Penilaian Kualitas Data *Broadband Customer Profiling* (BCP) Pelanggan *Fixed Broadband* PT Telekomunikasi Indonesia Tbk.

Data Quality Assessment of Broadband Customer Profiling (BCP) of Fixed Broadband Customer of PT Telekomunikasi Indonesia Tbk.

Ines Dwi Andini[1], Yova Ruldeviyani[2], Ahmad Hendra Maulana[3], Arif Hidayat[4]
[1,2,3,4]Prodi Magister Teknologi Informasi, Fakultas Ilmu Komputer, Universitas Indonesia
Kampus UI Salemba Jl. Salemba Raya No. 4, Jakarta, Indonesia

[1]ines.dwi@ui.ac.id, [2]yova@ui.ac.id, [3]ahmad.hendra@ui.ac.id, [4]arif.hidayat91@ui.ac.id

Naskah diterima: 22 Februari 2020, direvisi: 15 April 2020, disetujui: 6 Juni 2020

Abstract

A company's success in increasing revenue and managing risk of loss depends on data. High quality data results in good quality decision making at top management level. Telkom is the largest telecommunications company in Indonesia with fixed broadband as its main portfolio. BCP data contains customer access history that can be analyzed to produce user profiling. The results of profiling rely upon the quality of BCP data. The purpose of this study was to assess, identify the main causes of the problem, and provide recommendations to improve the quality of BCP data. The framework used was Total Data Quality Management (TDQM). Results of data quality assessment indicated four main causes of BCP data quality problems and five preventive and corrective recommendations. With these improvements, the quality of BCP data is expected to improve, so that insights generated from processing BCP data are more accurate.

Keywords: *data quality, data quality assessment, total data quality management, fixed broadband.*

Abstrak

Kesuksesan perusahaan dalam meningkatkan pendapatan dan mengelola risiko terjadinya kerugian bergantung pada data. Kualitas data yang bagus menghasilkan pengambilan keputusan yang berkualitas pula di tingkat manajemen. Telkom merupakan perusahaan telekomunikasi terbesar di Indonesia dengan fixed broadband sebagai portofolio utamanya. Data BCP berisi riwayat akses pelanggan yang sangat potensial dianalisis untuk menghasilkan profil pengguna. Bagus tidaknya hasil profil pengguna bergantung pada kualitas data BCP. Tujuan penelitian ini adalah menilai, mengetahui penyebab utama permasalahan, serta memberikan rekomendasi untuk meningkatkan kualitas data BCP. Kerangka kerja yang digunakan adalah Total Data Quality Management (TDQM). Hasil penilaian kualitas data menunjukkan adanya empat penyebab utama permasalahan kualitas data BCP serta lima rekomendasi aksi preventif dan korektif yang perlu dilakukan. Dengan perbaikan pada lima hal tersebut, diharapkan kualitas data BCP dapat meningkat sehingga informasi yang dihasilkan dari pengolahan data BCP semakin akurat dan berkualitas.

Kata kunci: kualitas data, penilaian kualitas data, *total data quality management*, *fixed broadband.*

PENDAHULUAN

Dalam perspektif teknologi informasi, data diartikan sebagai informasi yang tersimpan dalam bentuk digital. Data juga diartikan sebagai fakta yang dikumpulkan, baik oleh individu maupun organisasi dari suatu kejadian di dunia. Data merupakan aset penting bagi suatu perusahaan, karena sangat berperan dalam pengambilan keputusan dan dapat menciptakan keunggulan kompetitif (DAMA International 2017). Organisasi di bidang industri menggunakan data dalam peningkatan kualitas produk, layanan, laba, efisiensi biaya, serta pengelolaan risiko. Sementara itu, organisasi di bidang pemerintahan mengganggap data sebagai sumber daya yang dapat memberikan nilai tambah dalam menjalankan misinya (Fleckenstein dan Fellows 2018).

Istilah kualitas data telah banyak didefinisikan oleh para ahli. Salah satu diantaranya mengidentifikasikan kualitas data sebagai data yang sesuai kebutuhan untuk digunakan oleh konsumen data (Bertoni et al. 2009). Kualitas data juga dapat didefinisikan berdasarkan seberapa baik data mempresentasikan suatu objek, peristiwa dan konsep diciptakannya data tersebut (Wang 1998). Guna memenuhi kesesuaian tersebut dibutuhkan data yang akurat, tepat waktu, relevan, lengkap, dapat dipercaya dan juga dapat dipahami dengan baik (Wang, Ziad and Lee 2002).

Terdapat tiga tantangan pada kualitas data menurut (Strong, Lee, and Wang 1997). Pertama, terdapat banyak tipe sumber data yang membawa tipe data yang berbeda-beda dengan struktur yang kompleks, hal ini menyebabkan sulitnya mengintegrasikan data di antara sistem yang berbeda. Kedua, jumlah data dalam sistem memiliki volume yang besar, sehingga menyebabkan kesulitan dalam menilai kualitas data sesuai dengan waktu yang ditentukan. Ketiga, data berubah dengan cepat, sedangkan ketepatan waktu sangat singkat yang menyebabkan data menjadi usang atau informasi yang dihasilkan menjadi tidak valid.

Dalam meningkatkan kualitas data, strategi yang dapat dilakukan dibedakan menjadi dua, yaitu *data-driven* dan *process-driven* (DAMA International 2017). Strategi *data-driven* meningkatkan kualitas data dengan cara memodifikasi nilai dari data tersebut secara langsung. Adapun strategi *process-driven* meningkatkan kualitas dengan cara melakukan *design* ulang terhadap proses pembuatan atau pengubahan data (Strong, Lee, and Wang 1997).

Dimensi kualitas data merupakan sifat kualitas data yang dapat diukur untuk merepresentasikan beberapa aspek dari data (Laranjeiro, Soydemir, and Bernardino 2015). Untuk mengukur dan menganalisa kualitas data, sebuah perusahaan harus mendefinisikan dimensi kualitas data yang dibutuhkan. Dimensi kualitas data merupakan atribut jika diukur dengan baik, maka dapat menghasilkan level kualitas data secara keseluruhan (Coleman 2013). Kualitas data dapat dianalisis dari beberapa dimensi. (Laranjeiro, Soydemir, and Bernardino 2015). Enam dimensi inti dari kualitas data menurut DAMA International (2017) adalah:

a. Kelengkapan (*Completeness*), yaitu berapa proporsi atau jumlah data yang disimpan bila dibandingkan dengan potensi dari keseluruhan data yang dapat direkam.
b. Keunikan (*Uniqueness*), yaitu tidak ada data yang direkam lebih dari satu kali berdasarkan pada identitas yang membedakan data ini.
c. Ketepatan Waktu (*Timeliness*), yaitu sejauh mana data mewakili keadaaan pada waktu data tersebut diperlukan.
d. Validitas (*Validity*), yaitu data hanya valid jika sesuai dengan sintaks (format, jenis, rentang) dari definisinya.
e. Keakuratan (*Accuracy*), yaitu sejauh mana data dengan benar menggambarkan objek atau peristiwa pada "dunia nyata".

f. Konsistensi (*Consistency*), yaitu tidak adanya perbedaan, ketika membandingkan dua atau lebih representasi dari suatu data yang sama.

Data Quality Management (DQM) menjadi isu yang sedang berkembang di dunia akademis dan profesional. Saat ini, perusahaan mulai menaruh perhatian besar pada kualitas datanya. Data menjadi aset berharga bagi suatu perusahaan yang dapat memberi manfaat tidak hanya di sisi operasional, tetapi juga di sisi strategis. Data dengan kualitas buruk menghasilkan informasi yang tidak akurat, di mana dapat menimbulkan pemborosan sumber daya serta merugikan perusahaan secara eksternal, misalnya hubungan perusahaan dengan pelanggannya (Lucas 2010). Pada dasarnya, sebuah organisasi sudah menyadari bahwa *insight* dari data yang mereka miliki sebagian besar dapat menguntungkan kinerja bisnis mereka melalui teknik, salah satunya adalah *business intelligence* (Jorge et al. 2016). Sebuah penelitian yang dilakukan oleh IBM pada tahun 2016 di Amerika Serikat menyebutkan bahwa kualitas data yang buruk pada suatu perusahaan dapat menyebabkan biaya tahunan mencapai lebih dari tiga triliun Dollar Amerika (Redman 2016). Kesadaran akan kualitas data yang meningkat dalam beberapa tahun terakhir, ternyata tidak diikuti dengan banyaknya jumlah penelitian tentang tingkat kualitas data dalam suatu organisasi (Nagle, Redman, and Sammon 2020). Penelitian (Nagle, Redman, and Sammon 2020) menunjukkan rata-rata 47% data yang ada dalam suatu organisasi memiliki setidaknya satu kesalahan.

PT Telekomunikasi Indonesia Tbk. (Telkom) merupakan perusahaan telekomunikasi terbesar di Indonesia, yang memiliki visi "*the King of Digital in the Region*". Untuk mendukung visi tersebut, maka Telkom merumuskan *Strategic Initiative* 5, yaitu *transform into smart enabler 'hub' platform for the digital ecosystem*, di mana implementasinya memerlukan tata kelola mekanisme pemanfaatan data dan informasi di lingkungan Telkom *Group*. Oleh karena itu, untuk mendukung strategi dan bisnis Telkom tersebut dibutuhkan data dan informasi yang berkualitas tinggi sebagai komitmen Telkom untuk memenangkan kompetisi dalam dinamika ekosistem bisnis yang selalu berubah (Direktur Network IT and Solution 2016). Menurut Menteri Badan Usaha Milik Negara Republik Indonesia (2018), data dan informasi yang berkualitas tinggi harus dapat menjamin kelengkapan (*completeness*), akurasi (*accuracy*), validitas (*validity*), dan autorisasi (*authorization*) data.

Fixed broadband merupakan salah satu portofolio utama Telkom dengan jumlah pelanggan mencapai delapan juta di seluruh Indonesia. Produk *fixed broadband* atau biasa dikenal dengan Indihome merupakan layanan yang menyediakan internet rumah, telepon rumah, dan televisi interaktif (PT Telekomunikasi Indonesia 2019). Data *history* akses pelanggan *fixed broadband* disimpan sebagai data *Broadband Customer Profiling* (BCP). Setiap harinya data *history* akses pelanggan yang masuk mencapai empat hingga delapan milyar baris data dan rata-rata kenaikan data setiap harinya mencapai 33%. Dengan jumlah data yang sangat besar tersebut, investasi *storage* yang dibutuhkan juga sangat besar dan dapat terus bertambah seiring dengan bertambahnya data *history* akses pelanggan setiap harinya. Di sisi lain, data BCP merupakan data yang sangat potensial untuk diolah dan dianalisis menjadi *insight* yang bermanfaat, misalnya untuk segmentasi pelanggan, profil pelanggan, analisa jaringan, *usage*, dan masih banyak lagi. Kualitas data BCP yang bagus menghasilkan *insight* yang akurat yang dapat dimanfaatkan untuk mengembangkan bisnis Telkom. Namun, berdasarkan hasil wawan-cara dengan unit *Big Data Management* Telkom, diperoleh fakta bahwa belum pernah dilakukan pengukuran kualitas data BCP serta untuk menganalisis data BCP diperlukan tahap *pre-processing* yang memakan waktu cukup lama dikarenakan kualitas data yang kurang baik dan kuantitas data yang sangat besar. Selanjutnya, dilakukan observasi data BCP secara langsung dan ditemukan beberapa data dan informasi yang tidak lengkap, tidak valid, dan tidak akurat. Hal ini

yang mendorong dilakukannya penelitian untuk menilai dan mengevaluasi kualitas data BCP Telkom.

Terdapat banyak jenis metode penilaian kualitas data. Perbandingan tiga metode penilaian kualitas data, yaitu *Total Data Quality Management* (TDQM), *Data Quality Assessment* (DQA), dan *Complete Data Quality* (CDQ) (Dama Internasional 2017; Batini et al. 2009; Wang 1998) dapat dilihat pada Tabel 1.

Tabel 1. Perbandingan Metode Penilaian Kualitas Data

Kriteria	TDQM	DQA	CDQ
Akronim	*Total Data Quality Management*	*Data Quality Assessment*	*Complete Data Quality*
Komponen Utama	Tahapan: • Pendefinisian • Pengukuran • Analisis • Peningkatan *tahapan proses yang kontinu (berkelanjutan) Fokus: Informasi produk	• Penilaian secara subjektif dan objektif • Analisis perbandingan • Analisis sebab akibat • Peningkatan	Tahapan: • *State reconstruction* • Pengukuran • Peningkatan *tahapan proses yang tidak kontinyu (tidak berlanjut)
Tipe Data	• Data terstruktur • Data semi terstruktur	Data terstruktur	• Data terstruktur • Data semi terstruktur
Assessment & Improvement	Komprehensif, dilihat juga dari perspektif implementasi	Membuat perbedaan antara metrik kualitas data subyektif dan obyektif	Pemilihan proses peningkatan kualitas data dilakukan melalui *cost-benefit analysis*

Sumber: Batini et al. (2009); Wang (1998)

Dalam melakukan penilaian kualitas data, kerangka kerja yang digunakan adalah *Total Data Quality Management* (TDQM). TDQM menyediakan kerangka kerja umum yang melakukan peningkatan kualitas data dengan menggunakan pendekatan DQM. TDQM mendefinisikan empat area yang menjadi tantangan dalam kualitas data, yaitu sistem, proses, prosedur, dan kebijakan (DAMA International 2017). TDQM merupakan pendekatan yang terstruktur dan komprehensif untuk manajemen suatu perusahaan dalam meningkatkan kualitas data (Batini et al. 2009; Wijayanti et al. 2018).

Terdapat dua pendekatan untuk menghasilkan data dengan kualitas yang baik, yaitu aksi preventif dan aksi korektif (DAMA International 2017). Aksi preventif dilakukan dengan cara mencegah data yang berkualitas buruk masuk. Tindakan pencegahan ini dapat menghentikan terjadinya kesalahan. Aksi preventif terdiri dari enam jenis, yaitu:

1. *Establish data entry controls*, membuat aturan untuk memasukkan data ke dalam sistem yang dapat mencegah data tidak valid dan tidak akurat.
2. *Train data producers*, memastikan semua karyawan memahami dampak dengan adanya data yang kualitasnya buruk dan memberikan insentif kepada karyawan berdasarkan keakuratan dan kelengkapan data yang dimasukkan ke dalam sistem.
3. *Define and enforce rules*, membuat aturan mengenai kualitas data terhadap bisnis serta menginformasikan kepada tim *analyst* jika data memiliki kualitas yang di bawah standar.
4. *Demand high quality data from data suppliers*, memeriksa proses yang terjadi pada penyedia data eksternal terkait struktur, definisi, data sumber datanya, kemudian melakukan penilaian mengenai seberapa baik data tersebut dapat diintegrasikan, sehingga dapat mencegah penggunaan data yang tidak sesuai dengan peruntukkannya.
5. *Implement data governance and stewardship*, memastikan peran dan tanggung jawab

telah mendeskripsikan dan melaksanakan aturan untuk pengelolaan aset data dan informasi secara efektif.

6. *Institute formal change control*, memastikan semua perubahan pada data yang tersimpan telah didefinisikan dan diuji sebelum diimplementasikan.

Aksi korektif merupakan tindakan yang diimplementasikan setelah permasalahan kualitas data terjadi dan dapat dideteksi. Menyelesaikan permasalahan ketika permasalahan tersebut terjadi merupakan *best practice* dalam DQM. Dalam aksi korektif secara tidak langsung juga dilakukan aksi preventif untuk mencegah terjadinya permasalahan yang sama pada kualitas data. Aksi korektif terdiri dari tiga jenis, yaitu:

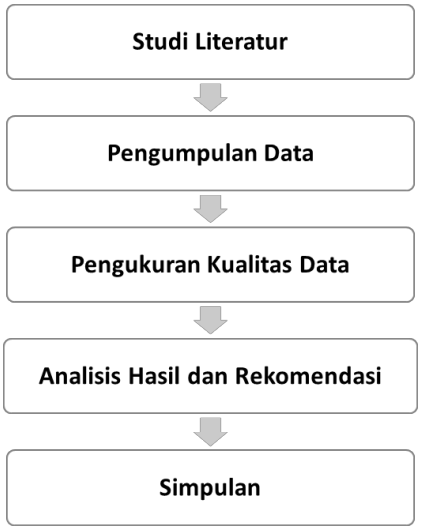
1. *Automated correction*, meliputi standardisasi, normalisasi, dan koreksi berbasis aturan. Nilai yang dimodifikasi dihasilkan tanpa adanya intervensi secara manual.
2. *Manually-directed correction,* menggunakan *tools* yang secara otomatis dapat memulihkan dan memperbaiki data, namun tetap membutuhkan tinjauan manual sebelum melakukan koreksi ke dalam penyimpanan tetap.
3. *Manual correction*, merupakan metode yang tidak dianjurkan. Namun, ketika tidak ada pilihan lain yang bisa diambil atau perubahan hanya bisa dilakukan secara manual oleh manusia, maka metode ini bisa dilakukan.

Pertanyaan penelitian yang dijawab dalam penelitian ini adalah: (1) Seberapa baik kualitas data BCP Telkom? (2) Apa saja penyebab utama permasalahan kualitas data BCP Telkom? dan (3) Apa rekomendasi yang dapat diberikan untuk meningkatkan kualitas data BCP Telkom?

Penelitian ini disusun sebagai berikut: pada bagian pendahuluan dijabarkan latar belakang penelitian, tujuan, review dari studi literatur yang berkaitan dengan penelitian ini, serta sistematika penulisan hasil penelitian. Pada bagian metode dijabarkan instrumen penelitian yang digunakan, prosedur pengumpulan data, serta metode untuk menilai kualitas data. Pada bagian hasil dan pembahasan dijabarkan hasil yang diperoleh setelah penilaian kualitas data dilakukan, diskusi dan implikasi hasil penilaian kualitas data, serta rekomendasi yang bisa diberikan untuk meningkatkan kualitas data. Pada bagian akhir, diberikan simpulan dari hasil penelitian yang dilakukan.

METODE

Penelitian ini menggunakan pendekatan kualitatif, yakni proses penelitian yang menghasilkan data deskriptif berupa kata-kata tertulis atau lisan dari orang dan perilaku yang diamati. Alur metode penelitian yang digunakan adalah sebagai berikut: studi literatur dilakukan dari berbagai sumber akademik dan dokumen-dokumen perusahaan. Pengumpulan data yang sifatnya wawancara dilakukan dengan menggunakan instrumen wawancara. Wawancara dilakukan dengan *Manager Big Data & Analytics Development* dan stafnya, yang berada di bawah Divisi *Information Technology* (DIT), serta dengan *Head of Big Data Management, Manager Data Quality Management, Manager Datamart dan Data Mining* beserta staf, yang berada di bawah Divisi *Digital Service* (DDS). Adapun observasi dilakukan dengan *query* secara langsung ke dalam Hive-Hadoop, tempat di mana data BCP disimpan. Dengan demikian, proses pengumpulan data untuk penelitian ini menggunakan metode triangulasi, di mana dengan metode tersebut dapat mengumpulkan data sekaligus menguji dan mengecek kredibilitas data dengan teknik pengumpulan yang berbeda-beda untuk mendapatkan data dari sumber yang sama (Sugiyono 2013).



Gambar 1. Bagan Metode Penelitian

Penilaian kualitas data dilakukan dengan menggunakan kerangka kerja TDQM yang diadopsi dari penelitian sebelumnya (Jiang and Zhao 2012). Menurut Strong, Lee dan Wang (1997) metode TDQM adalah metode umum pertama yang diterbitkan dalam literatur kualitas data (Batini et al. 2009). TDQM adalah hasil penelitian akademik yang banyak digunakan sebagai panduan untuk *reenginering* data organisasi. Tujuan mendasar dari TDQM adalah untuk memperluas kualitas data, yang merupakan prinsip dari *Total Quality Management* (TQM) (Wijayanti, et al. 2018).

Dikutip dari Cichy and Rass (2019) TDQM dapat diringkas menjadi empat proses dasar, sebagai berikut:

a. *Define*: di mana pada proses ini menyelaraskan tujuan kualitas data dengan tujuan strategis perusahaan, mengidentifikasi dan menganalisis siklus informasi yang mencirikan proses produksinya, mengidentifikasi kualitas yang dirasakan dan diinginkan oleh informasi konsumen, serta mengevaluasi kepuasan pengguna terkait dengan kualitas data.
b. *Measurer*: Mendefinisikan metrik objektif untuk kualitas data, formula, dan menerapkan metrik tersebut ke berbagai sumber data dan titik siklus informasi melibatkan pembersihan data dan proses desain ulang.
c. *Analyze*: mengidentifikasi penyebab masalah kualitas untuk mendukung perencanaan peningkatan kualitas, yang mungkin melibatkan pembersihan data dan proses desain ulang.
d. *Improve*: memprioritaskan bidang-bidang utama dan menguraikan perencanaan proses perbaikan. Dengan demikian, TDQM adalah metode yang berfokus baik pada proses produksi informasi, dengan mengidentifikasi siklus hidupnya, dan pada penilaian konten data. Namun, hal ini tidak sepenuhnya merinci masalah organisasi yang secara langsung memengaruhi kualitas data.

Dalam menilai kualitas data, TDQM telah mempertimbangkan empat tahapan yang ada di DQM, yaitu mendefinisikan dimensi dari kualitas data, mengukur metrik kualitas data, menganalisa hasil, dan melakukan perbaikan. Adapun tahapan utama yang perlu dilakukan adalah:

1. Mengidentifikasi data yang dinilai.
2. Menentukan dimensi kualitas data yang digunakan.
3. Mengukur kualitas data dengan melakukan *query* secara langsung berdasarkan *entity*

yang telah ditentukan sebelumnya.

4. Menganalisa hasil penilaian kualitas data dan penyebab terjadinya anomali data dalam sistem basis data.

Analisis hasil dan rekomendasi dilakukan berdasarkan hasil yang diperoleh pada tahap penilaian kualitas data. Pada bagian akhir, membuat simpulan dari penelitian yang telah dilakukan. Alur metode penelitian yang digunakan dapat dilihat pada Gambar 1.

HASIL DAN PEMBAHASAN

Penelitian ini menggunakan kerangka kerja TDQM yang merupakan kerangka kerja umum yang terstruktur dan komprehensif untuk melakukan penilaian kualitas data dengan pendekatan DQM. Tahap pertama yang dilakukan adalah *define*, yaitu dengan mengidentifikasi data yang dinilai. Data yang dinilai adalah data BCP, yaitu data *history* akses pengguna *fixed broadband* Telkom ke internet. Data tersebut di-*capture* dari *network* Telkom yang tersebar di wilayah *network* Telkom, terpusat di Jatinegara. Kemudian, setiap 30 menit, Hadoop menarik data BCP dari Jatinegara tersebut. Setiap baris data BCP, mengandung informasi waktu akses, pelanggan (yang telah di-*masking*), *device* yang digunakan untuk akses, apa yang diakses, *usage*, *delay*, *failure*, dan *packet*. Data BCP yang dijadikan *sample* adalah data bulan Oktober 2019, dengan 42 kolom dan 2 milyar baris data. Penilaian kualitas data BCP dilakukan dengan cara *query* secara langsung ke dalam Hive-Hadoop Telkom melalui Ambari.

Selanjutnya, tahap kedua yang dilakukan adalah *measurer*, yaitu dengan menentukan dimensi kualitas data yang akan dinilai. Dalam penelitian ini, penilaian kualitas data BCP menggunakan tiga dimensi, yaitu kelengkapan (*completeness*), akurasi (*accuracy*), dan validitas (*validity*). Pemilihan tiga dimensi tersebut didasarkan pada kebijakan pengelolaan sumber daya data dan informasi dalam Badan Usaha Milik Negara (BUMN) (Menteri Badan Usaha Milik Negara Republik Indonesia 2018). Selain itu, pada tahap ini juga dilakukan penilaian kualitas data dengan melakukan *query* secara langsung berdasarkan *entity* yang telah ditentukan sebelumnya. Tahap ketiga yang dilakukan adalah *analyze*, yaitu dengan menganalisa hasil penilaian kualitas data dan penyebab terjadinya anomali data dalam sistem basis data. Hasil penilaian kualitas data pada masing-masing dimensi serta analisa terhadap hasil penilaian kualitas data akan dijelaskan pada sub bab bagian ini.

Dimensi *Completeness*

Penilaian dimensi *completeness* berhubungan dengan representasi statistik dari suatu data dan keterwakilan periode pengumpulan data (Bicalho et al. 2017). Penilaian dimensi *completeness* dilakukan pada 42 kolom di tabel BCP. Kesemua kolom tersebut memiliki peran dalam analisis data untuk menghasilkan *insight* yang bermanfaat. Misalnya, kolom *device_name* dan *vendor_name* yang berisi merk dan tipe *device* pengguna yang digunakan untuk mengakses internet, secara tidak langsung dapat dimanfaatkan untuk melihat tingkat ekonomi pengguna dari harga *device* yang dimilikinya.

Tabel 2. Data Persentase *Completeness*

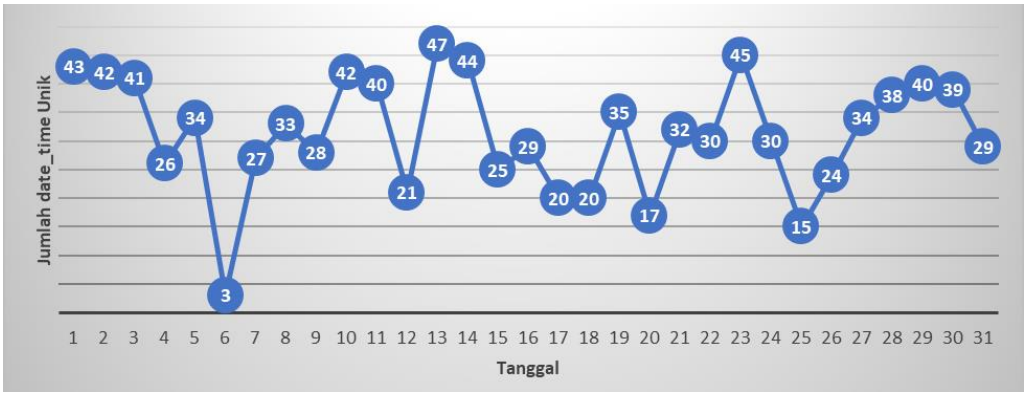
Persentase	Total Kolom	Nama Kolom
100%	34	• user_id • activity_sec • volume_in • volume_out

Persentase	Total Kolom	Nama Kolom
		• rxmit_volume_in • rxmit_volume_out • client_delay • server_delay • client_delay_sum • server_delay_sum • client_delay_samples • server_delay_samples • first_data_delay_sum • first_data_delay_samples • connection_failures • session_failures • internal_ploss • external_ploss • packets_in • packets_out • rxmit_packets_in • rxmit_packets_out • total_flow • ds • periode • dll
98%	1	application_name
65%	1	date_time
62%	1	component_name
60%	1	category_name
12%	1	device_name
5%	1	os_name
0%	2	• browser_name • vendor_name

Sumber: Data BCP Bulan Oktober 2019, telah diolah kembali

Penilaian dimensi *completeness* melihat ada tidaknya *null value* (“NULL”), *blank value* (“”), serta data “*unclassified*” dalam suatu kolom. Hasil penilaian dimensi *completeness* menunjukkan 34 kolom memiliki persentase *completeness* 100% dan delapan kolom sisanya memiliki persentase antara 0% - 98%. Delapan kolom tersebut mayoritas diisi dengan data “*unclassified*”. Data “*unclassified*” merupakan data yang belum dikategorisasikan atau belum di-*index* dalam *library*, sehingga ketika data dari *network* masuk, maka disimpan sebagai data “*unclassified*”. Banyaknya jumlah baris data yang “*unclassified*” dapat menghambat dan membuat hasil analisis data menjadi samar. Hasil wawancara dengan unit Datamart dan Data Mining, data “*unclassified*” ini tidak diikutkan ketika menyusun analisis, karena dianggap sama dengan data *null* atau *blank value*. Detail hasil penilaian dimensi *completeness* dapat dilihat pada Tabel 2.

Selanjutnya, dilakukan observasi lebih lanjut pada kolom *date_time*. Oleh karena, Hadoop menarik data BCP setiap 30 menit, maka setiap hari seharusnya terdapat 48 baris *date_time* yang unik dan selama bulan Oktober 2019 ada 1.488 baris *date_time* yang unik. Setelah dilakukan observasi, diperoleh 973 *date_time* yang unik dalam tabel BCP bulan Oktober 2019 atau sekitar 65%. Total data *date_time* yang unik setiap harinya dapat dilihat pada Gambar 2.



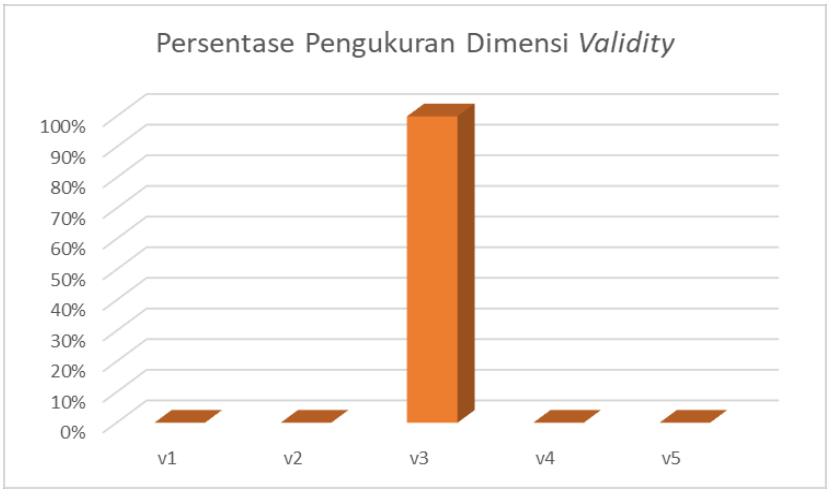
Gambar 2. Persebaran Jumlah *date_time* Unik Bulan Oktober 2019
Sumber: Data BCP Bulan Oktober 2019, telah diolah kembali

Dimensi *Validity*

Penilaian dimensi *validity* dilakukan berdasarkan lima kriteria yang merupakan hasil observasi dan wawancara dengan DIT dan DDS. Kriteria dimensi *validity* tersebut adalah:

a) Format kolom date_time adalah YYYY-MM-DD HH:MM (v1)
b) Format kolom ds adalah YYYYMMDD. (v2)
c) Format kolom periode adalah YYYYMM. (v3)
d) Kolom client_ip berada di rentang 0.0.0.0 dan 255.255.255.255. (v4)
e) Kolom user_id terdiri dari 32 digit. (v5)

Penilaian kualitas data pada dimensi *validity* dengan menggunakan lima kriteria di atas menghasilkan data yang ditunjukkan pada Gambar 3. Dari Gambar 3 terlihat hasil penilaian kualitas data pada dimensi *validity* kriteria v3 (format kolom periode adalah YYYYMM) sebesar 100%. Ini berarti semua data dalam kolom periode tidak menggunakan format YYYYMM. Seluruh baris dalam kolom periode berisi data "0", di mana data ini tidak mewakili makna kolom periode. Hal ini disebabkan oleh adanya *query* yang tidak jalan dengan baik ketika menarik data BCP ke dalam Hive-Hadoop



Gambar 3. Persentase Hasil Penilaian Dimensi *Validity*
Sumber: Data BCP Bulan Oktober 2019, telah diolah kembali

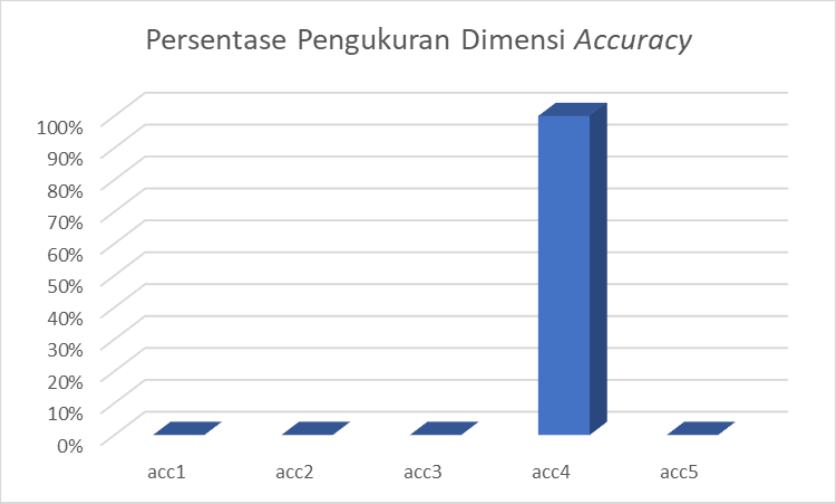
Dimensi *Accuracy*

Penilaian dimensi *accuracy* dilakuken dengan membandingkan data yang ada dengan *rule*

yang ditentukan. Adapun *rule* yang digunakan dalam penilaian ini adalah:

a) Kolom user_id hanya terdiri dari huruf kecil dan angka. (acc1)
b) Dalam satu baris data salah satu atau keduanya, kolom *volume_in* dan *volume_out* bernilai > 0. (acc2)
c) Kolom ds menunjukkan tanggal, bulan, dan tahun yang sama dengan kolom *date_time*. (acc3)
d) Kolom periode menunjukkan bulan dan tahun yang sama dengan kolom *date_time*. (acc4)
e) *Time* pada kolom *date_time* kelipatan 30 menit. (acc5)

Hasil penilaian kualitas data dimensi *accuracy* dapat dilihat pada Gambar 4. Dari Gambar 4 terlihat hasil penilaian kualitas data pada dimensi *accuracy rule* acc4 (kolom periode menunjukkan bulan dan tahun yang sama dengan kolom *date_time*) sebesar 100%. Ini berarti semua data dalam kolom periode tidak menunjukkan bulan dan tahun yang sama dengan kolom *date_time*. Sebagaimana telah dijelaskan pada bagian dimensi *validity*, bahwa seluruh baris dalam kolom periode berisi data "0", sehingga tidak menunjukkan informasi yang sama dengan kolom *date_time*. Kolom periode merupakan salah satu kolom partisi dalam data BCP. Jika persentase *accuracy* dari kolom partisi buruk, memengaruhi proses *query* ke data BCP menjadi tidak efektif dan efisien.



Gambar 4. Persentase Hasil Penilaian Dimensi *Accuracy*
Sumber: Data BCP Bulan Oktober 2019, telah diolah kembali

Analisis Penyebab Utama Permasalahan Kualitas Data

Berdasarkan penilaian yang dilakukan sesuai dengan kriteria yang telah ditetapkan pada masing-masing dimensi, terdapat beberapa hal yang menjadi penyebab utama permasalahan kualitas data. Hal tersebut dapat diidentifikasi dari alur data masuk ke sistem hingga dapat di-*query* melalui Hive-Hadoop.

Untuk mengidentifikasi penyebab utama permasalahan tersebut dilakukan wawancara dan observasi dengan DIT dan DDS yang merupakan divisi yang bertanggung jawab dalam *develop*, *manage*, dan memanfaatkan data BCP. Hal-hal yang mempengaruhi kualitas data BCP adalah sebagai berikut:

a. Dimensi *Completeness: Library* kamus data yang digunakan untuk mengategorisasikan kolom-kolom dalam data BCP belum *up to date*. Hal ini menyebabkan banyaknya data yang "*unclassified*".
b. Dimensi *Completeness:* Belum ada *recovery system* yang otomatis agar ketika terjadi

error, data otomatis ditarik ketika sistem sudah normal. Saat ini, jika terjadi *error*, maka data akan ditarik ketika terdapat *complain* terjadinya data tidak lengkap saja. Itupun jika data masih bisa ditarik.

c. *Data Governance*: Belum ada kebijakan mengenai DQM sehingga belum ada prosedur yang mengatur siapa yang bertanggung jawab untuk *monitoring* dan meningkatkan kualitas data BCP.
d. Dimensi *Validity* dan *Accuracy*: Ada *query* yang tidak jalan dengan baik ketika menarik data BCP ke dalam Hive-Hadoop. Hal ini menyebabkan kolom periode berisi data yang tidak sesuai.

KESIMPULAN

Hasil penilaian kualitas data BCP di Telkom menggunakan tahapan kerangka kerja TDQM dengan tiga dimensi yang diukur, yaitu *completeness*, *validity*, dan *accuracy* menunjukkan hasil: rata-rata persentase *completeness* (*null value*, *blank value,* dan *unclassified*) dari 42 kolom yang diukur adalah sebesar 88%. Selanjutnya, pada dimensi *validity* menunjukkan rata-rata 80% data *valid* memenuhi lima kriteria *validity* yang ditentukan. Pada dimensi *accuracy* menunjukkan rata-rata 80% data akurat memenuhi *rule accuracy* yang ditetapkan. Ketiga nilai rerata tersebut diperoleh dengan melakukan rata-rata terhadap persentase hasil penilaian dengan jumlah kolom atau kriteria yang dinilai pada masing-masing dimensi.

Ada empat penyebab utama terjadinya permasalahan pada kualitas data BCP, yaitu kamus data yang tidak *up to date*, belum adanya *recovery system* yang otomatis, belum adanya kebijakan *data quality management*, serta adanya *error* pada penarikan data. Untuk meningkatkan kualitas data BCP Telkom, maka lima aksi *preventive* dan *corrective* yang perlu dilakukan adalah membuat sistem pengecekan dan pemberitahuan adanya data baru yang "*unclassified*", melakukan *update library* secara berkala, merancang dan membangun *recovery system* yang otomatis, membuat kebijakan DQM untuk data BCP, serta melakukan pengecekan terhadap *query* penarikan data BCP ke dalam Hive-Hadoop. Dengan perbaikan pada lima hal tersebut diharapkan kualitas data BCP meningkat, sehingga *insight* yang dihasilkan dari pengolahan data BCP semakin akurat dan berkualitas.

Berdasarkan hasil penilaian kualitas data BCP yang dilakukan di Hive-Hadoop sebagai studi kasus dari Telkom, maka untuk meningkatkan kualitas data BCP dapat dilakukan dengan melakukan aksi yang mengacu pada teori data *Data Management Body of Knowledge* (DMBOK) sebagai berikut:

a. *Preventive Actions – Define and Enforce Rules*: Membuat sistem pengecekan dan pemberitahuan adanya data baru yang "*unclassified*".
b. *Corrective Actions – Manually Directed Correction*: Melakukan *update library* data yang digunakan untuk kategorisasi data BCP, jika terdapat data yang "*unclassified*".
c. *Preventive Actions – Define and Enforce Rules:* Merancang dan membangun *recovery system* yang dapat mengatasi penarikan data ketika terjadi *error*.
d. Preventive Actions – Implement Data Governance and Stewardship: Membuat kebijakan yang mengatur DQM BCP.
e. *Corrective Actions – Manual Correction:* Melakukan pengecekan ulang pada *query* penarikan data BCP ke dalam Hive-Hadoop.

Lima rekomendasi yang diberikan di atas masuk dalam strategi peningkatan kualitas data secara *process-driven*, di mana melakukan peningkatan kualitas data dengan cara melalui

Lima rekomendasi yang diberikan di atas masuk dalam strategi peningkatan kualitas data secara *process-driven*, di mana melakukan peningkatan kualitas data dengan cara melalui perancangan ulang terhadap proses pembuatan data, misalnya dengan *update library* kategorisasi data BCP, membangun *recovery system*, serta pengecekan *query* penarikan data BCP. Dengan implementasi kelima rekomendasi tersebut, maka terjadinya data tidak lengkap, tidak valid, dan tidak akurat dapat dihindari, sehingga kualitas data BCP akan semakin baik, dapat menghemat waktu dalam *pre-processing* data, serta meningkatkan keakuratan hasil *profiling* pelanggan Indihome.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Kementerian Komunikasi dan Informatika (Kominfo) sebagai pemberi beasiswa, Ibu Yova Ruldeviyani sebagai pembimbing, serta Bapak Ida Bagus Mahaputra dan Ibu Tutut Vaty Husnawati dari Unit Big Data Management (BDM), Divisi Digital Service (DDS) sebagai narasumber untuk penelitian ini.

DAFTAR PUSTAKA

Batini, Carlo, Cinzia Cappiello, Chiara Francalanci, och Andrea Maurino. 2009. ”Methodologies for Data Quality Assessment and Improvement.” *ACM Computing Surveys* 41 (3). doi:10.1145/1541880.1541883.

Bertoni, Matteo, Giuliano Furlini, Gianluca Gozzoli, Mariapaola Landini, Matteo Magnani, Antonio Messina, och Danilo Montesi. 2009. ”A Case Study on the Analysis of the Data Quality of a Large Medical Database.” *2009 20th International Workshop on Database and Expert Systems Application* (IEEE): 308-312. doi:10.1109/DEXA.2009.82.

Bicalho, Tereza, Ildo Sauer, Alexandre Rambaud, och Yulia Altukhova. 2017. ”LCA Data Quality: A Management Science Perspective.” *Journal of Cleaner Production:* 888-898. doi:10.1016/j.jclepro.2017.03.229.

Cichy, Corinna, och Stefan Rass. 2019. ”An Overview Data Quality Framework.” *IEEE Access* (IEEE) 7: 24634-24648. doi:10.1109/ACCESS.2019.2899751.

Coleman, Laura Sebastian. 2013. *Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework.* Morgan Kaufmann. doi:10.1016/C2011-0-07321-0.

DAMA International. 2017. *DAMA-DMBOK Data Management Body Of Knowledge.* 2nd. New Jersey: Technics Publications.

Direktur Network IT and Solution. 2016. *Peraturan Direktur Network IT and Solution Perusahaan Perseroan (Persero) PT Telekomunikasi Indonesia Tbk Nomor PR.404.03/r.00/HK.270/COP-D0031000/2016 Tentang Tata Kelola Data (Data Governance) Telkom Group.* Jakarta: PT Telekomunikasi Indonesia.

Fleckenstein, Mike, och Lorraine Fellows. 2018. *Modern Data Strategy.* Springer.

Jiang, Lizheng, och Jiantao Zhao. 2012. ”An Empirical Study on Risk Data Quality Management.” *International Conference on Information Management, Innovation Management, and Industrial Engineering* (IEEE): 511-514. doi:10.1109/ICIII.2012.6339714.

Jorge, Merino, Caballero Ismael, Rivas Bibiano, Serrano Manuel, och Piattini Mario. 2016. ”A Data Quality in Use model for Big Data.” *Future Generation Computer Systems* (Elsevier) 63: 123-130. doi:10.1016/j.future.2015.11.024.

Laranjeiro, Nuno, Seyma Nur Soydemir, och Jorge Bernardino. 2015. ”A Survey on Data Quality: Classifying Poor Data.” *2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing:* 179-188.

Menteri Badan Usaha Milik Negara Republik Indonesia. 2018. *Peraturan Menteri BUMN RI Nomor PER-03/MBU/02/2018 Tentang Perubahan Atas Peraturan Menteri BUMN Nomor PER-02/MBU/2013 Tentang Panduan Penyusunan Pengelolaan Teknologi Informasi BUMN.* Jakarta: Kementerian Badan Usaha Milik Negara.

Nagle, Tadhg, Tom Redman, och David Sammon. 2020. ”Assessing Data Quality: A Managerial Call to Action.” *Business Horizons.* doi:10.1016/j.bushor.2020.01.006.

PT Telekomunikasi Indonesia. 2019. *Indihome.* https://www.indihome.co.id/.

Redman, Thomas C. 2016. *Getting In Front On Data: Who Does What.* USA: Technics Publication.

Strong, Diane M., Yang W. Lee, och Richard Y Wang. 1997. ”Data Quality in Context.” May: 103-110. doi:10.1145/253769.253804.

Sugiyono. 2013. *Metode Penelitian Kuantitatif, Kualitatif, dan R&D.* Bandung: Alfabeta.

Wang, Richard Y. 1998. ”A Product Perspective On Total Data Quality Management.” doi:10.1145/269012.269022.

Wang, Richard Y., Mostapha Ziad, och Yang W. Lee. 2002. *Data Quality.* New York: Kluwer Academic Publisher.

Wijayanti, Wiluyaningtyas, Achmad Nizar Hidayanto, Nori Wilantika, Infaz Rizki Adawati, och Satrio B Yudhoatmojo. 2018. ”Data Quality Assessment on Higher Education: A Case Study of Institute of Statistics.” *2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI):* 231-236.