ITEJ Juli-2019, Volume 4 Nomor 1 Page 14 - 25

ITEJ
Information Technology Engineering Journals
eISSN : 2548-2157

Url : https://syekhnurjati.ac.id/journal/index.php/itej
Email : itej@syekhnurjati.ac.id

Development of Big Data Analytics Model

Saiful Rizal
School of Electrical Engineering and Informatics
Institut Teknologi Bandung Bandung, Indonesia
saiful.rizal@students.itb.ac.id

Abstract— The development of information technology produces very large data sizes, with various variations in data and complex data structures. Traditional data storage techniques are not sufficient for storage and analysis with very large volumes of data. Many researchers conducted their research in analyzing big data with various analytics models in big data. Therefore, the purpose of the survey paper is to provide an understanding of analytics models in big data for various uses using algorithms in data mining. Preprocessing big data is the key to turning big data into big value.

Keywords—big data, analytics model, algorithms, data mining

I. PENDAHULUAN

Pada konteks perkembangan informasi digital, semakin banyak perusahaan yang menganalisis sejumlah besar data yang disebut dengan *big data*. *Big data* adalah generasi berikutnya dari *data warehousing* dan *business analytics*. *Big data* memberikan *revenues cost* secara efisien bagi perusahaan [1]. *Big data* memiliki kumpulan data yang sangat kompleks, luas, dan data tidak terstruktur. *Big data* memberikan kemampuan untuk melakukan pengolah data yang bervariasi dengan jumlah yang besar sesuai dengan kebutuhan dari pebisnis, peneliti, dan pemerintah. *Big data* sangat membantu pengguna dalam mengolah data dengan cepat dan mendapatkan hasil data yang terpercaya. Tantangan sebenarnya adalah mengidentifikasi atau mengembangkan metode yang paling hemat biaya dan andal untuk mengekstraksi nilai dari semua *big data* yang sekarang tersedia, maka *data analytics* pada *big data* menjadi penting.

Big data mengubah proses bisnis yang belum pernah terjadi sebelumnya seputar analisis bisnis dan big data telah dihasilkan terutama dari komunitas web dan *e-commerce* [2]. *Web-based systems* (WBS), mulai dari sistem rekomendasi produk, platform *e-commerce*, jejaring sosial, permainan, hingga aplikasi CRM (Customer Relationship Management), dan SCM (Supply Chain Management), secara tradisional mengandalkan *data analytics* untuk beroperasi. Untuk WBS, *data analytics* adalah tentang menghasilkan prediksi dan wawasan yang dapat ditindaklanjuti untuk meningkatkan pengalaman pelanggan secara *real-time*, meningkatkan kecerdasan pasar dan pelanggan, memprediksi perilaku pelanggan, mengoptimalkan efisiensi operasional, mempersonalisasikan penyediaan layanan, mencegah ancaman keamanan dan penipuan, meminimalkan risiko, dan berinovasi proses dan layanan [3].

Data analytics pada *big data* mempunyai tujuan untuk mengubah *big data* menjadi *big value*. B*ig data analytics* adalah proses mengeksplorasi sejumlah besar data dan berbagai jenis data dengan tujuan memperoleh informasi yang berharga [4]. Media sosial dan sensor sebagai bagian dari

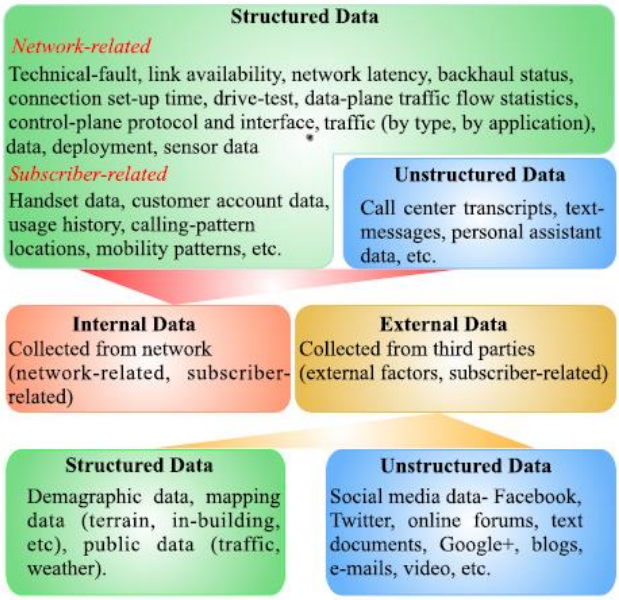
teknologi informasi memberikan sumber data triliunan byte. Sumber data dari teknologi sebagai sumber *big data*. Setiap kumpulan data set sebagai sumber *big data* memiliki ukurang yang sangat besar dengan kompleksitas datanya yang tinggi sehingga sulit untuk ditangani jika hanya menggunakan pemrosesan data secara tradisional atau menggunakan basis data manajemen biasa.

Dalam 25 tahun terakhir, data telah tumbuh secara masif di berbagai bidang dengan berbagai jenis. Menurut laporan statistik dari *International Data Corporation* (IDC), pada tahun 2011, volume data keseluruhan yang dibuat di dunia adalah 1,8ZB yang meningkat hampir sembilan kali dalam lima tahun berikutnya [5]. Pada tahun 2015 IDC melaporkan volume data keseluruhan yang dibuat di dunia adalah 7,9ZB [6].

Gambar 1. *Total Digital Universe*.

Dengan peningkatan volume data universal, teknologi *big data* dan proses analisisnya umumnya digunakan untuk memberikan deskripsi tentang dataset yang besar. Dibandingkan dengan dataset tradisional lainnya dan prosesnya, *big data* mencakup data semi terstruktur dan tidak terstruktur yang membutuhkan analisis secara terkini. *Big data* juga mendapatkan perincian tentang prospek baru untuk menentukan nilai-nilai baru, mendukung untuk meningkatkan pemahaman mendalam tentang nilai-nilai tersembunyi, dan juga menimbulkan tantangan baru, Misalnya, bagaimana mengatur dan memanipulasi dataset. Volume informasi dari berbagai sumber semakin besar, juga memberikan beberapa masalah yang menantang yang menuntut resolusi cepat. Proses visualisasi *big data* adalah proses penting lainnya yang mengambil tempat penting dalam masalah *big data analytics*. Karena melalui visualisasi data hanya laporan akhir *data analytics* yang akan divisualisasikan.

Tantangan teknis berkaitan dengan mengatasi 7V *big data* (*Volume, Variety, Velocity, Veracity, Variability, Value, Visualization*) untuk mendukung *data analytics*. B*ig data analytics* menggunaan set data yang sangat besar (dari Terabytes ke Exabytes) dan kompleks sehingga mereka membutuhkan teknologi penyimpanan, manajemen, analisis, dan untuk visualisasi data yang canggih dan unik [7]. Keberhasilan dalam *big data analytics* untuk WBS tergantung pada pembangunan infrastruktur yang secara efektif mengatur komponen teknologi untuk memproses, menyimpan, mengintegrasikan, dan memvisualisasikan sejumlah besar data dari berbagai jenis (terstruktur, terstruktur dan tidak terstruktur) dari internal dan eksternal sumber termasuk media sosial, *IoT*, *web logs*, dan *web-crawler information* [8].



Gambar 2. *Data Set* Dan Sumber Data Yang Tersedia Untuk Untuk *Big Data Analytics.*

(Gambar diambil dari [9]).

II. GAMBARAN UMUM MODEL ANALITICS PADA BIG DATA

A. Big Data

Mengingat popularitasnya saat ini, definisi *big data* sangat beragam. IDC merupakan pelopor dalam mempelajari data besar dan dampaknya, mendefinisikan *big data* dalam laporan 2011 yang disponsori oleh EMC (pemimpin *cloud computing*) [15]: `` Teknologi *big data* menggambarkan generasi baru teknologi dan arsitektur, yang dirancang untuk secara ekonomis mengekstraksi nilai dari volume yang sangat besar dari berbagai sumber data, dengan mengaktifkan pengambilan data dengan kecepatan tinggi, pencarian, dan atau analisis".

Pada tahun 2011, laporan Mckinsey [10] mendefinisikan *big data* sebagai kumpulan data yang mempunyai ukuran di atas kemampuan perangkat lunak basis data biasa untuk menangkap, menyimpan, mengelola, dan menganalisis. Definisi ini subjektif dan tidak mendefinisikan besar data dalam hal metrik tertentu. Namun, ia memasukkan aspek evolusi dalam definisi (dari waktu ke waktu atau lintas sektor) dari apa yang harus dianggap sebagai *big data.*

Menurut penelitian Han Hu dkk. [11] *big data* diciptakan untuk menangkap makna tren yang muncul dengan menunjukkan karakteristik unik lainnya dibandingkan dengan data tradisional, *big data* umumnya tidak terstruktur dan memerlukan lebih banyak analisis wakt. Perkembangan ini membutuhkan arsitektur sistem baru untuk akuisisi data, transmisi, penyimpanan, dan mekanisme pemrosesan data skala besar.

Tabel 1. Perbandingan antara *big data* dan data tradisional

	Traditional Data	**Big Data**
Volume	GB	constantly updated (TB or PB currently)
Generated Rate	per hour, day, ...	more rapid
Structure	structured	semi-structured or un-structured
Data Source	centralized	fully distributed
Data Integration	easy	difficult
Data Store	RDBMS	HDFS, NoSQL
Access	interactive	batch or near real-time

B. Arsitektur Sistem *Big Data*

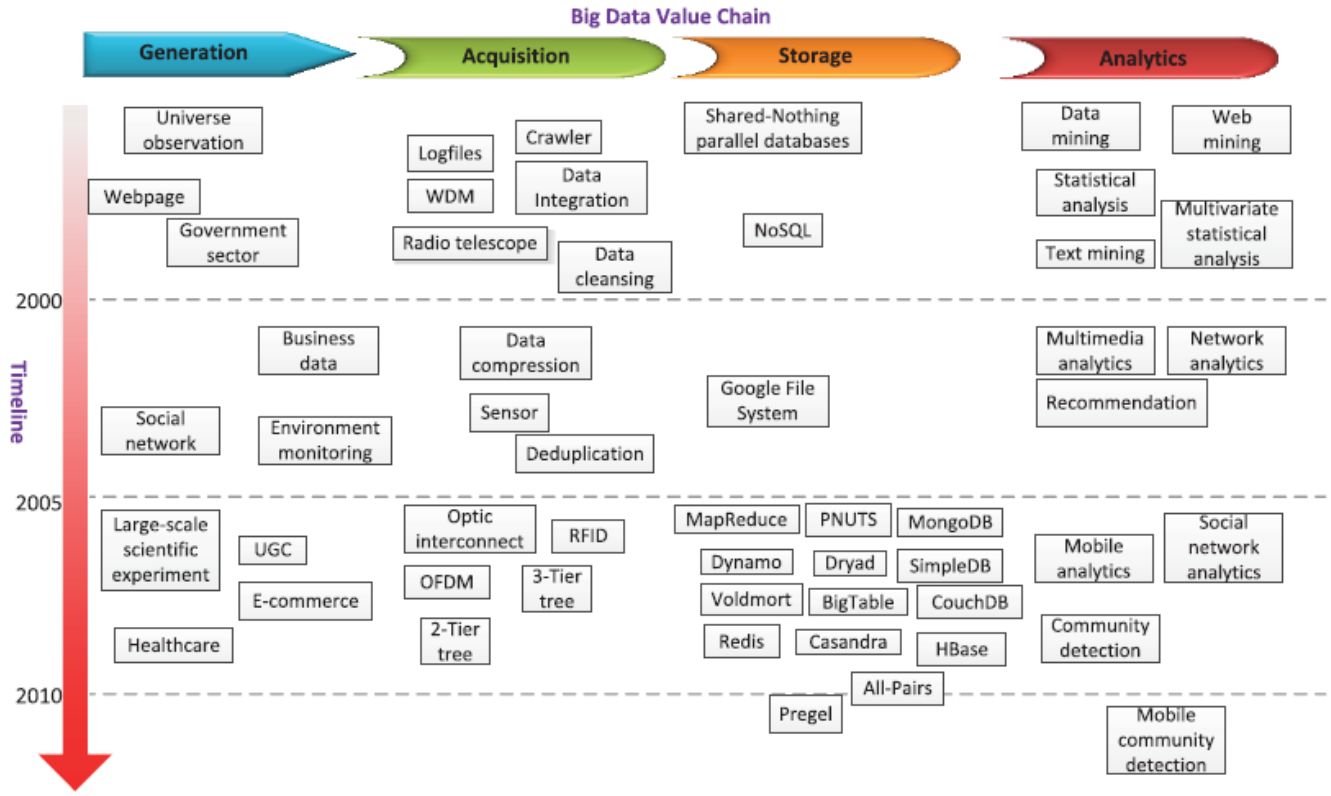
Menurut penelitian Han Hu dkk. [11] fokus pada *value chain* untuk *big data analytics*. Secara khusus, menggambarkan *big data value chain* yang terdiri dari empat tahap (*generation, acquisition, storage*, dan *processing*). Rincian untuk setiap fase dijelaskan sebagai berikut.

Data generation memperhatikan bagaimana data dihasilkan. Dalam hal ini, istilah *big data* dimaksudkan untuk berarti kumpulan data besar, beragam, dan kompleks yang dihasilkan dari berbagai sumber data longitudinal dan atau terdistribusi, termasuk sensor, video, dan sumber digital lain yang tersedia. Biasanya, kumpulan data ini dikaitkan dengan berbagai tingkatan nilai spesifik domain. Namun, ada banyak tantangan teknis dalam mengumpulkan, memproses, dan menganalisis kumpulan data yang menuntut solusi baru untuk merangkul kemajuan terbaru dalam teknologi informasi dan komunikasi.

Data acquisition mengacu pada proses memperoleh informasi dan dibagi lagi menjadi pengumpulan data, transmisi data, dan pra-pemrosesan data. Pertama, karena data dapat berasal dari beragam sumber, situs web yang menampung teks, gambar, dan video yang diformat. Pengumpulan data mengacu pada teknologi pengumpulan data khusus yang memperoleh data mentah dari lingkungan produksi data tertentu. Kedua, setelah mengumpulkan data mentah, kita memerlukan mekanisme transmisi berkecepatan tinggi untuk mengirimkan data ke sistem penyimpanan penyimpanan yang tepat untuk berbagai jenis aplikasi analitik. Akhirnya, kumpulan data yang dikumpulkan mungkin mengandung banyak data yang tidak berarti, yang secara tidak perlu meningkatkan jumlah ruang penyimpanan dan memengaruhi analisis data konsekuen. Sebagai contoh, redundansi adalah umum di sebagian besar kumpulan data yang dikumpulkan dari sensor yang digunakan untuk memantau lingkungan, dan kita dapat menggunakan teknologi kompresi data untuk mengatasi masalah ini. Dengan demikian, kita harus melakukan operasi pra-pemrosesan data untuk penyimpanan yang efisien.

Masalah *data storage* terus-menerus menyimpan dan mengelola kumpulan data dengan skala besar. Sistem penyimpanan data dapat dikategorikan menjadi dua bagian: infrastruktur perangkat keras dan manajemen data. Infrastruktur perangkat keras terdiri dari kumpulan sumber daya teknologi informasi dan komuniasi bersama yang diorganisasikan untuk berbagai tugas sebagai respon terhadap permintaan organisasi. Infrastruktur perangkat keras harus dapat ditingkatkan untuk dapat dikonfigurasi ulang secara dinamis dalam mengatasi berbagai jenis aplikasi yang ada. Perangkat lunak manajemen data digunakan di atas infrastruktur perangkat keras untuk memelihara dataset skala besar. Selain itu, untuk menganalisis atau berinteraksi dengan data yang disimpan, sistem penyimpanan harus menyediakan beberapa fungsi antarmuka, permintaan cepat, dan model pemrograman lainnya.

Analisis data memanfaatkan metode atau alat analitik untuk memeriksa, mengubah, dan memodelkan data untuk mengekstraksi nilai. Banyak bidang aplikasi memanfaatkan peluang yang disajikan oleh data berlimpah dan metode analisis khusus untuk memperoleh dampak yang diinginkan. Meskipun berbagai bidang menimbulkan persyaratan aplikasi dan karakteristik data yang berbeda, beberapa bidang ini dapat memanfaatkan teknologi mendasar yang serupa. Penelitian analitik yang muncul dapat diklasifikasikan ke dalam enam bidang teknis kritis: data analytics, text analytics, multimedia analytics, web analytics, networkanalytics, dan mobile analytics. Klasifikasi ini dimaksudkan untuk melihat karakteristik data utama dari setiap area.



Gambar 2. *Big Data Technology Map*

Penelitian big data adalah bidang luas yang terhubung dengan banyak teknologi yang memungkinkan. Pada gambar.2 menyajikan peta teknologi *big data*. Dalam peta teknologi ini, penelitian [11] mengaitkan daftar teknologi yang memungkinkan, baik sumber terbuka maupun kepemilikan, dengan berbagai tahapan dalam big data value chain.

Peta ini mencerminkan tren perkembangan *big data*. Pada tahap pembuatan data, struktur *big data* menjadi semakin kompleks, dari terstruktur atau tidak terstruktur menjadi campuran dari berbagai jenis, sedangkan sumber data menjadi semakin beragam. Pada tahap akuisisi data, pengumpulan data, pra-pemrosesan data, dan penelitian transmisi data muncul pada waktu yang berbeda. Selain itu, teknologi atau metode berkualitas yang terkait dengan tahapan yang berbeda dapat dipilih dari peta ini untuk menyesuaikan sistem *big data*.

C. Karakteristik Big Data

Menurut Doug Laney [12] karakteristik *big data* adalah *Volume*, *Velocity* dan *Variety*, yang dikenal sebagai 3V:

Volume: Organisasi mengumpulkan data berasal dari sumber data yang beragam, termasuk transaksi komersial, data media sosial dan informasi dari sensor atau data mesin-ke-mesin.

Velocity: Aliran data masuk dengan kecepatan yang tinggi dan harus dialokasikan dengan cara yang sesuai. Berbagai jenis sensor IoT, tag RFID, dan *smart metering* mendorong kebutuhan untuk menangani aliran data secara *real time*.

Variety: Data hadir dalam berbagai jenis format seperti terstruktur, data berbentuk numerik yang tersimpan dalam database tradisional hingga data dokumen teks tidak terstruktur, email, video, audio, stok dan transaksi keuangan.

Tetapi ketiga V ini berkembang menjadi tujuh V dengan menambahkan empat lagi V seperti , *Veracity*, *Variability*, *Value*, dan *Visualization*, yang dikenal dengan 7V [7]:

Veracity: *Veracity* mengacu pada tingkat di mana seseorang mempercayai informasi yang digunakan untuk mengambil keputusan. Singkatnya, akurasi data diperiksa untuk diproses lebih lanjut agar inkonsistensi dari kumpulan data dapat tertangani.

Variability: Ini menunjukkan variasi data dalam varietas tertentu. Ini juga mempertimbangkan inkonsistensi aliran data. Kualitas data yang diambil dapat sangat bervariasi, yang memengaruhi analisis yang akurat.

Value: Pengguna menjalankan *queries* terhadap data yang disimpan dan mengumpulkan hasil penting dari data yang difilter. Dengan value yang sesuai mendapatkan peringkat data sesuai dengan persyaratan.

Visualization: Menemukan cara untuk merepresentasikan informasi yang membuat temuan menjadi jelas. Ini adalah salah satu tantangan big data.

D. Analytics Model

Big data analytics adalah proses menggunakan algoritma analisis yang berjalan pada platform pendukung yang kuat untuk mengungkap potensi yang tersembunyi dalam *big data*, seperti pola tersembunyi atau korelasi yang tidak diketahui [11]. Selain itu *big data analytics* merupakan proses mengeksplorasi sejumlah besar data dan berbagai jenis data dengan tujuan memperoleh informasi yang berharga [4].

Terdapat beberapa *model analytics* dalam *big data analytics*, mulai dari *descriptive* analytics, *diagonostic analytics*, *prediktive analytics*, dan *preskriptive analytics* seperti yang ditunjukkan pada Gambar. 3 [9], di mana tiga (analisis deskriptif, prediktif, dan preskriptif) menjadi dominan.

Penelitian kibria dkk. [9] pada operator jaringan berada dalam *descriptive* analytics untuk menggunakan alat visualisasi untuk mendapatkan wawasan tentang apa yang telah terjadi, kinerja jaringan, profil lalu lintas, dll. *Descriptive* analytics mendeskripsikan dan merangkum fitur-fitur dasar yang telah tersirat pada seluruh populasi dari dataset yang diberikan. Ringkasan ditampilkan dalam berbentuk grafik untuk memfasilitasi proses pengambilan keputusan. Dataset yang diberikan juga disajikan sesuai kecenderungan utama, variabilitas dan penyebaran.

Operator jaringan dapat menggunakan *diagonostic analytics* untuk mengetahui akar penyebab dari anomali jaringan dan mencari tahu KPI yang salah dan fungsi elemen jaringan. Untuk mendapatkan *diagonostic analytics*, alat analisis menggunakan teknik seperti *drill-down*, pembelajaran mendalam, penemuan data, korelasi, dll.

Prediktive analytics adalah alat yang sangat baik untuk membuat prediksi dapat menghasilkan perkiraan tentang apa yang mungkin terjadi, misalnya, lokasi pelanggan di masa mendatang, pola lalu lintas masa depan dan kemacetan jaringan, dll. *Prediktive analytics* memberikan peristiwa prediksi berdasarkan pada data *real time* dan yang diarsipkan dengan memanfaatkan berbagai teknik statistik seperti *machine learning*, *data mining*, pemodelan sebagai beberapa proses analitik, dan *game-theoretic analysis*.

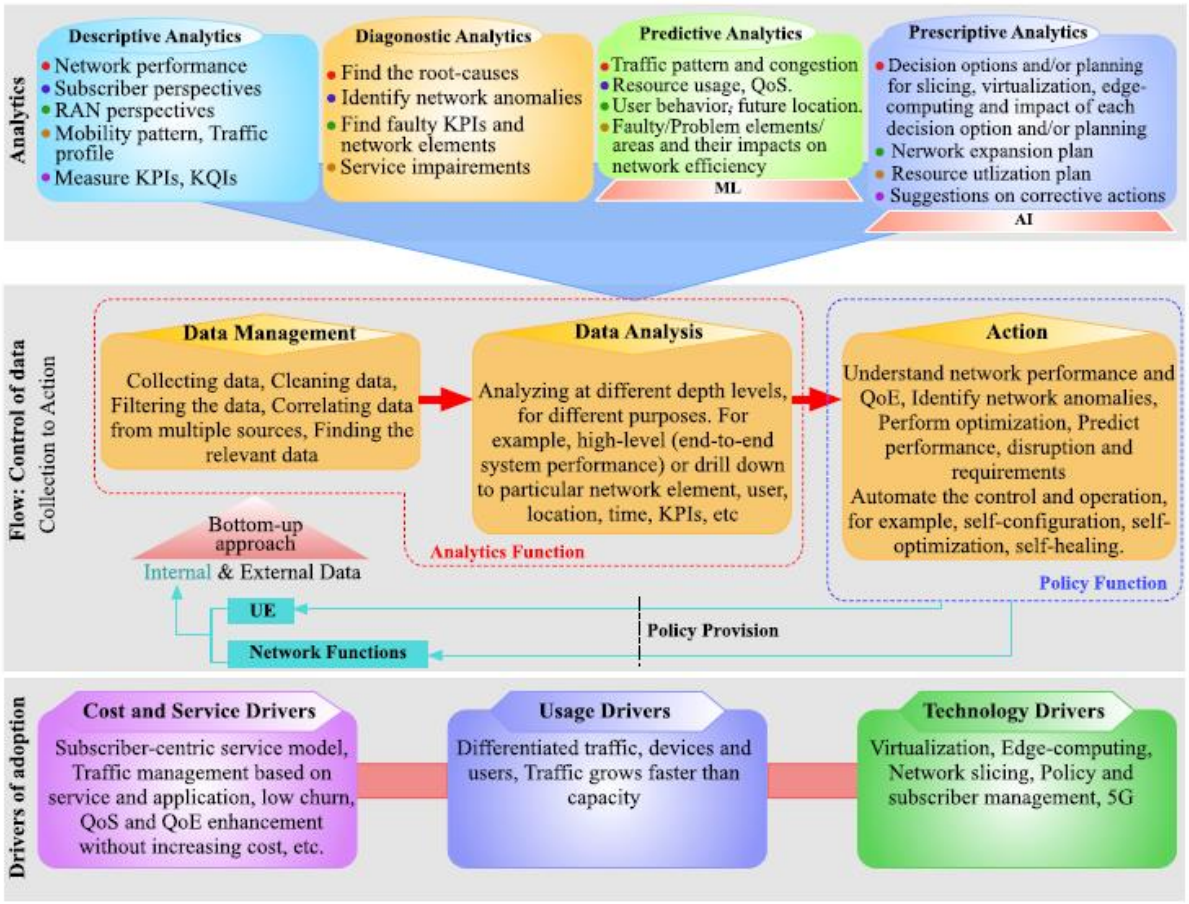
Preskriptive analytics berjalan selangkah lebih maju dari sekadar memprediksi peristiwa pada masa depan dengan menyarankan opsi memberikan keputusan, *virtualization*, *edge-computing*, dll., Bersama dengan implikasi dari setiap opsi keputusan. Oleh karena itu, *preskriptive analytics* memerlukan *predictive model*, data yang dapat ditindaklanjuti, dan sistem umpan balik untuk melacak hasil yang dihasilkan oleh tindakan yang diambil. Opsi keputusan (mis., Untuk perluasan jaringan, penggunaan sumber daya) diproduksi dengan mempertimbangkan jaringan preferensi operator, kendala sistem dll. *Preskriptive analytics* juga dapat menyarankan tindakan terbaik untuk setiap target yang telah ditentukan.

Salah satu tahap awal dari setiap upaya analitik adalah penggabungan studi investigasi sumber daya data. Dengan kata lain, sebelum laporan dibuat atau pemodelan kuantitatif dilakukan, seorang analis perlu lebih memahami apa yang ada dalam file data. Proses investigasi ini melibatkan untuk melakukan analisis distribusi berbagai variabel data, mungkin menghitung metrik maksimum, minimum, dan varians seperti standar deviasi. Proses ini memberikan karakter deskriptif dari apa variabel data terdiri dan membuat analisis tambahan lebih kuat, karena mengidentifikasi adanya masalah seperti bias data atau *outlier*, dan bahkan kesalahan dalam sumber daya data.

Algoritma *data mining* dan teknik analisis data memainkan peran penting dalam *big data analytics* dalam hal pengurangan dimensi, biaya komputasi, kebutuhan memori, manajemen dan akurasi hasil akhir [12]. Metode *data mining* sebagian besar berasal dari statistik, *machine learning, artificial intelligence*, dan *database systems* [13].

Big data memberikan tantangan dan peluang bagi *data mining* untuk mengembangkan model yang ditingkatkan. Peralatan komputasi analitis di dalam memori saat ini tidak lagi menghalangi ukuran data yang dapat untuk dianalisis. Keuntungan utama adalah dapat menganalisis lebih banyak populasi dengan berbagai analisis klasik dan modern. *Data mining* juga dapat mencoba lebih banyak opsi konfigurasi untuk algoritma tertentu, misalnya, topologi *neural network* termasuk fungsi aktivasi dan kombinasi yang berbeda, karena model berjalan lebih cepat dalam.

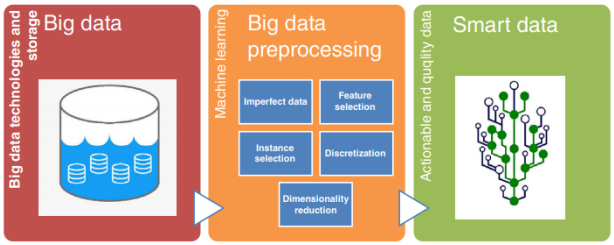
Menggunakan beberapa metode algoritma dalam *data mining* sangat penting dalam pengembangan model, sehingga *data mining* harus cepat dalam pengembangan model. Intinya adalah bahwa *big data* tumbuh semakin besar, dan *data mining* perlu secara signifikan mengurangi waktu siklus yang diperlukan untuk menganalisis data besar menjadi membuat model yang siap digunakan.



Gambar 3. Analytics Model, Control Data, Driver Adoption Pada Analitik Dalam Sistem Komunikasi Operator Seluler.

III. ANALYTICS MODEL PADA BIG DATA

Pada *International Conference on Data Mining '06* 21 desember 2006 melakukan pemungutan suara terbuka dengan seluruh 145 peserta menghasilkan 10 algoritma terbaik dalam *data mining* [14]. Hasil 10 algoritma terbaik yaitu, C4.5 dan tururannya , *k-means algorithm, support vector machines* (svm), *apriori algorithm, expectation maximization algorithm, pagerank, k-nearest neighbors, naıve bayes, classification and regression trees, adaboost.*



Gambar 4. Preprocessing *big data* adalah kunci untuk mengubah big data menjadi data yang berkualitas dan cerdas.

A. C4.5

C4.5 merupakan metode yang membangun pengklasifikasi yang umum digunakan menggunakan *decision tree* untuk melakukan prediksi. Metode *decision tree* berguna untuk menemukan hubungan antara sejumlah inputan variabel dengan sebuah target variabel. *Decision tree* menggabungkan antara permodelan dengan mengeksplorasi data sebagai proses model awal pada pemodelan dan dapat dijadikan sebagai model akhir dari beberapa teknik lain.

Pada penelitian Lu Han dkk. [16] berdasarkan jaringan elemen pengetahuan darurat serta metodologi representasi hirarkis untuk model respons darurat, penelitian ini telah menyajikan metode untuk memecahkan masalah yang terkait dengan manajemen darurat di dua tingkat, yaitu proses konstruksi model dan proses pemilihan model. Dalam proses konstruksi model, koefisien keseimbangan, yang dibangun secara dinamis dengan menyimpulkan jaringan pengetahuan peristiwa yang mendasarinya, ditambahkan ke algoritma C4.5 yang ditingkatkan untuk mencapai proses konstruksi model baru. Proses baru ini tidak hanya memanfaatkan algoritma C4.5 yang lebih baik dalam menyelesaikan masalah yang melekat pada penyimpangan atribut utama, tetapi juga mencakup kelemahan dalam koefisien kesetimbangan yang ditentukan oleh pengetahuan para ahli sebelumnya. Selain itu, proses baru memenuhi persyaratan untuk pemilihan dinamis yang didasarkan pada informasi kontekstual selama proses pemilihan model.

B. K-Means Algorithm

K-Means Algorithm menggunakan pendekatan non-hirarkis untuk membentuk kluster yang baik adalah dengan menentukan sebelumnya jumlah klaster yang diinginkan, k, dan menetapkan setiap kasus ke salah satu klaster k sehingga dapat meminimalkan ukuran dispersi di dalam kluster. Analisis *cluster* memiliki aplikasi yang luas, termasuk segmentasi pelanggan, pengenalan pola, studi biologi, dan klasifikasi dokumen web. Metode *K-Means Algorithm* memiliki proses untuk membagi data ke dalam sejumlah kluster sehingga kelompok data yang terbentuk berdasarkan kelompok data yang memiliki karakteristik yang ke dalam satu cluster yang sama dan karakteristik data yang berbeda di kelompokan ke dalam cluster yang lain.

Penelitian oleh Pengfei Shan [17] menggunakan metode entropi maksimum grey-gradient untuk mengekstraksi fitur dari gambar, menggunakan metode *k-means algorithm* untuk mengklasifikasikan gambar, dan menggunakan metode evaluasi *Average Precision* (AP) dan Intersection Over Union (IU) untuk mengevaluasi hasil. Hasil penelitian menunjukkan bahwa segmentasi gambar dapat direalisasikan dengan baik dengan menggunakan metode K-mean.

C. Support Vector Machines (SVM)

Support Vector Machines (SVM) merupakan salah satu teknik Machine Learning (ML) paling efektif. SVM telah menarik perhatian banyak peneliti dalam analisis geospasial. Sementara di sebagian besar studi geospasial, SVM telah diterapkan pada klasifikasi *remotely sensed data.* SVM tidak hanya mampu secara efektif mempertimbangkan variabel kontinu dan kategorikal, data terdistribusi tidak normal, hubungan non-linier, data *noisy* dan kompleks, dan *dataset* pelatihan dengan outlier, juga dapat menghindari overfitting dan memastikan kinerja generalisasi yang baik.

Penelitian oleh Firoozeh Karimi dkk. [17] menggunakan teknik SVM untuk memodelkan ekspansi kota. Berbagai model SVM dievaluasi untuk memilih model perluasan kota yang paling efisien di

Guilford County, NC, selama periode 2001-2011. Proses pemodelan meliputi eksplorasi berbagai metode pengambilan sampel, pemeriksaan berbagai variabel prediktor, penelitian parameterisasi SVM dan regulasi kernel, dan pengembangan berbagai metrik evaluasi. Penerapan tiga metode pengambilan sampel termasuk pengambilan sampel acak, pengambilan sampel semua sel yang berubah, dan pengambilan sampel berimbang mengungkapkan efek signifikan metode ini terhadap kinerja model. Model perluasan kota berbasis SVM dapat digunakan untuk mengevaluasi dampak ekspansi kota terhadap fragmentasi habitat, pencemaran lingkungan, masalah hidrologi, gangguan satwa liar, penggundulan hutan, perusakan ladang pertanian, dan pemanasan regional dan global. Oleh karena itu, model ini akan sangat membantu perencana kota, pembuat kebijakan lingkungan, dan ahli geografi untuk memperbaiki kegiatan yang berkaitan dengan interaksi antara lingkungan alam dan bangunan.

D. Apriori Algorithm

Apriori Algorithm adalah sebuah algoritma pencarian pola yang memiliki teknik *data mining* untuk mempelajari korelasi dan asosiasi antar variabel dalam database Algoritma ini ditujukan untuk mendapatkan kombinasi item yang mempunyai nilai berulang tertentu dengan kesesuaian kriteria yang diinginkan. Hasil yang didapat digunakan untuk membantu dalam pengambilan keputusan.

Penelitian oleh Santosh Kumar [18] membuat model sistem rekomendasi untuk memprediksi dan merekomendasikan konsumsi berbagai item pertanian menggunakan *apriori algorithm*. Sistem yang dikembangkan mampu membuat prediksi dan rekomendasi berdasarkan perilaku pembelian pelanggan sebelumnya dan rekomendasi item barang yang dibeli. Model yang dibuat dapat membuat prediksi barang yang dikonsumsi oleh semua pelanggan, sehingga petani dapat memproduksi barang sesuai pilihan mereka. Dengan demikian prediksi kumulatif dapat membantu petani untuk merencanakan dan membuat budidaya pertanian untuk musim apa pun, sehingga tidak akan ada pemborosan barang yang diproduksi oleh petani. Dengan demikian sistem yang dikembangkan, membantu pelanggan untuk merekomendasikan lebih banyak barang daripada yang diprediksi.

E. Expectation Maximization Algorithm

Expectation Maximization Algorithm (*EM Algorithm*) merupakan algoritma yang digunakan untuk menemukan nilai estimasi kemungkinan maksimal dari parameter dalam sebuah model probabilistik. *EM Algorithm* telah banyak digunakan untuk estimasi parameter dalam identifikasi proses data. *EM Algorithm* adalah algoritma untuk estimasi kemungkinan maksimum dari parameter dan memastikan konvergensi fungsi kemungkinan. Dengan adanya variabel yang hilang dan dalam masalah yang tidak terkondisikan, algoritma EM sangat membantu desain algoritma identifikasi yang lebih kuat. Situasi seperti itu sering terjadi di lingkungan industri. Pengamatan yang hilang karena kerusakan sensor, beberapa kondisi operasi proses dan informasi waktu tunda yang tidak diketahui adalah beberapa contoh yang dapat menggunakan algoritma EM.

Penelitian oleh Nima Sammaknejad dkk. [19] melakukan ulasan tentang aplikasi algoritma EM untuk mengatasi masalah pengamatan variabel yang hilang akibat kerusakan sensor. Dalam artikel ini, aplikasi algoritma EM dalam identifikasi proses data-driven telah diperkenalkan. Metode identifikasi untuk sistem linear parame-ter varying / switching, waktu tunda, ruang keadaan dan model yang dapat diperbarui serta hidden markov model melalui langkah-langkah Expectation (E) dan Maximization (M) ditinjau. Keuntungan langkah E untuk estimasi parameter yang kuat dengan adanya outlier dan data yang hilang telah dijelaskan. Algoritma EM menjamin peningkatan monotonik dari fungsi kemungkinan yang diharapkan. Algoritma EM pada dasarnya adalah varian dari estimasi kemungkinan maksimum dan mampu menangani data yang hilang, penerapannya dalam deteksi kesalahan, deteksi sinyal, dan penyaringan di hadapan data yang hilang adalah arah yang menarik untuk dijelajahi. Masalah hilangnya data sering terjadi pada praktis *engineering systems*. Dengan meningkatnya dimensi dan volume data dalam analisis *big data*, efisiensi komputasi dari algoritma EM dan kemampuannya dalam menangani data dimensi tinggi. Kinerja algoritma EM sensitif terhadap nilai awal. Meskipun banyak upaya dalam menemukan nilai awal yang tepat untuk memulai iterasi EM, pilihan terbaik dari nilai awal tetap sebagai masalah terbuka.

F. PageRank

PageRank adalah algoritma pencarian yang mempunyai fungsi untuk mendapatkan situs web yang lebih populer. Implementasi p*agerank* terdapat pada fitur utama dalam Google untuk melakukan pencarian. PageRank merupakan algoritma analisis *hyperlink* yang dirancang untuk menstandarisasi signifikan relatif dari beberapa objek yang terhubung dalam jaringan objek data. Algoritma ini memproses jenis analisis jaringan yang ingin mengeksplorasi hubungan antara objek dan peringkatnya.

Penelitian oleh Eonho Kim dan Minsoo Jeon [20] mengusulkan model PageRank diusulkan untuk mengatasi masalah yang ada dalam memperkirakan peringkat olimpiade taekwondo. Sebagai hasilnya, peneliti mengidentifikasi kemungkinan menerapkan model PageRank untuk menghitung peringkat olimpiade taekwondo dengan peringkat dapat dihitung lebih akurat dan wajar.

*G. K-Nearest Neighbors (*k-NN*)*

K-Nearest Neighbors (k-NN) merupakan sebuah metode klasifikasi terhadap sekumpulan data berdasarkan pembelajaran data yang sudah terklasifikasikan sebelumya. Termasuk dalam *supervised learning*, dimana hasil *query instance* yang baru diklasifikasikan berdasarkan mayoritas kedekatan jarak dari kategori yang ada dalam k-NN. K-NN termasuk kelompok *instance-based learning*. Algoritma ini juga merupakan salah satu teknik *lazy learning* dikarenakan hanya menyimpan sebagian atau seluruh data latih, kemudian menggunakan data latih tersebut ketika proses prediksi. k-NN dilakukan dengan mencari kelompok k objek dalam *data training* yang paling dekat (mirip) dengan objek pada data baru atau *data testing.*

Penelitian oleh Isaac Triguero dkk. [15] telah membahas peran k-NN sebagai alat yang baik untuk mendapatkan *smart data* yang merupakan data berkualitas tinggi untuk dilakukan analisis. Banyak dari teknik preprocessing data ini didasarkan pada kerja yang mendasari algoritma k-NN yang memungkinkan proses preprocessing yang sederhana namun efektif. Proses-proses ini ternyata bermanfaat tidak hanya untuk algoritma k-NN untuk apa yang awalnya dirancang, tetapi juga untuk banyak teknik *data mining* lainnya.

H. Naıve Bayes

Naıve Bayes merupakan algoritma yang digunakan untuk memprediksi peluang yang mungkin terjadi di masa yang akan datang yang didasarkan pada pengalaman/kejadian di masa sebelumnya. Naive Bayes menggunakan metode yang sama untuk memprediksi probabilitas kelas yang berbeda berdasarkan berbagai atribut. Algoritma ini sebagian besar digunakan dalam klasifikasi teks dan dengan masalah memiliki beberapa kelas.

Penelitian oleh Kevin Joy Dsouza dan Zahid Ahmed Ansari [21] dilakukan dengan menggunakan pengklasifikasi *naıve bayes* untuk mengklasifikasikan data medis. Kesesuaian pengklasifikasian dan akurasi klasifikasi diukur menggunakan kriteria kinerja yang berbeda. Penelitian ini bermanfaat bagi para peneliti dan pengembang dalam memahami dan menggunakan teknik klasifikasi dalam diagnosis medis.

*I. Classification and Regression Trees (*CART*)*

Classification and Regression Trees (CART) yaitu metode pohon regresi dan pohon klasifikasi. Tujuan klasifikasi pada metode CART menghasilkam sebuah pohon klasifikasi yang diperoleh melalui penyekatan berulang terhadap sstruktur data. Metode CART sangat efektif untuk diterapkan pada pengamatan dengan data yang relatif banyak.

Dari Penelitian oleh Y Ji dkk [22] beberapa kesimpulan dapat ditarik, algoritma CART dengan fungsi ganda klasifikasi dan regresi sangat cocok untuk memecahkan masalah dengan menghitung penggunaan energi per jam dari unit terminal HVAC yang tidak dapat diukur secara langsung. Model generik dan proses terstandardisasi dibuat, dan tipe variabel input dan jumlah data pelatihan dianalisis dalam penelitian ini. Metode ini dapat diprogram sepenuhnya dan sangat cocok untuk memproses sejumlah besar data di berbagai jenis bangunan..

J. Adaboost

Adaboost bertujuan meningkatkan kemampuan *weak classifier*, untuk meningkatkan akurasi model. Algoritma *adaboost* termasuk dalam *boosting algorithm*, yang artinya merupakan *combined learning algorithm*. Gagasan utama dari algoritma ini adalah untuk membangun beberapa model "*weak*" dan menggabungkannya di bawah aturan tertentu untuk mendapatkan model baru berkekuatan tinggi. Model "*weak*" tidak perlu sangat akurat. Ketika algoritma adaboost membangun model "lemah", bobot data dengan kesalahan tinggi diperbesar. Maka data dengan kesalahan tinggi lebih efektif. Dengan algoritma iterasi berulang dan hasil akhirnya optimal dan akurat.

Dalam algoritma *adaboost*, setiap set sampel *training* diperoleh dengan mengubah bobot distribusi data sampel *training*. Pada awal algoritma, setiap set sampel *training* ditugaskan dengan bobot yang sama. Pada akhir langkah *training*, bobot set sampel *training* dengan kesalahan diatas ambang besar tetap akan ditingkatkan dan bobot set sampel *training* yang kesalahannya kurang dari ambang kesalahan tetap akan berkurang. Dengan menyesuaikan semua bobot set sampel, total data sampel *training* baru diperoleh. Dengan cara ini, sampel yang diukur sebagai kesalahan tinggi menunjukkan lebih penting dalam proses pelatihan model "*weak*" di setiap waktu. Pada akhir setiap loop, model "*weak*" yang relevan dengan data sampel dengan kesalahan tinggi diperoleh. Menurut kesalahan model "*weak*", bobot model dihitung.

Penelitian oleh Yinlei Wen dkk [22] metode analisis baru dikembangkan berdasarkan algoritma adaboost. Dengan menggunakan *neural network* dengan struktur tetap, serangkaian model dibangun yang mungkin tidak akurat. Tingkat kesalahan model dihitung untuk mendapatkan dan menyesuaikan bobot setiap model. Model akurat yang lebih tinggi dibangun oleh model dan bobot. Dibandingkan dengan metode *neural network* tradisional, metode berbasis adaboost ini tidak perlu menyesuaikan jumlah simpul *neural network*. Selain itu, tetap akurat dan mengurangi kompleksitas.

IV. CONCLUSION

Dalam survei literatur ini, *big data* dan berbagai konsepnya meliputi arsitektur sistem *big data*, karakteristik *big data*, teknik analisis *big data*, dan algoritma analisis *big data* telah dipelajari. *Big data* memiliki kemampuan untuk mempengaruhi berbagai bidang teknologi informasi, dari ilmu sosial ke ilmu politik, dari industri keuangan ke bisnis, dari ilmu kedokteran ke kesehatan masyarakat, dari perawatan kesehatan ke genetika, dan dari obat-obatan yang dipersonalisasi ke pasien / kebiasaan. *Data science* menawarkan peluang baru dalam komputasi. Tetapi sementara itu, *analytics model pada big data* hadir dengan banyak tantangan yang memungkinkan peningkatan pemahaman tentang pemanfataan *big data*. Alat analitik yang berkembang saat ini dan pemanfaatan sumber daya terbuka secara cerdas adalah kunci untuk meningkatkan nilai sebenarnya dari *big data* secara *real time* untuk pengambilan keputusan yang dapat ditindaklanjuti dan hasil yang lebih akurat.

REFERENCES

[1] M. Minelli, M. Chambers, and A. Dhiraj, *Big Data, Big Analytics : Emerging Business Intelligence And Analytic Trends For Today's Businesses*. Hoboken, New Jersey: Wiley & Sons, Inc., 2013.

[2] H. Chen, R. Chiang, and V. Storey, "Business intelligence and analytics: From big data to big impact," *MIS Q.*, vol. 36, no. 4, 2012.

[3] H.-M. Chen, R. Kazman, and S. Haziyev, "Agile Big Data Analytics for Web-Based Systems: An Architecture-Centric Approach," *IEEE Trans. Big Data*, vol. 2, no. 3, pp. 234–248, 2016.

[4] B. Santosa and A. Umam, *Data Mining dan Big Data Analytics, Teori dan Implementasi Menggunakan Python dan Apache Spark*, Cetakan 1. Penebar Media Pustaka, 2018.

[5] S. Jiang, X. Qian, T. Mei, and Y. Fu, "Personalized Travel Sequence Recommendation on Multi-Source Big Social Media," *IEEE Trans. Big Data*, vol. 2, no. 1, pp. 43–56, 2016.

[6] J. Gantz and D. Reinsel, "Extracting Value from Chaos State of the Universe," 2011.

[7] S. Gaikwad, P. Nale, and R. Bachate, "Survey on Big data Analytics for digital world," in *2016 IEEE International Conference on Advances in Electronics, Communication and Computer Technology, ICAECCT 2016*, 2017, pp. 180–186.

[8] H. M. Chen, R. Kazman, S. Haziyev, and O. Hrytsay, “Big Data System Development: An Embedded Case Study with a Global Outsourcing Firm,” in *Proceedings - 1st International Workshop on Big Data Software Engineering, BIGDSE 2015*, 2015, pp. 44–50.

[9] M. G. Kibria, K. Nguyen, G. P. Villardi, O. Zhao, K. Ishizu, and F. Kojima, “Big Data Analytics, Machine Learning, and Artificial Intelligence in Next-Generation Wireless Networks,” *IEEE Access*, vol. 6, pp. 32328–32338, 2018.

[10] J. Manyika, B. Chui, M., J. B., Bughin, R. Dobbs, C. Roxburgh, and A. Hung Byers, “Big data: The next frontier for innovation, competition and productivity,” 2011.

[11] Han Hu, Yonggang Wen, Tat-Seng Chua, and Xuelong Li, “Toward Scalable Systems for Big Data Analytics: A Technology Tutorial,” *IEEE Access*, vol. 2, pp. 652–687, 2014.

[12] D. Laney, *3D Data Management: Controlling Data Volume, Velocity, and Variety, Gartner*, no. February 2001. CRC Press Taylor & Francis Group, 2014.

[13] C. K.-S. Leung, *Big Data Mining and Analytics*. 2014.

[14] H. Motoda *et al.*, *Top 10 algorithms in data mining*, vol. 14, no. 1. London: Springer, 2007.

[15] I. Triguero, D. García-Gil, J. Maillo, J. Luengo, S. García, and F. Herrera, “Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data,” in *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2019, vol. 9, no. 2, pp. 1–24.

[16] L. Han, W. Li, and Z. Su, “An assertive reasoning method for emergency response management based on knowledge elements C4.5 decision tree,” in *Expert Systems with Applications*, 2019, vol. 122, pp. 65–74.

[17] P. Shan, “Image segmentation method based on K-mean algorithm,” *Eurasip J. Image Video Process.*, vol. 2018, no. 1, 2018.

[18] M. B. S. Kumar and K. Balakrishnan, “Development of a Model Recommender System for Agriculture Using Apriori Algorithm,” in *Cognitive Informatics and Soft Computing*, 2019, vol. 768.

[19] N. Sammaknejad, Y. Zhao, and B. Huang, “A review of the Expectation Maximization algorithm in data-driven process identification,” *J. Process Control*, vol. 73, pp. 123–136, 2019.

[20] E. Kim and M. Jeon, “Proposal for implementation of a ranking model for Olympic Taekwondo competitions using PageRank,” *Int. J. Perform. Anal. Sport*, vol. 19, no. 2, pp. 227–235, 2019.

[21] K. J. D’souza and Z. Ansari, “Big Data Science in Building Medical Data Classifier Using Naïve Bayes Model,” pp. 76–80, 2019.

[22] Y. Ji, P. Xu, and J. Y. Chen, “An hourly electricity consumption calculation method for hvac terminal units with classification and regression tree on the basis of sub-metering,” *IOP Conf. Ser. Earth Environ. Sci.*, vol. 238, p. 012002, 2019.