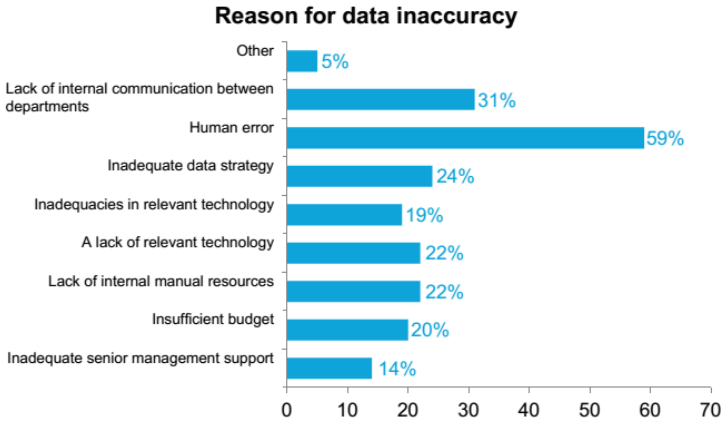
BAB I PENDAHULUAN

I.1 Latar Belakang

Data merupakan komponen penting dalam perusahaan yang dapat mempengaruhi tujuan dan pengambilan keputusan perusahaan tersebut. Pada dasarnya sebuah data menjelaskan fakta dan angka-angka yang diproses perusahaan setiap hari. Jika data dioperasikan setiap hari mengakibatkan banyaknya format dan bentuk data yang ada untuk diolah maka kemampuan untuk menganalisis dan mengelola sebuah data sangat penting. Selain itu sebuah data harus memiliki kualitas yang baik, yang menjadi faktor penentu penting dalam hal efektivitas suatu organisasi untuk memberikan nilai bisnis (MacDonald, 2011).

Dengan demikian kualitas data (*data quality*) menjadi sebuah tantangan bagi perusahaan untuk meningkatkan nilai bisnis. Sebuah riset oleh Experian Information Solutions, Inc tahun 2013 di Amerika Serikat menunjukkan terdapat 91% perusahaan menderita kesalahan data umum. Kesalahan yang sering ditemui adalah data tidak lengkap atau hilang, informasi usang dan data yang tidak akurat.

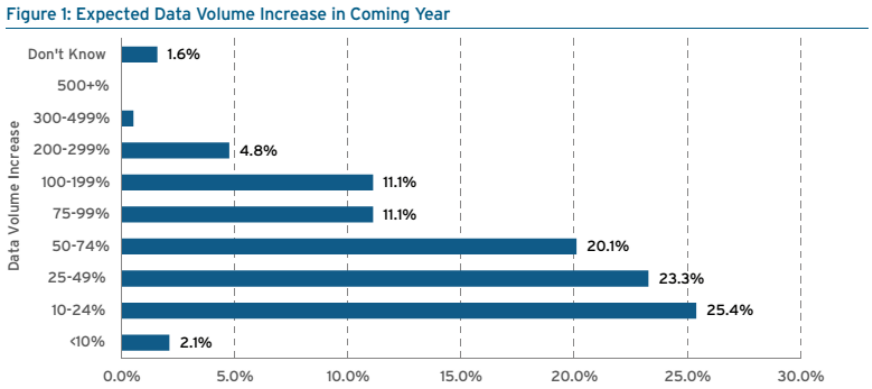


Gambar I-1 Grafik penyebab data tidak akurat di Amerika Serikat (Thomas, 2013)

Berdasarkan Gambar I-1 *human error* merupakan masalah utama sebagai alasan tidak akuratnya sebuah data. Sumber data yang berasal dari berbagai saluran (*channel*) banyak terkena kesalahan manusia saat proses memasukkan data (*entry data*). Secara kolektif di Amerika Serikat sebesar 78% dari perusahaan mempunyai masalah dengan kualitas data yang mereka ambil dari berbagai saluran (*channel*) (Thomas, 2013). Bahwasannya masalah yang terjadi membutuhkan sebuah

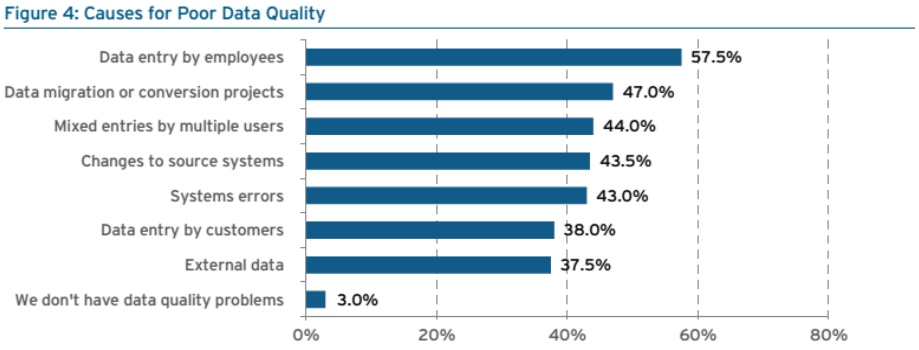
pengelolaan kualitas data (*data quality management*). Dalam pengelolaannya terdapat landasan dan pembangunan peran, tanggung jawab, aturan, dan prosedur untuk memastikan akuisisi, perbaikan, penyebaran dan disposisi dari data. Sehingga menjadikan *data quality management* bagian dari sebuah Big Data, yang akan diakses oleh banyak pengguna dan banyak *channel* data berbeda serta berelasi (Juddoo, 2015).

Selain itu melalui survei yang dilalukan oleh Blazent *company* kepada 200 IT *decision-makers* dan *influencers* memperkirakan bahwa kenaikan perkembangan data pada tahun yang akan datang sekitar 10% - 24% (Lehmann, Roy, & Winter, 2016).



Gambar I-2 Perkiraan kenaikan volume data tahun 2016 (Lehmann, Roy, & Winter, 2016)

Blaznet meninjau pula mengenai masalah kualitas data yang mayoritas dikarenkan oleh *human error*. Kesalahan terjadi dalam melakukan pekerjaan yang berkaitan dengan praktik IT seperti pemindahan data, perubahan sistem dan kesalahan sistem. Hasilnya menunjukkan bahwa tidak adanya hubungan antara tanggung jawab dan akuntabilitas mengenai kualitas data (Lehmann, Roy, & Winter, 2016).



Gambar I-3 Penyebab lemahnya kualitas data tahun 2016 (Lehmann, Roy, & Winter, 2016)

Pada Gambar I-3 terlihat penyebab lemahnya data entry paling dominan yaitu, proses pemasukan data oleh pihak internal perusahaan ketika mereka melakukan secara bersama-sama (Lehmann, Roy, & Winter, 2016). Oleh karena itu sebuah pengelolaan kualitas data (*data quality management*) memang dibutuhkan.

Dalam mengelola kualitas data terdapat beberapa proses, salah satunya adalah data profiling. Data profiling merupakan sebuah proses pencarian ketidaknormalan di dalam data yang merusak nilai data tersebut. Ketidaknormalan data mampu merusak sebuah aturan bisnis yang telah disesuaikan dengan kinerja aplikasi. Jika ditemukan satu ketidaknormalan dari ribuan data dalam *database* maka menyebabkan proses data profiling yang salah (Patel & Patel, 2012).

Gambar I-4 Runtunan melakukan data quality management (www.rcgglobalservice.com)

Data profiling mendasari bermacam-macam program manajemen informasi termasuk penilaian kualitas data, validasi kualitas data, manajemen metadata, integrasi data dan proses *extraction*, *transformation*, dan *loading* (ETL), perpindahan data dan modernisasi projek. Data profiling memungkinkan serangkaian analisis dan penilian algoritma yang bila diterapkan dalam konteks yang tepat dapat memberikan bukti untuk potensi masalah yang ada dalam sebuah *data set*. Keutamaan fungsi profiling terletak pada kemampuan meringkas rincian mengenai *data set* yang besar dari berbagai sudut (Loshin, 2012).

Oleh karena itu data profiling merupakan langkah awal pada *data quality management*, guna memahami seluruh kelayakan sumber-sumber data dan kualitas

setiap sumber data saat ini (Oracle Corporation, 2013). Dalam pendekatanya, proses data profiling dapat menggunakan sebuah *tools* yang mampu menangani hal ini dengan masukan yang berupa data yang kompleks dan beberapa *tools* membutuhkan spesifikasi yang kuat untuk melakukan tugas profil data (Abedjan, Golab, & Naumann, 2016).

Pada penelitian ini, terdapat sebuah Organisasi XYZ yang merupakan perusahaan dimana memiliki tanggung jawab dalam pengecekan keaslian sebuah makanan dan minuman, obat dan kosmetik di Indonesia. Kondisi Organisasi XYZ saat ini telah memiliki beberapa aplikasi eksisting untuk menunjang membantu melakukan pengecekan keaslian sebuah produk pangan, obat atau kosmetik. Namun sayangnya aplikasi yang diterapkan masih relatif bersifat independen dan belum ada pengecekan kesalahan pemasukan data pada aplikasi serta tiap aplikasi memiliki databasenya sendiri.

Dikarenakan setiap aplikasi terdapat *personal database* hal tersebut dapat mengakibatkan data yang dimasukkan pada database menjadi tidak relevan dan pada saat proses memasukkan data seringkali mengalami perbedaan standar penulisan yang pada umumnya merupakan kesalahan manusia. Kedua masalah tersebut akan menimbulkan masalah beruntun yang mengacu pada kualitas data. Diketahui bahwa Organisasi XYZ belum memiliki sebuah aplikasi personal yang digunakan sebagai pengecekan keakuratan data sebelum data tersebut diolah pada tahap selanjutnya. Akibatnya sering terjadinya redudansi antara data baru dengan data lama pada *database* yang kemudian ditampilkan ke halaman *website* organisasi. Sehingga membuat pihak internal maupun eksternal menyayangkan peristiwa tersebut. Untuk memberikan solusi pada kasus tersebut diperlukan adanya perancangan aplikasi untuk meningkatkan kualitas data pada Organisasi XYZ.

Berkaitan dengan kondisi Organisasi XYZ sedemikian rupa sehingga dibutuhkan sebuah *tool* yang mampu memberikan solusi yang mumpuni. Banyak *tool* atau aplikasi yang ditawarkan berbagai pihak, mulai dari berbayar hingga *open source*. Penelitian ini berpedoman pada aplikasi *open source*, salah satu acuannya adalah

OpenRefine dari Google. Penerapan aplikasi *open source* pada studi kasus data profiling Organisasi XYZ didukung dengan metode komparasi dan melakukan perbandingan aplikasi eksisting guna memberikan ketepatan analisa dalam pengambilan keputusan untuk penentuan aplikasi open source yang akan dijadikan *benchmark*. Untuk selanjutnya sebagai pengembangan dan penyesesuaian dengan kebutuhan perusahaan.

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 model arsitektur aplikasi untuk data profiling *tool* berbasis open source?
2. Seperti apakah pemodelan desain arsitektur aplikasi pada Organisasi XYZ jika berbasis open source?

I.3 Tujuan Penelitian

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

1. Mengembangkan arsitektur aplikasi untuk mendukung proses data profiling.
2. Dapat mengetahui arsitektur aplikasi menggunakan pendekatan OpenRefine yang disesuaikan dengan kebutuhan Organisasi XYZ

I.4 Batasan Penelitian

Adapula batasan dalam melakukan penelitian ini, sebagai berikut:

1. Penelitian hanya dilakukan pada lingkungan Organisasi XYZ.
2. Penelitian dilakukan mengenai proses data profiling.
3. Fitur data profiling yang diimplementasi berupa clustering string (text facet) pada kolom nama dan perhitungan data null/blank pada kolom telepon dari sebuah tabel uji, yaitu tabel pabrik dan tabel trader.
4. Algoritma clustering yang diterapkan pada penelitian menggunakan algoritma yang tersedia pada OpenRefine, yaitu fingerprint.

5. Untuk proses penarikan database dan tabel belum dinamis.
6. Pada tampilan monitoring perubahan nama tabel dan nama kolom belum dinamis.
7. Untuk implementasi Job scheduling diatur secara berulang dengan waktu setiap minggu untuk mengeksekusi perubahan data.

I.5 Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini meliputi manfaat secara keilmuan. Manfaat keilmuan yang diharapkan adalah mampu memberikan kontribusi terhadap penambahan konsep baru dalam perancangan arsitektur aplikasi berbasis *open source* pada proses data profiling. Manfaat lainnya yaitu melalui kontribusi perancangan aplikasi data profiling pada Organisasi XYZ.

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 – 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 model dan sistematika penelitian yang digunakan pada risetyang dilakukan.
d) BAB IV – ANALISIS DAN DESAIN, berisi tentang perhitungan sebuah model analisis yang digunakan untuk pengambilan keputusan dari *brenchmark* sebuah aplikasi, penggambaran arsitektur aplikasi usulan dan penggambaran desain diagram UML.
e) BAB V – IMPLEMENTASI DAN PENGUJIAN, berisi tentang implementasi pembuatan logika dan pembuatan aplikasi, pengujian, dan evaluasi.

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