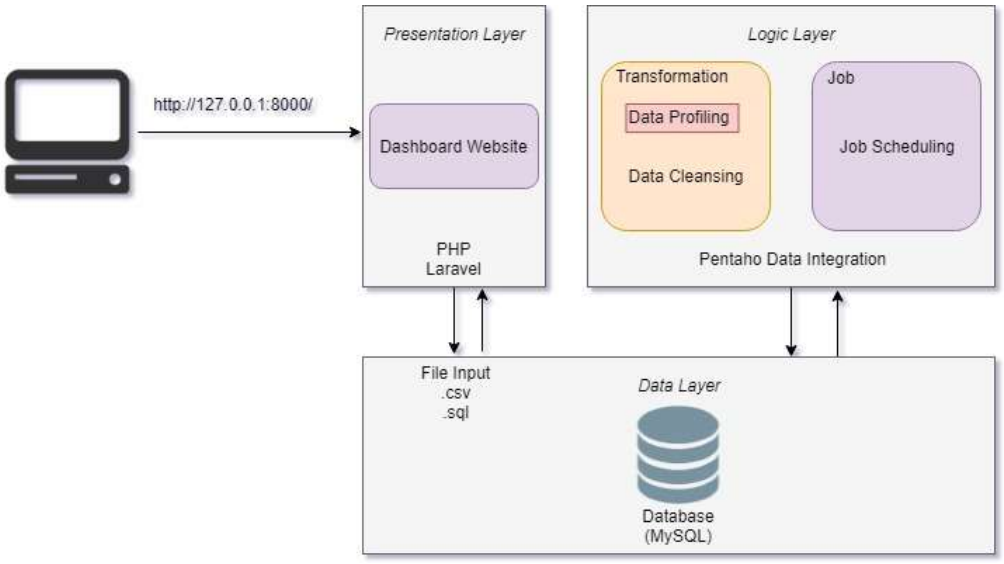
Gambar I-2 Grafik penyebab ketidakakuratan data (Watson, 2012)

Menurut penelitian yang dilakukan oleh Experian Information Solution, Inc pada tahun 2014, bahwa secara global jumlah data yang tidak akurat telah meningkat dari 17 persen menjadi 22 persen hanya dalam kurun waktu 12 bulan, dan perusahaan – perusahaan di Amerika percaya bawha mereka memiliki tingkat ketidakakuratan yang paling tinggi yaitu sebesar 25 persen. Alasan utama terjadinya ketidakakuratan data tersebut adalah kesalahan pada manusia atau *human error.* Tingkat ketidakakuratan data berkaitan dengan kurangnya strategi pengelolaan kualitas data (*Data Quality Management*) yang canggih (Desai, 2014).

Dalam pengelolaan kualitas data terdapat beberapa faktor antara lain, pemahaman teknis tentang proses pengumpulan data dan sumber. Selain itu, terdapat pembentukan proses pelaporan yang digunakan untuk memantau perubahan pada kualitas data serta peran, tanggung jawab untuk mengembangkan lingkungan yang mendukung peningkatan kualitas data (Phil, 2002)

Dalam mengelola kualitas data terdapat beberapa proses, yaitu *Quality Assessment, Quality Design, Quality Transformation* dan *Quality Monitoring.* Pada proses *Quality Assessment*, sumber data ditentukan kemudian data yang ada pada sumber tersebut dimuat. Setelah data dimuat, digunakanlah *data profiling* untuk menilai kualitas data tersebut. *Data profiling* adalah teknik analisa informasi pada data yang tersimpan di dalam basis data. *Data profiling* memastikan bahwa kualitas data tetap terjaga, untuk mengukur akurasi dan konsistensi data, dan mendeteksi duplikasi data dengan tujuan untuk memperoleh data dengan nilai yang tepat yang dapat digunakan dalam pemilihan keputusan. (Oracle Corporation, 2009)

Terdapat penelitian sebelumnya yang dilakukan oleh Febri Dwiandriani dan Alfi Nuri Khoirunisa menggunakan Pentaho Data Integration atau Kettle untuk *data profiling* dan *framework PHP Laravel* untuk integrasi web. Arsitektur pada aplikasi web tersebut menggunakan arsitektur *three-tier*, dimana arsitektur tersebut terdiri dari tiga bagian yaitu, *presentation layer* sebagai layer yang berhubungan dengan pengguna karena menampung tampilan aplikasi*, logic layer* sebagai layer yang berperan untuk menjalankan dan mengeksekusi aplikasi, dan *data layer* sebagai layer yang berperan dalam menampung sumber data yang digunakan. (Dwiandriani, 2017; Khoirunisa, 2017)



Gambar I-3 Arsitektur aplikasi web *Three-Tier* (Dwiandriani, 2017)

Pada penelitian ini, permasalahan yang sering terjadi terhadap kualitas data pada banyak perusahaan terutama di perusahaan BUMN adalah tidak tersaringnya data yang diinputkan, sehingga terdapat permalahan data mengenai data kosong, pola yang tidak terstandar dan penyebaran data yang tidak merata. SBerkaitan dengan permasalahan tersebut, maka data yang

bersih itu diperlukan untuk kepentingan *master data management*. Untuk memenuhi *master data management* diperlukan pembersihan data yang dimana data tersebut harus diketahui penyebab dari data tersebut kotor dengan menggunakan proses *data profiling*. Pada *data profiling* terdapat dua analisis yang dilakukan yaitu *single-column analysis* dan *multi-column analysis. Single-column analysis* adalah analisis terhadap satu kolom saja dalam satu waktu. *Single-column analysis* sendiri dibagi menjadi beberapa metode yaitu *cardinalities* untuk permasalahan data kosong, *data pattern* untuk permasalahan pola data yang tidak terstandar dan *value distribution* untuk permasalahan penyebaran data. Pada penelitian ini *tool* open source yang digunakan mengacu pada Google OpenRefine. Logika aplikasi yang akan diimplementasikan dalam perangkat open source akan dibandingkan dengan dengan *tool* open source lainnya.

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 implementasi metode *Single Column cardinalities, data pattern* dan *value distribution* dengan menggunakan platform berbasis *open source* untuk keperluan data *profiling*?

2. Bagaimana hasil komparasi algoritma untuk metode *Single Column cardinalities, data pattern* dan *value distribution* yang diimplementasikan dengan fungsi yang dimiliki oleh *tool open source*?

I.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang ada, tujuan yang ingin dicapai dari penelitian ini adalah sebagai berikut :

1. Implementasi metode *Single Column cardinalities, data pattern* dan *value distribution* menggunakan platform open source untuk keperluan *data profiling*.

2. Melakukan komparasi metode *Single Column cardinalities, data pattern* dan *value distribution* dengan fungsi *tool* open source.

I.4 Batasan Penelitian

Adapun batasan dalam melakukan penelitian ini adalah sebagai berikut :

1. Dataset yang digunakan adalah data Badan Pemerintahan Indonesia tahun 2018
2. Implementasi logika analisis menggunakan *tools* Pentaho Data Integration
3. Semua metode hanya dilakukan pada satu kolom.
4. C*ardinalities* hanya dapat dilakukan pada kolom yang memiliki tipe data Integer.

I.5 Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini adalah membantu permasalahan yang dihadapi perusahaan saat ini dalam pengelolaan kualitas data sehingga data dapat memberikan nilai maksimal bagi proses bisnis perusahaan. Manfaat keilmuan yang diharapkan adalah mampu memberikan kontribusi terhadap penambahan konsep baru dalam penerapan logika analisis *single column* dan platform berbasis *open source*

I.6 Sistematika Pelaporan

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, batasan penelitian, manfaat penelitian dan sistematika penelitian
b) BAB II – LANDASAN TEORI, bab ini berisi penjelasan kajian – kajian literature pendukung untuk riset, dan metode yang digunakan dalam penelitian yang dilakukan
c) BAB III – METODE PENELITIAN, bab ini berisi penjelasan mengenai model konseptual dan sistematika penelitian yang digunakan pada riset yang dilakukan.
d) BAB IV – ANALISIS DAN DESAIN, bab ini berisikan tentang perhitungan sebuah model analisis yang digunakan untuk pengambilan keputusan
e) BAB V – IMPLEMENTASI DAN PENGUJIAN, berisi tentang implementasi pembuatan logika, pengujian dan evaluasi
f) BAB VI – KESIMPULAN DAN SARAN, berisi tentang kesimpulan dari hasil penelitian yang dilakukan dan saran yang dapat dipertimbangkan untuk penelitian kedepannya.