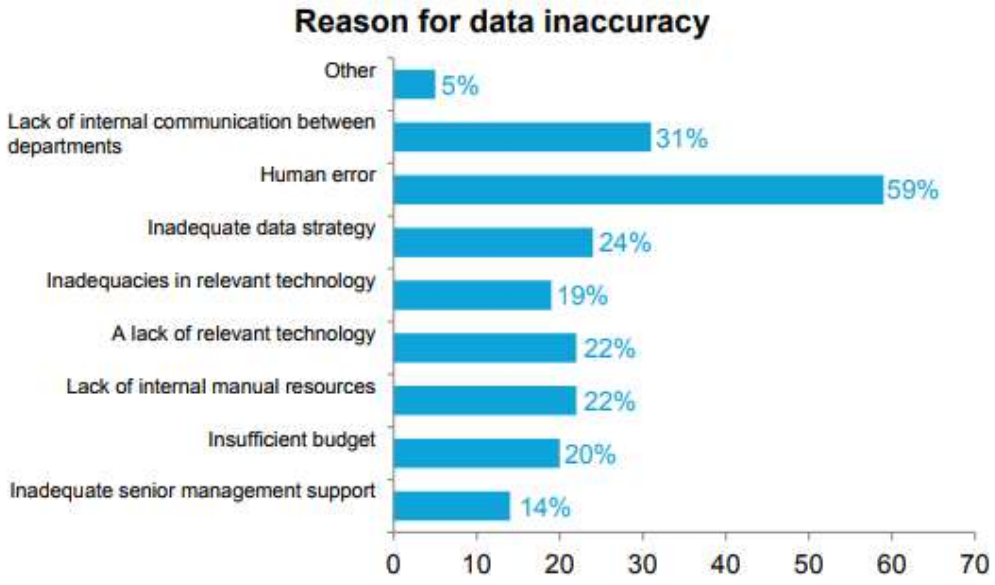
Bab I Pendahuluan

1.1 Latar Belakang

Data pada dasarnya merupakan fakta yang dikumpulkan pada suatu kegiatan, dalam konteks bisnis data merupakan fakta dan angka – angka yang di olah oleh suatu perusahaan setiap hari. Data yang sudah di olah oleh suatu perusahaan dapat berubah menjadi informasi dimana informasi tersebut dapat digunakan oleh perusahaan untuk melakukan analisis. Informasi yang saling berkaitan dapat menciptakan pengetahuan/*knowledge* yang akhirnya dapat digunakan untuk mendapatkan pembuatan keputusan dengan benar, namun untuk mendapatkan *knowledge* data harus sudah mengalami proses pengolahan yang panjang dilain itu kualitas data juga harus diperhatikan untuk mendapatkan hasil yang maksimal (MacDonald, 2011)

Kualitas data memiliki peranan yang sangat tinggi di dalam dunia bisnis, kualitas data merupakan suatu penilaian atau presepsi terhadap suatu data untuk digunakan dalam tujuan atau konteks tertentu, dimana terdapat beberapa aspek dalam kualitas data yaitu ketepatan data, kelengkapan data, umur data hingga reabilitas suatu data. Pada suatu organisasi tingkat kualitas data yang dapat diterima sangatlah berpengaruh terhadap proses operasional dan transaksional yang nantinya hal tersebut akan masuk ke dalam laporan pada *business analytics* dan *business intelligence* (Rouse, 2005)

Banyaknya data yang diolah memberikan informasi bahwa terdapat banyak data yang tidak relevan, menurut penelitian Experian Information Solutions, Inc menunjukkan bahwa rata – rata tingkat ketidakakuratan data pada organisasi di Amerika Serikat berada pada tingkat 25 persen, dan hingga 91 persen dari perusahaan yang ada di Amerika Serikat menderita kesalahan data umum dimana kesalahan tersebut.



Gambar I-1Grafik alasan dari ketidakakuratan data di Amerika Serikat (Desai, 2014)

Penyebab utama dari ketidakakuratan data masih di sebabkan oleh kesalahan pada manusia, yang dimana penyebab tersebut sudah menjadi penyebab utama dari ketidakuratan data dalam 3 tahun terakhir, dimana di ikuti dengan penyebab kedua yaitu strategi data yang tidak memadai (Desai, 2014)

Dalam pengelolaan kualitas data terdapat beberapa proses tata kelola data yang dapat mempengaruhi data yang ada di dalam suatu organisasi dimana salah satunya adalah *data quality management (DQM)* dimana DQM merupakan bagian penting dari keseluruhan strategi dalam tata kelola data. Dimana alur yang tersedia dari dalam DQM terdapat pada gambar dibawah (IBM, 2007)

Gambar I-2 Alur strategi data quality management (www.rcgglobalservice.com)

Menurut wikipedia Data *profiling* merupakan salah satu proses pemeriksaan data yang tersedia dari suatu sumber informasi dan mengambil statistik serta informasi terhadap data tersebut (Wikipedia, 2016)

Penggunaan data *profiling* juga dibutuhkan sebelum data dapat masuk ke proses *cleansing* dimana dengan tujuan untuk menguji kualitas data dengan mendeteksi apakah sumber data yang ada sudah mengikuti *business rule* atau belum, data *profiling* juga dapat dilakukan menggunakan banyak teknik analisis tergantung dari elemen data yang akan di analisis (Kusumasari & Fitria, 2016)

Data *profiling* dapat diraih dengan menggunakan *tools* yang terdapat secara berbayar maupun gratis, dalam hal ini beberapa *tools* yang dapat digunakan yaitu Talend, DataCleaner, WinPure, DataPreparator, Data Match, DataMartist, Pentaho Kettle, SQL Power Architect, SQL Power DQguru, dan DQ Analyzer, suatu penelitian oleh Venkata Sai Venkatesh Pulla dan Cihan Varol and Murat Al melakukan percobaan untuk melakukan komparasi antara *tools* tersebut berdasarkan fitur utama yang dimiliki dan performa dalam melakukan data *profiling, integration* dan *cleansing* (Sai, Pulla, Varol, & Al, 2016)

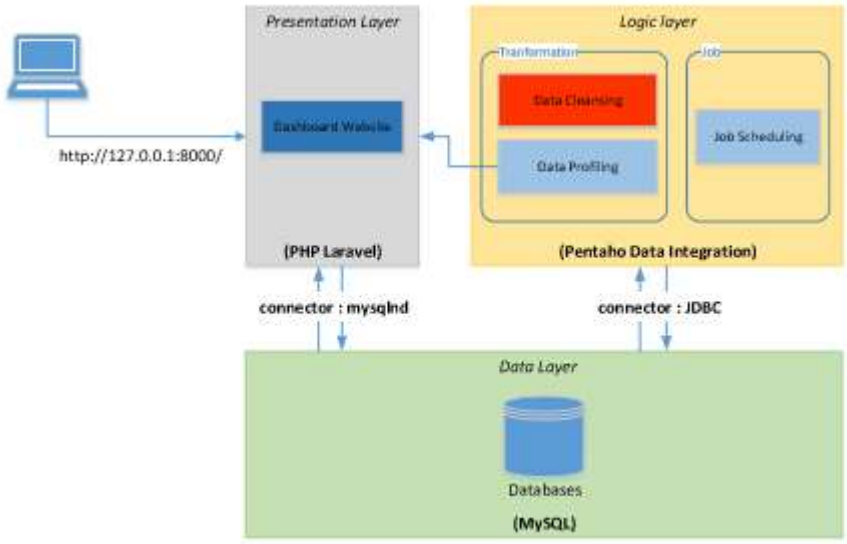
Tools	Data Source Connectivity	Dataset Mng.	Metadata Repository	GUI	Development Platform	Reporting
Talend Open Studio	All	ELT	Y	G	Perl/Java	T/Gr
Data Cleaner	JDBC	ETL	Y	G	Java	T/Gr
WinPure	JDBC	ETL	N	G	Java	T
Data Preparator	N	ELT	N	G	Java	T/Gr
Data Match	All	ELT	N	G	Java	T/Gr
DataMartist	All	ETL	N	G	Java	T
Pentaho Kettle	JDBC	ETL	Y	G	Java	T
SQL Power Architect	JDBC	ETL	Y	G	Java	T/Gr
SQL Power DQguru	JDBC	ELT	Y	G	Java	T/Gr
DQ Analyzer	JDBC	N	Y	G	Java	T/Gr

Gambar I-3 Performance Table Criteria (Sai et al., 2016)

Dalam tabel diatas menjelaskan bahwa (ALL) kemampuan untuk melakukan koneksi ke berbagai macam sumber data. (ELT) *Extract, Load* dan *Transform.* (ETL) *Extract, Transform* dan *Load*. (Y) Yes. (N) No. (G) *Graphical User*

Interface. (T) dapat menghasilkan laporan dalam bentuk tabel. (Gr) Dapat menghasilkan laporan dalam bantuk gambar. (Sai et al., 2016)

Pada penelitian sebelumnya yang telah dilakukan oleh Dwiandriani bahwa integrasi antara Pentaho Data Integration dengan *web site* menggunakan *framework* laravel sebagai penunjang untuk menampilkan hasil dari pengolahan data. Arsitektur aplikasi yang digunakan pada *web* tersebut menggunakan arsitektur *three-tier*, dimana arsitektur *three-tier* terdiri dari 3 bagian yaitu *presentation layer*, *logic layer*, *data layer* (Dwiandriani, 2017)



Gambar I-4 Peracangan arsitektur aplikasi web *Three-tier* (Dwiandriani, 2017)

Tahapan pada 3 layer tersebut memiliki peranan masing – masing dimana pada *presentation layer* merupakan bagian yang berhubungan langsung dengan pengguna dimana di dalamnya terdapat tampilan aplikasi, lalu pada *logic layer* merupakan bagian yang berisi logika – logika dan tempat terjadinya pengolahan data, dan terakhir pada *data layer* bagian ini menampung sumber data yang akan digunakan (Dwiandriani, 2017)

Pada masalah di dunia nyata banyak aplikasi dari organisasi BUMN yang memulai memperhatikan data yang dimiliki oleh organisasi tersebut namun karena banyaknya aplikasi yang memiliki basis data yang bersifat *silo*, hal tersebut

membuat data menjadi redundan dan kotor sehingga organisasi tersebut tidak dapat langsung mengolah data yang sudah dimiliki, salah satu dampaknya adalah dikarenakan setiap aplikasi memiliki basis data sendiri sehingga secara tidak langsung masing – masing dari data memiliki business rule yang berbeda, Untuk mengatasi masalah ini maka dibutuhkan *tool* yang dapat menawarkan solusi. Banyaknya tool data profiling baik secara berbayar hingga *open source*. Penelitian ini menggunakan *open source tool* yang bernama Pentaho Data Integration. Penerapan logika untuk diimplementasikan pada *open source tool* akan menjadi pembanding dengan dilakukannya komparasi guna memberikan keputusan dalam penentuan *open source tool*.

1.2 Rumusan Masalah

Berdasarkan latar belakang diatas, terdapat beberapa rumusah masalah yang dapat diambil diantaranya :

1. Bagaimana implementasi *multi column analysis : value completeness & value similarity* pada *open source platform* untuk penggunaan *data profiling?*
2. Bagaimana melakukan pengujian *profiling* metode *value completeness* dan *value similarity* dengan data text menggunakan dua aplikasi *open source?*
3. Bagaimana hasil komparasi dari implementasi *multi column* dan *value similarity* dengan aplikasi *open source*?

1.3 Tujuan Penelitian

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

1. Implementasi analisis *multi column analysis : value completeness & value similarity* pada *opensource platform.*
2. Melakukan pengujian metode *profiling* dengan aplikasi *open source*

1.4 Batasan Penelitian

Adapula batasan dalam penelitian ini sebagai berikut :

1. Metode analisis multi kolom yang digunakan hanya menggunakan *data completeness & value similarity.*
2. Data yang digunakan dalam penelitian ini diambil dari data Pemerintah Indonesia tahun 2017.

3. Kedua metode hanya menggunakan algoritma Jaro-Winkler dalam pencarian kemiripan atau pengolahan kesamaan data.

1.5 Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini terdapat 2 bagian yaitu manfaat teknikal dimana penelitian ini dapat berkontribusi kepada permasalahan yang dihadapi oleh organisasi saat ini dimana semakin banyak organisasi yang memulai melakukan tata kelola data, dan dengan adanya penelitian ini diharapkan memudahkan banyak organisasi untuk melakukan data *profiling* menggunakan *tools open source* dengan hasil yang dapat dibandingkan pada hasil dengan menggunakan *tools* yang berbayar. Dan manfaat keilmuan dimana peneliti harapkan hasil dari penelitian ini dapat dikembangkan lebih lanjut untuk membantu menyelesaikan permasalahan yang terdapat di masa yang akan dating.

1.6 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, 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.