BAB I PENDAHULUAN

I.1 Latar Belakang

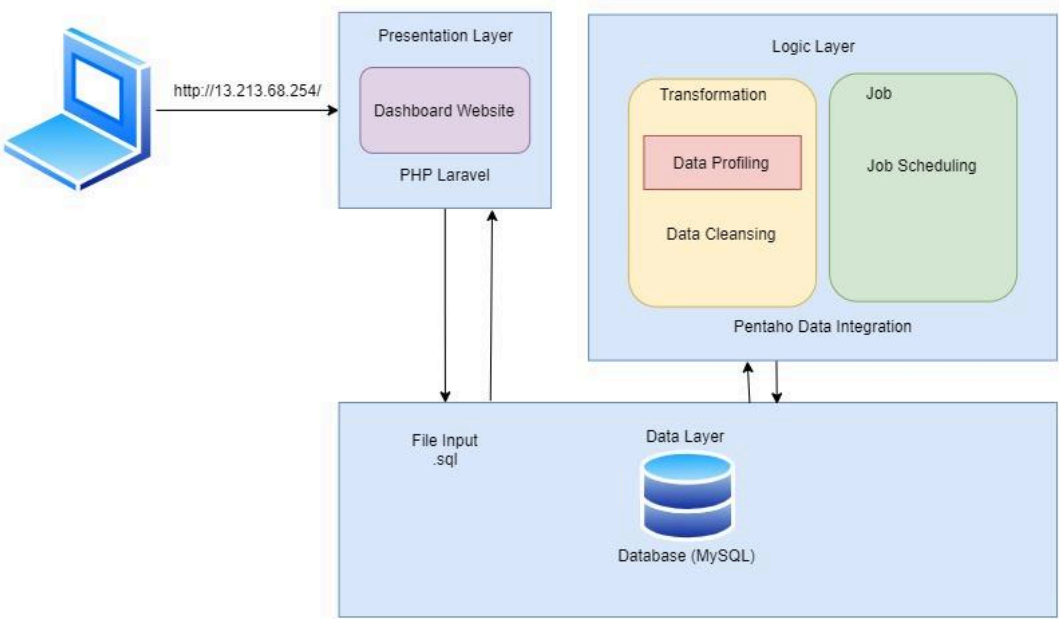
Data quality adalah level data yang menyatakan data tersebut akurat (*accurate*), lengkap (*complete*), terbaru (*update*), konsisten (*consistent*) sesuai dengan kebutuhan pada peraturan bisnis dan relevan (Mosley, Mark, 2008). Data yang berkualitas memiliki beberapa manfaat. Manfaatnya dapat berupa peningkatan kepercayaan diri dalam pengambilan keputusan, perbaikan pelayanan kepada *customer*, peningkatan kesempatan memperbaiki kinerja, mengurangi resiko dari keputusan yang salah, mengurangi biaya, meningkatkan produktivitas dengan memangkas beberapa proses, dan menghindari efek komplikasi dari data yang terkontaminasi.

Data quality memiliki karakteristik yaitu *accuracy* (data yang tersimpan sesuai dengan aslinya), *relevancy* (Sesuai dengan kebutuhan), *completeness* (berisi informasi yang lengkap), *timeliness* (data harus *up to date*), *consistency* (data dapat diandalkan), *granularity* (data harus cukup dan rinci), *uniqueness* (data yang ada tidak sama persis dengan data yang lain) (Makesh, 2020). Dalam mencapai *data quality* yang baik, masalah pada data harus diselesaikan terlebih dahulu. Permasalahan umum pada data yang banyak adalah pada saat memasukkan data pertama kali, baik secara manual ataupun otomatis. Terkadang data yang sudah ada akan menjadi masalah ketika di pindahkan kedalam *database* baru. *Duplicate* data juga menjadi salah satu masalah yang mempengaruhi *data quality*. Data yang sama lebih dari satu akan membuat data menjadi *redundancy*.

Seringkali data yang masuk dalam perusahaan tidak sesuai dengan standar aturan perusahaan. Penulisan masih ada yang *typo*, *uppercase* atau *lowercase* pada penulisan nama dan alamat masih sering berbeda – beda. Maka dari itu perlu dilakukan penyesuaian sehingga data yang ada menjadi data yang berkualitas. Data yang berkualitas dapat membantu perusahaan dalam pengambilan keputusan dan meningkatkan performa dari perusahaan. (Amethyst,2018)

Mengatasi masalah pada data yang banyak dapat dilakukan menggunakan *data profiling*. Secara umum, *data profiling* dibagi menjadi empat (*single column profiling, multi column, multi table, data rule validation*). *Single column profiling* adalah metode untuk menentukan distribusi frekuensi dan pola nilai data dalam suatu kolom data untuk menemukan *max of error*. *Multi column* adalah metode untuk menemukan kombinasi kolom dengan jumlah nilai yang ditentukan. Dalam *multi column* dapat menganalisis ketergantungan antara atribut data dalam tabel yang sama. *Multi table* adalah metode yang dapat membandingkan semua kolom disemua tabel yang dipilih. *Data rule validation* adalah metode untuk verifikasi data valid dalam aturan yang ada. (Ziawasch, Lukasz, & Felix, 2017)

Semakin lama perusahaan beroperasi, maka data yang ada pada perusahaan juga akan bertambah banyak. Data yang banyak terkadang mempunyai masalah dalam proses penginputannya. Dalam penelitian sebelumnya, dilakukan profiling menggunakan *single column* dan dilakukan perbaikan sehingga data dapat diproses dengan waktu performansi yang cepat. Tetapi, pada penelitian sebelumnya masih perlu dilakukan pengembangan lebih lanjut. Terutama pada klasifikasi *profiling*, *file input* yang hanya bisa digunakan untuk data dan database tertentu. (Karnia,2020)



Gambar I.1 Arsitektur Aplikasi *Three-Tier*

Arsitektur aplikasi *Three-tier* ini berisi tiga bagian utama yaitu *presentation layer, logic layer, data layer*. *Presentation layer* ini berfungsi untuk menampilkan aplikasi kedalam website agar dapat berinteraksi dengan *user*. *Logic layer* digunakan untuk menjalankan dan mengeksekusi aplikasi menggunakan pentaho data integration. *Data layer* berfungsi untuk menampung data yang digunakan.

Pada penelitian ini ditambahkan 3 *package* pada *single column* untuk melengkapi klasifikasi profiling penelitian sebelumnya. Klasifikasi profiling sebelumnya terdapat *single column* dan *multicolumn*. *Single column* terdapat *package value distribution, data completeness, show null,* dan *clustering*. *Multi column* terdapat *value similarity,* dan *data deduplication*. *Cardinalities, domain analysis*, dan *data type* akan ditambahkan pada *single column* untuk melengkapi klasifikasi profiling. Pentaho Data Integration dipilih karena mudah digunakan. Konsep Pentaho Data Integration menggunakan *drag and drop*.

I.2 Perumusan Masalah

Berdasarkan latar belakang yang telah disampaikan diatas, berikut ini merupakan permasalahan yang akan dikaji dalam penelitian ini:

1. Bagaimana proses penambahan klasifikasi dan analisis data *profiling* berdasarkan proses eksisting sebelumnya?
2. Bagaimana cara menambahkan proses *profiling* yang belum ada?

I.3 Tujuan Penelitian

Dengan adanya rumusan masalah, maka terdapat tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut :

1. Mengetahui proses klasifikasi dan analisis data *profiling* berdasarkan proses eksisting sebelumnya.
2. Mengetahui cara menambahkan proses *profiling* yang belum ada.

I.4 Batasan Penelitian

Adapun batasan dalam penelitian ini adalah sebagai berikut :

1. Dataset yang digunakan adalah data BPOM
2. Implementasi logika analisis menggunakan *tools Pentaho Data Integration*
3. Analisis yang dilakukan menggunakan *single column* berupa *cardinalities*, *domain analysis*, dan *data type*.
4. Analisis menggunakan satu kolom saja.
5. Analisis *cardinalities* hanya dapat digunakan untuk tipe data integer

I.5 Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini adalah membantu permasalahan yang telah dihadapi perusahaan dalam pengelolaan kualitas data, sehingga data yang ada dapat memberikan nilai maksimal bagi proses bisnis pengelolaan seluruh data diperusahaan. Kualitas data yang baik, dikelola berdasarkan proses klasifikasi yang terstruktur. Manfaat keilmuan yang diharapkan adalah dengan memberikan kontribusi terhadap penambahan konsep baru dalam analisis data menggunakan metode *profiling* dengan Pentaho Data Integration. Penelitian ini juga bermanfaat dalam proses implementasi *profiling* data dengan menggunakan berbagai macam kondisi.

I.6 Sistematika Penulisan

Penelitian ini diuraikan dengan sistematika penulisan sebagai berikut:

Bab I Pendahuluan

Pada bab ini dijelaskan mengenai konteks permasalahan yang diangkat, latar belakang penelitian, perumusan masalah, tujuan penelitian, batasan penelitian, manfaat penelitian, dan sistematika penulisan.

Bab II Tinjauan Pustaka

Pada Bab ini berisi literatur yang relevan dengan permasalahan yang mendukung penelitian, referensi dari penelitian yang terdahulu sebagai pedoman penyelesaian permasalahan.

Bab III Metodologi Penelitian.

Pada bab ini, menjelaskan pelaksanaan penelitian dengan gambaran metode konseptual dan sistematika penelitian.

Bab IV Analisis dan Perancangan

Pada bab ini, disajikan hasil rancangan, temuan, analisis dan pengolahan data.

Bab V Implementasi dan Pengujian

Pada bab ini, menjelaskan mengenai implementasi dan pengujian profiling menggunakan pentaho

Bab VI Kesimpulan dan Saran

Pada bab ini dijelaskan kesimpulan dari penelitian yang dilakukan serta saran penelitian yang dapat digunakan sebagai penelitian selanjutnya.