BIG DATA: INKONSISTENSI DATA DAN SOLUSINYA

Bawono Adi Sanjaya[1)], Selo Sulistyo[2]

[1), 2)] Magister Teknik Informatika UGM Yogyakarta
Jl Ring road Utara, Condongcatur, Sleman, Yogyakarta 55281
Email : yayatsanjaya@gmail.com[1)], selo@ugm.ac.id[2)]

Abstrak

Produksi data dunia saat ini mengalami peningkatan eksponensial dari tahun ke tahun. Pertumbuhan volume data yang pesat seiring dengan transaksi data di internet yang terjadi secara masif, bervariasi, serta memiliki struktur yang kompleks ketika data masuk ke media penyimpanan dikenal dengan istilah big data. Big data datang dari berbagai sumber dan memiliki sifat yang heterogen. Karena sifatnya yang heterogen, data bisa saja mengalami ketidakkonsistenan. Data yang mengalami konflik sangat menyulitkan dalam analisis big data untuk pengambilan keputusan. Tulisan ini akan membahas tentang konflik atau inkonsistensi data dalam big data. Di bagian akhir akan dibahas tentang beberapa metode dan algoritma yang digunakan untuk menyelesaikan problem ini.

Kata kunci:*big data, inkonsistensi, heterogen, volume..*

1. Pendahuluan

Big data saat ini bukan lagi sekedar istilah populer yang berkembang dalam dunia teknologi dan bisnis. Big data telah berkembang dengan pesat dan menyentuh hampir seluruh bidang yang berkaitan dengan aktivitas manusia. Sumber data dalam big data berasal dari domain kedokteran, pendidikan, komunikasi dan media massa, pemerintahan, asuransi, jasa keuangan, energi dan industri, lingkungan, transportasi, layanan berbasis lokasi (*location-based services*), manufaktur dan bisnis ritel.

Data-data tersebut menyebar di internet dengan berbagai macam format diantaranya transaksi online, surat elektronik (email), video, audio, gambar, kata kunci untuk mesin pencari, rekaman aktivitas kejadian (logs), interaksi di media sosial, data *sciences* , sensor serta data-data yang dikirimkan dari perangkat bergerak (mobile devices) dan aplikasinya [1].

Di tahun 2012, jumlah data digital di dunia telah mencapai 2.72 *zettabytes* (10^{21}) [1]. Data diprediksikan akan mencapai kenaikan hingga 8 *zettabytes* di tahun 2015 [2]. Survey IBM mengemukakan bahwa setiap hari sebanyak 2.5 *exabytes* data telah diciptakan oleh seluruh manusia di dunia, dan 90% dari data tersebut diproduksi di tahun 2011 [3]. Jenis data yang paling banyak beredar di internet adalah multimedia dan mengalami kenaikan sebesar 70% ditahun 2013 [4]. Google adalah satu-satunya perusahaan yang bergerak di bidang teknologi yang memiliki satu juta server menyebar diseluruh dunia. Mereka memiliki langganan mobile sebanyak 6 miliyar di tahun 2013 dan setiap harinya 10 miliyar pesan teks dikirimkan lewat server yang dimilki oleh google [1].

Data yang berasal dari berbagai macam sumber dan dengan format yang berbeda-beda disebut data heterogen. Data sumber yang bersifat heterogen dapat digunakan untuk keperluan analisis big data sehingga pengguna dapat mengekstrak informasi yang dibutuhkan saja. Sayangnya, sebagian besar dari informasi yang didapatkan oleh pengguna, organisasi, ataupun perusahaan enterprise tidak sesuai dengan apa yang mereka inginkan. Di Amerika Serikat, tingkat kebocoran dana untuk mendapatkan data yang valid mencapai $600 miliyar setiap tahunnya [5]. Perusahaan *enterprise* sering menemukan data yang tidak relevan dengan nilai 1-5%, dan untuk beberapa perusahaan nilai tersebut mencapai 30% [6, 7]. Pada kebanyakan perusahaan berbasis *data warehouse*, proses pembersihan dan pemilihan data (data cleaning) mencapai statistik 30-80% dari waktu pengerjaan proyek, sehingga mereka sering menaikkan anggaran untuk mendapatkan data yang berkualitas daripada membangun sistem yang baik dan tertata [8].
Untuk menangani masalah tersebut diperlukan manajemen data yang baik untuk mendapatkan data yang berkualitas. Jika manajemen data tidak teroganisir dengan rapi, setiap kesalahan kecil yang muncul seperti aktivitas input, proses dan output bisa terakumulasi menjadi besar dan mengakibatkan hilangnya pendapatan. Dampaknya adalah dunia industri mengalami proses inefisiensi dan kegagalan di bidang industri dan regulasi pemerintah tidak berjalan dengan sempurna seperti yang dipaparkan dalam teori *butterfly effect* [9].

Beberapa tahapan untuk mendapatkan data yang berkualitas adalah: membuat aturan atau basis pengetahuan, pengecekan inkonsistensi dan perbaikan data. Dampak dari proses tersebut akan bermuara pada permasalahan efisiensi dan akurasi (*efficiency vs accuracy trade-off*) [8]. Efisiensi penting ketika proses komputasi dilakukan, tetapi tidak mengesampingkan akurasi untuk mendapatkan data yang valid sesuai dengan keinginan dari pengguna.

Penyaringan data terhadap data sumber (data sources) sebelum diolah sangat penting untuk mendapatkan data

yang valid untuk keperluan analisis big data. Data yang menyebar di internet saat ini sangat beragam, terdiri dari banyak sumber dan seringkali memiliki nilai yang spesifik walapun berada pada domain yang sama. Seringkali juga terjadi pemahaman yang kontradiktif terhadap suatu data, sehingga data harus dibuat konsisten untuk mendapatkan makna yang seragam. Kondisi ini akan memudahkan pengguna ketika mereka mencoba mengekstrak informasi sesuai dengan yang mereka inginkan. Oleh karena itu, pembahasan dari paper ini akan berfokus pada inkonsistensi data dan solusi penanganannya.

Pembahasan dalam paper ini diorganisasikan dalam beberapa bagian. Bagian pertama membahas tentang data dan cakupan *big data* secara umum. Bagian kedua membahas tentang istilah-istilah khusus yang berkaitan dengan *big data*. Bagian ketiga tentang inkonsistensi big data. Bagian keempat menguraikan tentang beberapa metode dan algoritma untuk menyelesaikan masalah inkonsistensi dan bagian kelima adalah kesimpulan.

2. Pembahasan

A. Isu-Isu Penting dalam *Big Data*

Istilah big data selalu berkaitan erat dengan jumlah data yang sangat besar dan melewati tahapan yang rumit ketika data tersebut disimpan, diproses dan dianalisa pada teknologi database tradisional [10]. Dalam pembahasan yang lebih kompleks [11, 12], big data dipengaruhi oleh ukuran/jumlah (*volume*), keberagaman (*variety*), dan kecepatan (*velocity*). Belakangan ini muncul masalah bagaimana mendapatkan data yang valid. Dari sekian banyak data yang diperolah dari berbagai sumber, sangat mungkin jika data tersebut memiliki makna yang beragam (ambigu), bahkan ada indikasi juga data memiliki kerancuan. Sebagai contoh sebuah kalimat yang sama tetapi maknanya berbeda. Solusi dari permasalahan ini adalah dengan menguji kualitas dari data tersebut. Sehingga muncullah istilah *veracity* karena terjadi ketidakpastian terhadap data [13]. Dan yang terakhir adalah *value* yaitu aspek yg berkaitan dengan ekstraksi informasi. Kelimanya dikenal dengan istilah "*the 5V's of big data analytics*" [14].

Volume

Sekumpulan tipe data yang di-*generate* dari banyak sumber data dan akan mengalami perkembangan dalam kuantitas secara terus menerus. Data di-generate oleh mesin (*machines*), jaringan (*networks*) dan manusia (*human*) yang saling berinteraksi satu sama lain yang jumlahnya terus bertambah. Total data yang ada saat ini telah mencapai ukuran *zettabytes* dan media jejaring sosial adalah penyumbang terbesar dengan memproduksi dalam jumlah *terabytes* setiap harinya [15]. Hal yang sangat menyulitkan tentunya jika ditangani dengan sistem tradisional.

Variety

Sangat banyak sumber data yang beredar di internet saat ini dan data-data tersebut dikelompokkan menjadi data terstruktur (*structured*), semi terstruktur (*semi-structured*) dan tidak terstruktur (*unstructured*). Data terstruktur adalah data yang diorganisasikan ke dalam entitas-entitas yang memiliki makna tertentu. Entitas sejenis akan dikelompokkan menjadi satu grup yang akan saling berelasi satu sama lain dan dimodelkan dalam bentuk kelas (class). Entitas yang berada dalam grup yang sama memiliki atribut yang sama juga. Data terstruktur biasanya dikelola oleh *data warehouse* dan dipresentasikan dalam bentuk tabel. Contohnya adalah data-data pelanggan yang disimpan beserta atributnya.

Data tidak terstruktur terdiri dari berbagai macam tipe data. Data jenis ini tidak dikelompokkan sebagaimana data terstruktur dan tidak membutuhkan format atau aturan yang spesifik. Ciri khasnya adalah data tidak bisa diprediksi dan dalam kehidupan sehari-hari data ini menguasai trafik sosial media. Data tidak terstruktur di-*generate* oleh mesin dan manusia. Beberapa contoh data yang di-*generate* oleh mesin adalah gambar satelit yang digunakan dalam aplikasi *google earth*, data sains misalnya tentang aktivitas kegempaan dan gambar atmosfer, fotografi dan video misalnya video kepadatan lalu-lintas, video pengawasan (*surveillance*), dan radar atau sonar misalnya aktivitas yang berhubungan dengan metereologi dan oseanografi.

Data semi terstruktur memiliki kemiripan dengan data terstruktur dimana data dikelompokkan dalam semantik entitas. Entitas yang memiliki kemiripan disatukan dalam sebuah grup. Tetapi tidak semua entitas dalam grup yang sama memiliki atribut yang sama. Pengurutan atribut juga bukan masalah yang penting. Ciri khasnya adalah ukuran (*size*) dan tipe (*type*) atribut yang sama dalam sebuah grup bisa saja berbeda. Contoh dari data ini adalah *BibTex file* dan standar dokumen SGML.

Velocity

Lalu-lintas data di internet setiap waktu mengalami peningkatan kecepatan linier. Karakteristik kecepatan pada *big data* tidak hanya terjadi pada waktu kedatangan data tersebut, tetapi juga terjadi pada proses dalam aliran data. *Velocity* berhubungan dengan "kecepatan" saat data dibuat, diproses, dan dianalisis. Jika data mengalir dari sumber ke tujuan dengan kecepatan yang konstan, maka akan tercipta lingkungan *real-time* dan bisa mempercepat proses pengambilan keputusan. Terkadang dalam perpindahan data terjadi *latency*. *Latency* terjadi ketika data dibuat atau diperoleh, dan ketika data tersebut dapat diakses [16]. Masalah ini harus ditangani dengan menciptakan lingkungan yang lebih *real-time* sehingga tidak akan mengganggu *core* bisnis perusahaan atau organisasi[17].

Veracity

Aspek ini muncul karena terjadi perbedaan kualitas data dan level sekuriti yang melekat pada data tersebut. Data seringkali tidak memiliki atribut yang lengkap atau bisa jadi data yang tidak lengkap (*missing information*) sehingga data menjadi tidak baik untuk dianalisis. *Big data* membutuhkan data yang pasti untuk diolah dan tidak selalu menggunakan metode *data cleansing* untuk memilah data-data yang tidak dapat diprediksi. Dibutuhkan solusi yang lebih matang untuk menjawab problem tersebut.

Value

Aspek *value* sangat penting terhadap perkembangan big data di era ini. *Value* mengacu pada proses eksplorasi massal terhadap nilai-nilai yang tersembunyi pada suatu kumpulan data yang sangat besar [13]. Ketika analisis terhadap *big data* dilakukan, pengguna dapat mengekstrak pengetahuan atau makna yang terkandung di dalamnya [14]. Pengguna mengajukan *query* ke penyimpanan data dan mengambil kesimpulan penting terhadap data-data yang telah difilter. Selanjutnya melakukan pengurutan sesuai dengan kondisi dan format tertentu untuk memperoleh infomasi penting didalamnya.

B. Dimensi dan Inkonsistensi *Big Data*

Inkonsistensi data muncul karena perbedaan dan konflik yang terjadi pada data yang sama yang disimpan di tempat yang berbeda. Inkonsistensi data menghasilkan informasi yang tidak dapat dipercaya, karena sangat sulit untuk menentukan informasi mana yang benar. Problem ini sangat menghambat dalam proses pengambilan keputusan karena terjadi konflik informasi didalamnya. Inkonsistensi data terjadi ketika ada data yang redundan. Data redundan adalah menumpuknya data-data yang sama yang tidak dibutuhkan di dalam *database*. Oleh karena itu desain *database* yang baik mengandung metode untuk mengeliminasi redundansi data.

Unsur pengetahuan (knowledge) sangat membantu dalam analisis *big data* untuk pengambilan keputusan yang tepat. Du Zang [18] telah membagi "pengetahuan" dalam beberapa level, dimana setiap level adalah ruang lingkup konten dari *big data*. Level pengetahuan dikelompokkan dalam *expertise*, *meta-knowledge*, *knowledge*, *information* dan *data*. *Expertise* adalah atribut khusus yang melekat pada sebuah *task* secara spesifik dan sifatnya relatif tidak berubah-ubah. Meta-knowledge adalah pengetahuan yang menjelaskan tentang pengetahuan. *Knowledge* merepresentasikan informasi khusus tentang beberapa domain dan sangat berpengaruh dalam proses pengambilan keputusan terhadap analisis *big* data. Tabel 1 menjelaskan tentang *knowledge content* dan propertinya-propertinya. Sementara tabel 2 berisi penjelasan tentang jenis penalaran terhadap *big data knowledge content*.

Du Zang [18] kemudian mengelompokkan inkonsistensi menjadi beberapa bagian, yaitu: *temporal inconsistencies*, *spatial inconsistencies*, *text inconsistencies* dan *functional dependency inconsistencies*. *Temporal inconsistencies* adalah inkonsistensi yang berkaitan dengan waktu. *Spatial inconsistencies* berkaitan dengan ruang/lokasi. *Text inconsistencies* berkaitan dengan inkonsistensi teks. *Functional dependency inconsistencies* adalah inkonsistensi yang mengikat aturan-aturan (*rules*) untuk berelasi dalam *database*.

Ada dua tipe inkonsistensi data yang sering dijumpai di banyak literatur saat ini [19]. Pertama adalah kejadian dimana data yang sama direpresentasikan berbeda di lokasi *database* yang berbeda sehingga menyebabkan keragaman skematik. Solusi untuk masalah ini dilakukan dengan melakukan penamaan ulang atribut, memetakan kembali asalnya (*domain mapping*), konversi nilai dan transformasi struktur. Kedua adalah kegagalan dalam pemeliharaan *database*. Realitanya adalah ketika item data yang sama terletak pada *database* yang berbeda, diharapkan memiliki nilai yang sama, tetapi disimpan dalam nilai yang berbeda. Solusinya bisa dengan melakukan *update* berkala terhadap data ataupun pengetahuan.

Analisis terhadap big data selalu dilengkapi dengan basis pengetahuan atau aturan. Jika aturan-aturan baku telah ditetapkan dan aturan tersebut dilanggar, maka terjadi inkonsistensi data. Di level aplikasi, sering terdapat pilihan untuk melakukan perbaikan terhadap data yang tidak konsisten, atau membiarkan data tersebut tetap inkonsisten.

Ada beberapa cara untuk melakukan deteksi inkonsistensi [8], yaitu: perbaikan inkonsistensi pada *relational database*, perbaikan problem struktural pada *semi-structured data*, mendeteksi inkonsistensi pada data terdistribusi dan data *streaming* dan memvalidasi XML atau dokumen web.

Banyaknya aturan yang ditetapkan tidak menjamin bahwa data akan bebas dari masalah inkonsistensi. Jika aturan atau basis pengetahuan disusun berdasarkan data yang tidak valid, maka inkonsistensi yang terjadi adalah akibat dari kegegalan terhadap aturan yang telah ditetapkan. Oleh karena itu, untuk menjamin data yang diperoleh tetap valid, seharusnya perubahan dilakukan pada data yang tidak konsisten atau aturan yang tidak tepat.

C. Algoritma untuk Menangani Inkonsistensi *Big Data*

Beberapa algoritma saat ini telah dikembangkan untuk menangani masalah inkonsistensi *big data*. Sebuah algoritma dikembangkan dalam bentuk framework yang diberi nama *inconsistency-induced learning* [20, 21, 22] atau i^2Learning yang merupakan algoritma adopsi dari *machine learning*. Framework ini mengakomodasi

pembelajaran yang terjadi secara terus-menerus. Tipe pembelajaran yang dilakukan berulang-kali tanpa henti sangat rentan untuk bertemu dengan inkonsistensi pada data. Ada sebuah agen yang bertugas untuk memecahkan masalah inkonsistensi pada setiap rekam kejadian ketika pembelajaran terjadi secara terus menerus untuk mencapai kesempurnaan. Sehingga kinerja agen bisa ditingkatkan di tiap level rekam kejadian yang telah diselesaikan olehnya.

Ide menarik dibalik pengembangan i^{2}Learning adalah mengidentifikasi penyebab inkonsistensi dan melakukan pencarian penyebab spesifik tersebut untuk menyelesaikan masalah inkonsistensi. Dalam ruang lingkup analisis *big data*, i^{2}Learning memegang peranan penting untuk meningkatkan kualitas data dengan mencegah inkonsistensi yang ada pada *dataset*, menambah pengetahuan ketika proses analisis berjalan, pemodelan atau interpretasi data yang besar, serta dapat membantu meningkatkan kinerja aplikasi big data [18].

Ada dua pendekatan untuk menyelesaikan masalah inkonsistensi data [8]. Pertama adalah dengan melakukan elminasi beberapa data yang tidak konsisten pada *database* dan kedua adalah pertanyaan adalah jawaban yang diselesaikan dengan metode sewajarnya tanpa memperbaiki kesalahan yang muncul pada *database*.

Algoritma yang mengadopsi penyelesaian masalah dengan pendekatan pertama adalah [23]. Proses menyelesaikan masalahnya adalah pertama membiarkan pengguna menentukan aturan kualitas, kemudian melakukan penolakan terhadap batasan dengan predikat *ad-hoc*. Bahasa yang digunakan adalah bebas dan formal tetapi menggunakan aturan-aturan numerik, misalnya dengan nilai "lebih besar dari" atau "kurang dari". Paradigma yang digunakan adalah menyelesaikan masalah secara holistik yang diyakini bisa mengkalkulasi secara otomatis dan perbaikan kualitas yang lebih baik. Hasil penelitian pada sampel *dataset* menunjukkan bahwa pendekatan secara holistik memiliki kualitas penyelesaian masalah inkonsistensi lebih baik dan lebih efisien dibanding algoritma dengan pendekatan lainnya.

Algoritma [24] menggunakan pendekatan kedua yang mengkhususkan penanganan pada inkonsistensi teks.Pembahasan dalam jurnal difokuskan pada metode *approximate string matching*. Sebagai contoh, diberikan sebuah himpunan string S dan *query string* v, dengan tujuan menemukan semua string s € S dimana pengguna bisa menspesifikasikan derajat kemiripan terhadap v. Sebuah himpunan S yang merupakan sekumpulan dokumen, sekumpulan halaman web, atau sebuah atribut dari tabel relasional dalam *database*. Kemiripin antara string sering diidentikkan dengan kesamaan fungsi yang dipilih berdasarkan dari karakteristik dari data dan aplikasi yang eksis saat ini. Prosedur penyelesaian masalah adalah dengan kebalikan indeks (*inverted indexes*), teknik filter (*filtering technique*) dan struktur data pohon (*tree data structures*) untuk mengevaluasi keberagaman himpunan dan kesamaan fungsi yang dimiliki.

3. Kesimpulan

Pembahasan di bab sebelumnya telah mengklasifikasiskan jenis inkonsistensi yaitu *temporal inconsistency*, *spatial inconsistency*, *text inconsistency* dan *functional dependency inconsistency*. Beberapa algoritma juga telah diusulkan untuk menyelesaikan masalah inkonsistensi data. Metode yang digunakan adalah dengan pembelajaran secara terus-menerus(*i^{2}Learning*), elminasi data, dan yang terakhir adalah metode dengan mengajukan pertanyaan adalah jawaban yang diselesaikan dengan metode sewajarnya tanpa memperbaiki kesalahan yang muncul pada *database*. Inkonsistensi data bukanlah masalah yang sederhana. Jika kita menginginkan kualitas data yang baik, maka problem inkonsistensi harus diselesaikan terlebih dahulu sebelum masuk ke area analisis *big data*.

Daftar Pustaka

[1] S. Sagiroglu and D. Sinanc, "Big Data: a Review", *in Collaboration Technologies and Systems (CTS) International Conference*, pp. 42-47, May 20-24, 2013.

[2] Intel IT Center, "Planning Guide: Getting Started with Hadoop", Steps IT Managers Can Take to Move Forward with Big Data Analytics, June 2012. http://www.intel.com/content/dam/www/public/us/en/documents/guides/getting-started-with-hadoop-planning-guide.pdf

[3] S. Singh and N. Singh, "Big Data Analytics", *in 2012 International Conference on Communication, Information & Computing Technology Mumbai India, IEEE*, October 2011.

[4] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh and A.H. Byers, "Big data: The next frontier for innovation, competition, and productivity", McKinsey Global Institute, 2011. http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx

[5] W. W. Eckerson: "Data quality and the bottom line: achieving business success through a commitment to high quality data". *Data Warehousing Institute*, 2002.

[6] W. Fan and F. Geerts, "Foundations of data quality management". *Morgan & Claypool* 2012.

[7] T. Redman, "The impact of poor data quality on the typical enterprise". *Commun. ACM* 1998.

[8] B. Saha and D. Srivastava, "Data quality: The other face of Big Data", *2014 IEEE 30th International Conference on Data Engineering (ICDE)*, pp. 1294-1297, March 31 to April 4, 2014.

[9] S. Sarsfield, "The butterfly effect of data quality", *The Fifth MIT Information Quality Industry Symposium*, 2011.

[10] I.A.T. Hashem, I. Yaqoob, N.B. Anuar, S. Mokhtar, A. Gani, S.U. Khan, The Rise of Big Data on Cloud Computing: Review and Open Research Issues, Elsevier Information System, pp. 98-115, January, 2015.

[11] P. Zikopoulos, K. Parasuraman, T. Deutsch, J. Giles, D. Corrigan, Harness the Power of Big Data The IBM Big Data Platform, McGraw Hill Professional, 2012.

[12] J.J. Berman, *Introduction, in: Principles of Big Data*, Morgan Kaufmann, Boston, pp. xix–xxvi, 2013.

[13] J. Tee: Handling the four 'V's of big data: volume, velocity, variety,and veracity, TheServerSide.com, 2013.

[14] Y. Zhai, Y.S. Ong, I.W. Tsang, "The Emerging Big Dimensionality", *IEEE Computational Intelligence Magazine*, August, 2014.

[15] A. Katal, M. Wazid, R.H. Goudar, "Big Data: Issues, Challenges, Tools and Good Practices", *Contemporary Computing (IC3) 2013 Sixth International Conference*, pp. 404-409, August 8-10, 2013.
[16] Analytics: The Real-World Use of Big Data, Executive Report, IBM Global Business Analytics and Optimization.
[17] M. Chen, S. Mao, Y. Liu, "Big data: a survey", *Mob. Netw. Appl*. 19 (2), pp. 1–39, 2014.
[18] D. Zhang, "Inconsistencies in Big Data", *Cognitive Informatics & Cognitive Computing (ICCI*CC)*, *2013 12th IEEE International Conference*, pp. 61-67, July 16-18, 2013.
[19] K. Wang and W. Zhang, "Detecting Data Inconsistency for Multi Database".
[20] D. Zhang and M. Lu, "Inconsistency-induced learning for perpetual learners", *International Journal of Software Science and Computational Intelligence*, Vol.3, No.4, pp.33-51, 2011.
[21] D. Zhang, "i^2Learning: perpetual learning through bias shifting", *in Proc. of the 24th International Conference on Software Engineering and Knowledge Engineering*, pp. 249-255, July, 2012.
[22] D. Zhang and M. Lu, "Learning through Overcoming Inheritance Inconsistencies", *in Proc. of the 13th IEEE International Conference on Information Reuse and Integration*, pp. 201-206, August, 2012.
[23] X. Chu, I.F. Ilyas and P. Papotti, "Holistic data cleaning: putting violations into context", *ICDE*, pp. 458-469, 2013.
[24] M. Hadjieleftheriou and D. Srivastava, "Approximate string processing", *Foundations and Trends in Databases*, pp. 267-402, 2011.

Biodata Penulis

Bawono Adi Sanjaya, memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Teknik Informatika UII Yogyakartadi tahun 2012. Saat ini sedang menyelesaikan pendidikan S2 di Program Pasca Sarjana Magister Teknik Informatika UGM Yogyakarta.

Selo Sulistyo, memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Teknik Elektro UGM tahun 1996. Gelar M. Eng diraihnya di tahun 2000 pada jurusan yang sama dan gelar M.Sc diperoleh dari Adger University, Norway di tahun 2003. Menyelesaikan program doktoral di Adger University pada tahun 2012. Saat ini aktif sebagai staf pengajar di JTETI UGM.