ITEJ Juli-2019, Volume 4 Nomor 1 Page 1 - 13

ITEJ
Information Technology Engineering Journals
eISSN : 2548-2157

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

A Survey on Big Data Analytics Model

Candra Kurniawan
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung
candrak@student.itb.ac.id

Abstract—Topic about big data analytics have received a lot of attention and interest at this time. There are many topics can be discussed related to the analytical model, tools, and technology used. Big data analytics model involves many processes with various technologies used. Skills in handling big data, extracting mining, and developing insight are needed in applying big data analytics. Suitable analytical hardware and software also needed in decision making. Big data analytics is a key to a business strategy, but only a small portion of big data is currently used to support their business strategy. Big data analitycs can answer many questions about how to manage costs, time, and development or optimization strategies, and other decision making choices. However, there are many challenges in big data analytics technology. This survey paper addresses topics related to the analytical model, tools, and technology used. This paper also discusses the application of big data analytics in various fields.

Keywords—big data, big data analytics model, applications of big data analytics

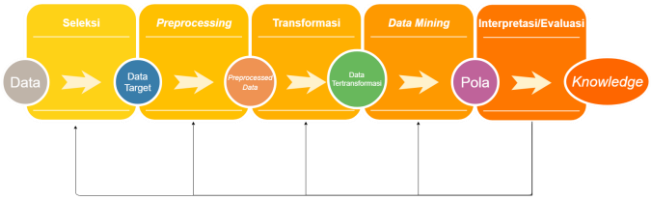
I. PENDAHULUAN

Data didefinisikan sebagai representasi terstruktur dari entitas tunggal [1]. Saat ini, baik sistem maupun penggunanya telah menggunakan generasi data dengan ukuran yang besar dan akan terus tumbuh secara eksponensial. Pertumbuhan ini dapat dilihat dari ukuran data yang digunakan, dimana ukuran paling kecil data yang pernah digunakan adalah bit (*binary digit*) dan kini telah mencapai *petabytes* (PB), dan diperkirakan akan terus tumbuh melampaui ukuran yang saat ini bisa diukur oleh manusia. Pertumbuhan data yang kuat ini menyebabkan lahirnya istilah *big data*.

Big data memang identik dengan istilah yang menggambarkan volume data yang besar. Lebih dari itu, *big data* tidak hanya merujuk pada volume datanya saja tetapi juga fitur lain yang dimiliki *big data* membuat istilah ini berbeda dengan konsep data masif dan data yang sangat besar [2]. Perbedaannya dapat dilihat berdasarkan tiga tipe definisi, yaitu: 1) definisi atribut yang menjelaskan dimensi dari *big data* (dijelaskan lebih lanjut pada BAB II); 2) definisi komparatif dimana *big data* didefinisikan sebagai data yang ukurannya di luar kemampuan perangkat lunak basis data biasa dalam menangkap, menyimpan, mengelola, dan menganalisis; dan 3) definisi arsitektur yang menyatakan bahwa *big data* adalah tempat dimana volume data, kecepatan akuisisi, atau representasi data membatasi kemampuan untuk melakukan analisis yang efektif menggunakan pendekatan tradisional.

Tentu saja, baik data maupun *big data* tidak akan menghasilkan informasi ataupun pengetahuan apabila tidak diproses. Vercellis [1] menyatakan bahwa informasi adalah hasil dari kegiatan ekstraksi dan pemrosesan data, dan informasi ditransformasikan menjadi pengetahuan/wawasan ketika

digunakan untuk membuat keputusan dan mengembangkan tindakan yang sesuai. *Knowledge Discovery in Databases* (KDD) atau penemuan pengetahuan/wawasan dalam basis data adalah proses untuk mengidentifikasi kumpulan data sehingga dapat dipahami [3]. Untuk mendapatkan penemuan pengetahuan/wawasan tersebut dibutuhkan proses dengan urutan berulang dan interaktif dari langkah-langkah utamanya, digambarkan pada Gambar 1 berikut.



Gambar 1 Proses dalam KDD

Big data analytics (BDA) dalam terminologi Layman adalah analisis data yang melibatkan data dalam jumlah besar (sampai triliunan baris) atau analisis masalah yang sulit dipecahkan.Tujuan utama dari *big data analytics* adalah untuk mendapatkan *big value*/informasi yang berharga [4], sehingga dapat dimanfaatkan dalam berbagai bidang yang mengarah pada pengambilan keputusan dan strategi bisnis yang lebih baik. BDA sebagai proses untuk menemukan dan mengelola informasi, pola, dan kesimpulan yang berguna dari *big data* [5] memiliki dampak penting pada proses pengambilan keputusan dengan menerapkan metode ilmiah untuk memecahkan masalah yang sebelumnya sulit untuk dipecahkan [6]. Analisis statistik tradisional berbasis hipotesis pada awalnya digunakan sebagai basis utama dalam BDA, namun kini BDA telah melibatkan pembelajaran mesin, pemodelan prediktif, alat pemrosesan yang lebih cepat, lingkungan analitik berkinerja tinggi, dan analitik visual [7].

Sayangnya tidak semua implementasi BDA berhasil, keterampilan dalam menangani *big data*, mengekstrak makna, dan mengembangkan wawasan sangat berpengaruh. Untuk mendapatkan wawasan yang dibutuhkan, maka implementasi BDA yang tepat, keterampilan analitik yang kompeten, dan kesesuaian perangkat keras atau perangkat lunak analitik yang dipakai, atau bisa disebut kecermatan dalam memilih fitur, semuanya diperlukan dalam pengambilan keputusan. Pemilihan fitur telah menunjukkan keefektifannya dalam banyak penerapan tetapi karakteristik unik dari *big data* membutuhkan keahlian dan tantangan tersendiri [8]. Sampai saat ini, hanya sebagian kecil saja *big data* dari jutaan bahkan triliunan sumber yang dimanfaatkan dengan melakukan analisis, padahal dengan BDA dalam sebuah bisnis dapat menjadi strategi yang baik dan menjadikan keuntungan yang terus mengalir. BDA dapat menemukan jawaban yang diinginkan, seperti bagaimana mengatur biaya, bagaimana mengatur waktu, strategi dalam mengembangkan produk baru, strategi pengoptimalan penawaran produk, dan pilihan pengambilan keputusan lainnya.

Sebagai hasilnya, survei ini bertujuan untuk memberikan tinjauan tentang model analitik pada *big data* yang berfokus pada penerapannya dalam berbagai bidang. Penulisan dalam survei literatur ini dibagi menjadi beberapa bab. Bab 2 membahas gambaran umum data analitik dan *big data*. Bab 3 melengkapi bagian sebelumnya dan membahas penerapan model analitik pada *big data* di berbagai bidang, lengkap dengan masalah dan peluangnya. Catatan kesimpulan diberikan di bagian 4 untuk merangkum hasil dan peningkatan di masa mendatang.

II. GAMBARAN UMUM MODEL ANALITIK DAN *BIG DATA*

A. Model Analitik

Dalam terminologi Layman, data analitik dideskripsikan sebagai kegiatan analisis data, baik besar atau kecil, untuk memahami dan melihat bagaimana menggunakan pengetahuan yang tersembunyi di dalamnya. Apabila dikaji berdasarkan modelnya dan diurutkan berdasarkan tingkat kesulitannya dari yang termudah sampai yang tersulit, terdapat empat jenis model analitik[9], yaitu :

1. Deskriptif

Model ini menjawab pertanyaan “apa yang terjadi”, analisis deskriptif mendeskripsikan dan merangkum fitur-fitur dasar yang telah tersirat pada seluruh populasi dari set data yang diberikan.

Ringkasan disajikan dalam bentuk grafik untuk memfasilitasi proses pengambilan keputusan. Set data yang diberikan juga disajikan sesuai kecenderungan utama, variabel, dan penyebaran data.

2. Diagnostik

Model diagnostik menjawab pertanyaan "mengapa terjadi", memungkinkan untuk menemukan akar permasalahan utama dari suatu kejadian. Kunci utama dalam model ini adalah mengidentifikasi anomali yang ada.

3. Prediktif

Model analitik prediktif menjawab pertanyaan "apa yang mungkin terjadi", analisis ini dapat menghasilkan apa saja yang bisa memengaruhi tren dan pola jika suatu data berubah. Lebih jauh lagi arah prediksi bukan hanya sekedar prediksi di masa depan, namun juga dapat diterapkan pada peristiwa yang tidak diketahui di masa lalu, dan masa sekarang.

4. Preskriptif

Menjawab pertanyaan "apa yang harus dilakukan selanjutnya", merupakan kombinasi analitik deskriptif dan prediktif.

Data mining adalah salah satu aktivitas dalam data analitik. *Data mining* juga sering disebut *Knowledge Discovery in Databases* (KDD) yang sekilas dibahas dalam Bab 1. Langkah yang harus dilalui sehingga data dapat menghasilkan pengetahuan adalah seleksi data, *preprocessing data*, transformasi data, *data mining*, dan interpretasi/evaluasi. Dalam seleksi data, data yang relevan dengan analisis diambil dari basis data. Proses ini dilakukan dengan memilih subset dari basis data atau sampel data dimana penemuan harus dilakukan. *Preprocessing data*, dalam kegiatan ini pembersihan data dan integrase data dilakukan. Pembersihan data dilakukan dengan menghilangkan data yang tidak konsisten dan data yang memiliki *noise*. Integrasi data adalah proses menggabungkan data yang berasal dari sumber yang berbeda. Transformasi data, pada langkah ini dilakukan fungsi ringkasan atau agregat, data ditransformasikan secara koheren untuk *mining* yang sesuai. Proses transformasi juga berarti mengurangi dan memproyeksikan data untuk memperoleh representasi yang sesuai dengan tugas spesifik yang akan dilakukan. *Data mining*, pada langkah ini, beberapa metode misalnya peringkasan, klasifikasi, pengelompokan, regresi, dan algoritma yang tepat digunakan untuk mengekstraksi pola yang sesuai dan mewakili hasil yang diharapkan. Langkah terakhir adalah interpretasi/evaluasi. Evaluasi dilakukan oleh pengguna untuk mengidentifikasi dan mengekstrak pengetahuan dari data yang diproses sehingga pada akhirnya menghasilkan pengetahuan/wawasan.

Ada tujuh teknik dasar dalam *data mining*, yaitu: karakterisasi dan diskriminasi, klasifikasi, regresi, analisis deret waktu, aturan asosiasi, pengelompokan/kluster, dan deskripsi dan visualisasi [1].

1. Karakterisasi dan diskriminasi adalah teknik yang menggeneralisasi, merangkum, dan mengkontraskan berdasarkan karakteristik datanya.
2. Klasifikasi adalah teknik yang paling umum digunakan yang berisi satu set sampel pra-klasifikasi untuk membuat model yang dapat mengklasifikasikan kumpulan data yang besar. Teknik ini membantu dalam memperoleh informasi penting tentang data dan metadatanya. Teknik ini terkait erat dengan teknik analisis kluster dan menggunakan pohon keputusan atau sistem jaringan saraf.
3. Regresi digunakan untuk membuat prediksi berdasarkan hubungan dalam kumpulan data.
4. Analisis deret waktu meliputi analisis regresi, penggalian pola sekuensial, analisis periodisitas, dan analisis berbasis kesamaan.
5. Aturan asosiasi membantu menemukan hubungan antara dua atau lebih item. Teknik ini membantu untuk mengetahui hubungan antara berbagai variabel dalam basis data.
6. Pengelompokan/kluster atau juga disebut segmentasi adalah salah satu teknik tertua yang digunakan dalam *data mining*, dimana proses identifikasi dilakukan dengan mengelompokkan data yang mirip satu sama lain untuk memahami perbedaan dan persamaan di antara data.
7. Teknik terakhir adalah deskripsi dan visualisasi, merupakan teknik yang paling berguna yang digunakan untuk menemukan pola dan mengubah data yang dianggap miskin menjadi data yang baik. Teknik ini digunakan pada awal proses *data mining*.

Algoritma yang digunakan dalam *data mining* bervariasi, disesuaikan dengan tujuan dan variabel yang tersedia. Terdapat 10 algoritma terpopuler dalam *data mining* [10], yaitu C4.5 dan turunannya, *K-means*, *Support Vector Machines* (SVM), Apriori, *Expectation-Maximization* (EM), *PageRank*, *AdaBoost*, *k-Nearest Neighbor* (kNN), *Naive Bayes*, *Classification and Regression Tree* (CART).

1. C4.5 dan turunannya, merupakan teknik pengklasifikasian yang merupakan teknik yang umum digunakan dalam data mining. Masukan berupa kumpulan kasus dijadikan sebagai dasar dimana masing-masing kasus mewakili sejumlah kecil kelas dan dideskripsikan dengan nilai-nilainya untuk

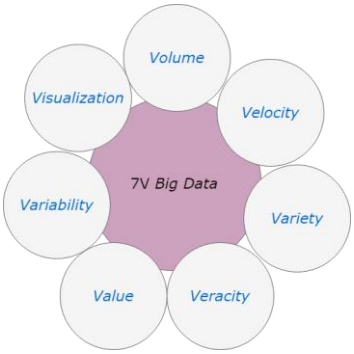
sekumpulan atribut yang tetap, sehingga dapat menghasilkan sebuah *classifier* yang dapat secara akurat memprediksi kelas di mana sebuah kasus baru berada.

2. *K-means* adalah metode iteratif sederhana untuk mempartisi set data yang diberikan ke sejumlah kluster yang ditentukan pengguna. Algoritma ini telah ditemukan oleh beberapa peneliti dari berbagai disiplin ilmu.
3. *Support Vector Machines* (SVM) menawarkan salah satu metode yang paling kuat dan akurat di antara beberapa algoritma yang ada. SVM memiliki landasan teori yang kuat, *data training* yang kecil, dan tidak peka terhadap jumlah dimensi. Selain itu, metode yang efisien untuk SVM juga terus dikembangkan.
4. Apriori adalah satu pendekatan *data mining* dengan menemukan itemet yang sering muncul dari set data transaksi, sehingga aturan asosiasi dapat dijalankan.
5. *Expectation-Maximization* (EM) merupakan pendekatan yang fleksibel dan berbasis matematika untuk pemodelan dan pengelompokan data yang diamati pada fenomena acak.
6. *PageRank* adalah algoritma pencarian berbasis peringkat menggunakan *hyperlink* di *Web*. Google adalah contoh perusahaan yang telah sukses besar menggunakan algoritma ini. Saat ini, setiap mesin pencarian memiliki metode peringkat berdasarkan tautannya sendiri.
7. *AdaBoost* memanfaatkan *ensemble learning*, dimana metode ini yang menggunakan banyak *data training* untuk memecahkan masalah. Kemampuan generalisasi sebuah *ensemble* biasanya jauh lebih baik daripada yang konvensional saja, sehingga metode ensemble sangat menarik. Algoritma *AdaBoost* adalah salah satu metode *ensemble* yang memiliki fondasi teoretis yang kuat, prediksi yang sangat akurat, sederhanaan, dan pengaplikasian yang luas dan sukses.
8. *k-Nearest Neighbor* (kNN) adalah 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 kNN.
9. *Naive Bayes* merupakan pengklasifikasian dengan metode probabilitas dan statistik. Metode ini memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai teorema Bayes. Teorema tersebut dikombinasikan dengan *naive* dimana diasumsikan kondisi antar atribut saling bebas.
10. *Classification and Regression Tree* (CART) adalah metode pohon regresi dan pohon klasifikasi. Jika variabel dependen yang dimiliki bertipe kategorik maka CART menghasilkan pohon klasifikasi (*classification trees*), sedangkan jika variabel dependen yang dimiliki bertipe kontinu atau numerik maka CART menghasilkan pohon regresi (*regression trees*).

B. Big Data

Penelitian mengenai *big data* masih menarik untuk terus dikaji selama dua dasawarsa terakhir. *Big data* bukan hanya merepresentasikan tindakan pengumpulan dan penyimpanan data dengan ciri ukuran besar saja, tetapi juga dicirikan oleh kompleksitas struktural dan makna, tingkat produksi yang tinggi, dan keanekaragaman sifat dari sumber[11].

Doug Laney, pada tahun 2001 menyampaikan konsep dimensi *big data* yang menjadi karakteristik penting dari *big data*. Karakteristik yang digagas mencirikan tantangan dan peluang yang ada saat itu dan masih terus berlanjut di era sekarang. Doug Laney membagi dimensi *big data* menjadi 3 yaitu *volume*, *velocity*, dan *variety*[12] yang dikenal sebagai 3V dari *big data*. Dimensi-dimensi *big data* yang lain (V tambahan) terus ditambahkan seiring berjalannya waktu, namun masih tetap berpedoman pada 3V. *Veracity* dan *value*[13] ditambahkan oleh Russom menjadikan 5V dari *big data*. Kemudian *variability* dan *visualization* ditambahkan dan melengkapi dimensi mejadi 7V dari *big data* seperti terlihat pada Gambar 2. *Volume* menggambarkan ukuran data, *variety* menggambarkan keragaman jenis data, *velocity* menggambarkan laju pertumbuhan maupun perubahan data, *veracity* merujuk pada sejauh mana data dapat dipahami (*biases, noise, abnormality*), *value* berhubungan dengan manfaat data dalam membuat keputusan, *variability* menunjukkan variasi data dalam varietas tertentu, dan *visualization* berhubungan dengan cara untuk merepresentasikan data.



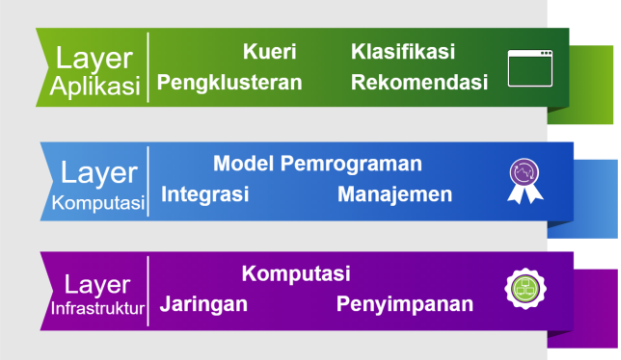
Gambar 2 Dimensi *Big Data*

Big data dapat dikelompokkan menjadi lima kategori berdasarkan sumbernya, yaitu *web* dan media sosial, mesin ke mesin, data transaksi besar, biometrik, dan data yang dihasilkan manusia [14].

1. Data yag berasal dari *web* dan media sosial adalah sumber utama yang menghasilkan *big data* saat ini. Data yang dimaksud berupa foto, video, data yang diunggah ke media sosial, pencarian konten *web* melalui mesin pencari, unggahan konten web, dan konten lainnya.
2. Data dari mesin ke mesin bervariasi dari berbagai sumber seperti pembacaan peralatan *smart meter*, pembacaan RFID di stasiun, kantor, bus atau alat transportasi lainnya, pembacaan sensor minyak, dan sinyal GPS dalam berbagai perangkat.
3. Data transaksi besar adalah kategori data yang dihasilkan dari perangkat *Healthcare*, catatan data panggilan dari berbagai operator telekomunikasi, catatan tagihan ponsel, saluran telepon, listrik, dan kartu prabayar kereta dan gerbang tol.
4. Data biometrik berasal dari perangkat biometrik untuk menyimpan data masuk dan keluar karyawan, kehadiran pegawai, sistem absensi siswa, dan sistem pengenalan wajah. Data buatan manusia contohnya adalah dari rekaman suara *call center*, surat elektronik, dan catatan medis elektronik.

Layer arsitektur dari sistem *big data* dapat didekomposisi menjadi tiga lapisan seperti terlihat pada Gambar 3, yaitu lapisan infrastruktur, lapisan komputasi, dan lapisan aplikasi [2].

1. Lapisan infrastruktur terdiri dari kumpulan sumber daya teknologi informasi yang dapat diatur oleh infrastruktur *cloud computing* dan didukung dengan teknologi virtualisasi. Sumber daya ini terhubung dengan lapisan atasnya dengan *service level agreement* (SLA) dalam memenuhi permintaan *big data*.
2. Lapisan komputasi merangkum berbagai alat yang digunakan dalam pemrosesan data ke dalam lapisan *middleware* yang menggunakan sumber daya teknologi informasi. Dalam konteks *big data*, alat pemrosesan data adalah alat yang digunakan dalam integrasi data, manajemen data, dan model pemrograman. Integrasi data berarti memperoleh data dari sumber yang berbeda dan mengintegrasikan set data ke dalam suatu formula yang sama dengan operasi pra-pemrosesan data yang diperlukan. Manajemen data mengacu pada mekanisme dan alat yang menyediakan penyimpanan data, seperti sistem file terdistribusi dan penyimpanan data SQL atau NoSQL. Model pemrograman mengimplementasikan logika aplikasi abstraksi dan memfasilitasi aplikasi analisis data, seperti *MapReduce*, *Dryad*, *Pregel*, dan *Dremel*.
3. Lapisan aplikasi mengeksploitasi antarmuka yang disediakan oleh model pemrograman untuk mengimplementasikan berbagai fungsi analisis data, termasuk kueri, analisis statistik, pengelompokan, dan klasifikasi yang kemudian digabungkan dengan metode analitik dasar untuk berbagai tujuan.



Gambar 3 Lapisan Arsitektur dari Sistem *Big Data*

C. Model Analitik pada Big Data

Big data analytics (BDA) seperti yang telah disebutkan dalam terminologi Layman pada bahasan sebelumnya adalah analisis data yang melibatkan data dalam jumlah besar atau analisis masalah yang sulit dipecahkan. Singkatnya pemrosesan data terdiri dari pengumpulan, pemrosesan, dan pengelolaan data untuk menghasilkan informasi baru bagi pengguna. Di dalam analisis *big data* dikenal 4A, dimana manajemen *big data* dibagi menjadi empat, yaitu: *access* (akses), *assemble* (menghimpun), *analyze* (menganalisis), dan *act* (aksi) [15].

4. Akses data atau sering juga disebut akuisisi data. Arsitektur big data harus memperoleh data dengan kecepatan tinggi dari berbagai sumber data dengan banyak protokol kontrol akses yang berbeda. Filter data yang dapat memilah data dengan hanya menyimpan data tertentu yang sesuai sangat penting dalam bagian ini.
5. Menghimpun data. Berbagai format data harus dapat diurai dan diekstrak informasinya pada langkah ini. Termasuk di dalamnya pembersihan data, penempatan ke dalam mode yang dapat dihitung, dan penyimpanan pada lokasi yang tepat, atau biasa dimisalkan seperti mekanisme *extract*, *transform*, dan *load*.
6. Analisis data. Dalam bagian ini kueri dijalankan, pemodelan dilakukan, dan algoritma dibangun untuk menemukan wawasan baru (proses KDD). Proses ini membutuhkan data yang sudah dibersihkan terlebih dahulu, terintegrasi, dan dapat dipercaya.
7. Aksi atau tindakan. Aktivitas terakhir adalah bagaimana suatu tindakan menghasilkan keputusan yang berharga berdasarkan hasil analisis pada langkah sebelumnya. Pengambilan keputusan merupakan hal yang sangat penting dalam bisnis karena menentukan keberlanjutan suatu organisasi selanjutnya.

Terdapat banyak *tools* yang bisa digunakan dalam *big data analytics*. *Tools* digunakan untuk manajemen *big data* mulai dari akuisisi data, penyimpanan data, hingga visualisasi data. Berikut ini adalah beberapa tools yang sering digunakan:

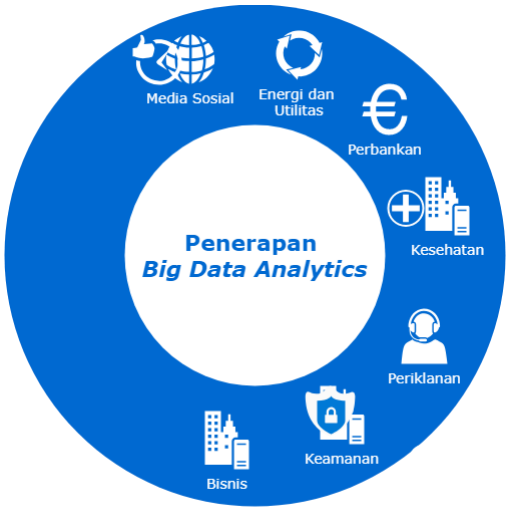
1. Apache Hadoop. Hadoop adalah *open source framework*, berbasis Java, yang mendukung pemrosesan set data besar dalam lingkungan terdistribusi. Kerangka dasar hadoop berisi empat modul utama, yaitu *Hadoop Distributed File System* (HDFS), YARN, MapReduce, dan Hadoop Common [16]. HDFS adalah sistem file terdistribusi yang menyediakan akses *throughput* tinggi ke data aplikasi. YARN adalah kerangka kerja untuk penjadwalan pekerjaan dan manajemen sumber daya. MapReduce adalah sistem berbasis YARN untuk pemrosesan parallel dari set data yang besar. Hadoop Common adalah utilitas umum yang mendukung modul Hadoop lainnya. Proyek terkait Hadoop lainnya yang terdapat di Apache adalah Ambari, Avro, Cassandra, Hbase, Hive, Mahout, Zookeeper, dan lainnya.
2. Map reduce. MapReduce adalah model pemrograman, diusulkan oleh Google, untuk komputasi paralel dalam memproses dan menganalisis set data yang besar. Proses utamanya sebelum menghasilkan keluaran adalah pemetaan, pengacakan, dan pengurangan.
3. Dryad. Dryad adalah model pemrograman yang juga digunakan untuk pemrosesan paralel set data besar dalam lingkungan terdistribusi. Ini terdiri dari sekelompok node komputasi, dan pengguna menggunakan sumber daya dari cluster komputer untuk menjalankan program mereka secara terdistribusi.

4. Apache Spark. Spark juga merupakan *open source framework* yang dibangun untuk pemrosesan yang cepat dan analitik yang canggih. Spark dirancang untuk mengatasi keterbatasan Hadoop MapReduce dan meningkatkan kinerja pada analisis *big data*. Fitur utama Spark yang membuatnya unik adalah kemampuannya untuk melakukan penghitungan memori. Spark memungkinkan penulisan aplikasi dengan cepat di java, scala, atau python. Selain itu, kueri SQL, *streaming* data, pembelajaran mesin, dan pemrosesan data grafik juga didukung dalam Spark.
5. Rapidminer. Rapidminer adalah *platform* perangkat lunak yang menyediakan integrase untuk pembelajaran mesin, *data mining*, *text mining*, analisis prediktif, dan analitik bisnis. Sering dimanfaatkan untuk aplikasi bisnis dan komersial juga untuk penelitian, pendidikan, pelatihan, *prototyping* cepat, dan pengembangan aplikasi. Integrasi yang ditawarkan mendukung semua prosedur proses *data mining* termasuk persiapan set data, validasi, visualisasi dan optimasi hasil.
6. Storm. Storm adalah sistem terdistribusi dan *fault tolerant* untuk memproses data *streaming* yang besar. *Fault tolerant* didefinisikan sebagai suatu fitur yang memungkinkan suatu sistem tetap berjalan normal meskipun ada kerusakan pada salah satu komponen di dalamnya. Berbeda dengan Hadoop yang dirancang untuk pemrosesan *batch*, storm dirancang khusus untuk pemrosesan *real-time*. Kelebihan yang lain adalah mudah untuk dioperasikan dan memberikan kinerja yang kompetitif karena *fault tolerant*.

Dilihat dari ukuran data yang sangat besar dan akan terus tumbuh menjadi lebih besar lagi, maka dibutuhkan *tools* yang menyediakan teknik penyimpanan yang efisien dan efektif. *Tools* tersebut antara lain Hbase, Skytree, dan NoSQL. HBase adalah produk dari Apache, *open source* yang menyediakan akses baca dan tulis ke basis data yang besar. HBase mampu menangani set data yang sangat besar dengan miliaran baris dan jutaan kolom dan dengan mudah menggabungkan sumber data yang menggunakan berbagai macam struktur dan skema yang berbeda. SkyTree adalah *platform* data analitik dan pembelajaran mesin berkinerja tinggi yang berfokus pada penanganan *big data*. NoSQL (*Non-Relational Databases*) sering juga disebut *not only* SQL atau bukan sekedar SQL. NoSQL adalah pendekatan administrasi data dan desain basis data dengan jumlah besar pada lingkungan terdistribusi. Basis data NoSQL yang paling populer dibangun menggunakan Apache Cassandra, basis data milik Facebook, dan kini sudah menjadi *open source*. Implementasi database NoSQL lainnya termasuk SimpleDB, Google BigTable, MemcacheDB, Cassandra, MongoDB dan Voldemort.

III. MODEL ANALITIK PADA *BIG DATA* DAN PENERAPANNYA

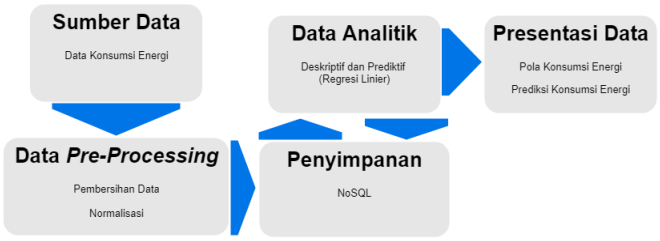
Penerapan b*ig data analytics* tlah terbukti banyak membantu di berbagai bidang. Survei ini mengumpulkan dan menganalisis beberapa makalah yang berhubungan dengan penerapan *big data analytics* dalam berbagai bidang dengan berbagai metodologi dan *tools* digunakan. Survei ini mengkategorikan setiap makalah yang dikaji ke dalam bidang-bidang spesifiknya sesuai dengan Gambar 4.



Gambar 4 :Bidang Penerapan *Big Data Analytics*

1. *Big data analytics* dalam bidang energi dan utilitas.

Peluang untuk implementasi model BDA yang berfokus pada sisi konsumen dari sektor energi cukup banyak. Dalam penelitian [9], pengembangan model BDA dilakukan untuk melacak dan memantau konsumsi listrik rumah tangga. Model analitik yang digunakan adalah analitik deskriptif dan preskriptif. Model ini bertujuan untuk memberikan hasil analisis yang dapat membantu pengguna untuk mengelola konsumsi listrik rumah tangga dengan lebih baik. Algoritma regresi digunakan penulis sebagai *data mining*. Regresi dipilih karena keluaran yang diperlukan untuk prediksi konsumsi listrik dalam model dalam format numerik kontinu. Selain itu, regresi juga dianggap sebagai algoritma yang populer.

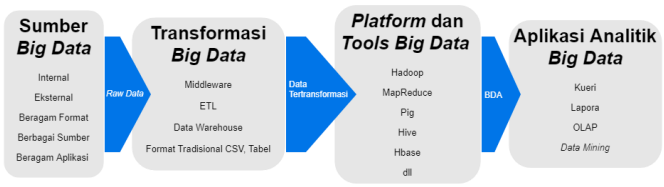


Gambar 5 Big Data Analytics Model untuk memonitor konsumsi listrik rumah tangga

Penulis membuat model, Gambar 5, yang mengadopsi KDD dengan penyesuaian di beberapa tahapan. *Tools* yang digunakan penulis adalah Hadoop dan NoSQL sebagai penyimpanannya. Berdasarkan evaluasi yang dilakukan, model BDA yang diusulkan untuk memonitor konsumsi listrik rumah tangga sudah dapat digunakan dan dapat memenuhi tujuannya, yaitu memberikan hasil analisis yang dapat membantu pengguna untuk mengelola konsumsi listrik rumah tangga dengan lebih baik. Peluang untuk pengembangan lebih lanjut di masa depan adalah mengembangkan model menjadi lebih luas cakupannya dengan memasukkan faktor-faktor eksternal seperti data cuaca.

2. *Big data analytics* dalam bidang kesehatan.

Big data analytics memiliki potensi mengubah cara penyedia layanan kesehatan dalam menggunakan teknologi untuk mendapatkan wawasan dari data klinis yang dimiliki dan membuat keputusan yang lebih baik berdasarkan informasi yang tersedia. Penemuan asosiasi dan pemahaman pola atau tren dalam data melalui BDA selain memiliki potensi untuk meningkatkan perawatan, juga dapat menyelamatkan banyak nyawa dan membuat biaya menjadi lebih rendah [17]. Penulis mengusung sebuah arsitektur, Gambar 6, yang merangkum kerangka konseptual BDA, mulai dari sumber data sampai dengan aplikasi BDA.



Gambar 6 Penerapan Arsitektur Konseptual *Big Data Analytics*

BDA dapat membantu mengurangi inefisiensi pada operasi klinis (diagnosis dan pengobatan pasien), penelitian & pengembangan (pemodelan prediktif), dan kesehatan masyarakat (kepentingan populasi) [18]. Selain itu, kontribusi BDA dalam perawatan kesehatan yang saat ini ada adalah sebagai berikut:

b. Pengobatan berbasis bukti. Yaitu menggabungkan dan menganalisis berbagai data terstruktur dan tidak terstruktur, data keuangan dan operasional, data klinis, dan data genom untuk mencocokkan perawatan dengan hasil, memprediksi pasien yang berisiko terhadap suatu penyakit, dan memberikan perawatan yang lebih efisien.

c. Analisis Genomik. Membuat proses lebih efisien dan hemat biaya. Analisis genomik dibutuhkan sebagai bagian dari proses pengambilan keputusan perawatan medis dan rekam medis pasien.

d. Analisis penipuan pra-ajudikasi: Secara cepat menganalisis sejumlah besar permintaan klaim untuk mengurangi penipuan, pemborosan, dan penyalahgunaan.

e. Pemantauan perangkat jarak jauh. Menangkap dan menganalisis data dalam volume besar dan bergerak cepat dari perangkat di rumah sakit, untuk pemantauan keamanan dan prediksi kejadian buruk.

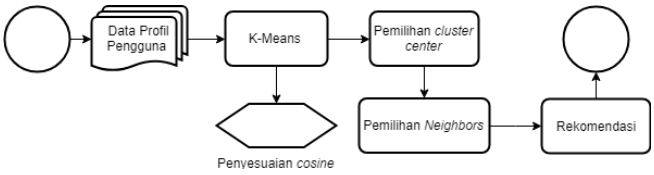
f. Analitik profil pasien: Menerapkan analitik lanjutan ke profil pasien untuk mengidentifikasi perawatan proaktif lanjutan/pencegahan dari perubahan gaya hidup pasien, risiko berdasarkan riwayat kesehatan, dan peluang terserang penyakit tertentu.

Memang sudah cepat dan luas penggunaan BDA di bidang kesehatan, namun beberapa tantangan ketika BDA menjadi lebih umum masih menjadi tantangan. Masalah seperti penjaminan privasi, keamanan, standar, tata kelola, dan peningkatan alat dan teknologi masih belum semua terselesaikan dan masih menarik perhatian.

3. *Big data analytics* dalam bidang periklanan.

Iklan merupakan salah satu faktor penting yang mempengaruhi tingkat penjualan suatu produk dalam suatu perusahaan. Dengan strategi periklanan yang tepat, maka akan meningkatkan penjualan produk, yang berimbas pada keuntungan perusahaan yang semakin besar. BDA mampu mengambil peran dalam upaya tersebut. Tantangan tersebut dapat diatasi dengan mengembangkan kerangka kerja BDA *mobile marketing* dan rekomendasi periklanan [19]. Pendekatan yang ditawarkan mendukung sistem rekomendasi iklan berbasis lokasi menggunakan teknologi terkini dan mampu menyajikan keputusan iklan yang relevan untuk pengguna.

Kerangka kerja yang ditawarkan mendukung operasi periklanan *offline* dan *online.* Teknik analisis yang dipilih digunakan untuk memberikan rekomendasi periklanan berdasarkan *big data* yang dikumpulkan pada profil pengguna seluler, perilaku akses, dan pola mobilitas. Sejumlah teknologi dan *tools* digunakan sebagai solusi untuk mendukung sistem rekomendasi yang dibangun, yaitu analisis *real-time* berdasarkan Spark, Informasi GEO yang terintegrasi dengan profil set data, algoritma K-Means (Gambar 7), dan teknik kluster.



Gambar 7 Prosedur K-Means yang Digunakan

Hasil pembahasan menunjukkan bahwa pendekatan yang dirumuskan masih membutuhkan peningkatan. Peningkatan komponen analisis data dari sistem, batasan atau ruang lingkup proyek, dan beberapa fitur masih bisa menjadi peluang di masa depan untuk menyempurnakan kerangka kerja yang digagas.

4. *Big data analytics* dalam bidang media sosial.

Aplikasi lain yang penting dan terus berkembang adalah media sosial. Analisis berdasarkan unggahan di situs-situs media sosial seperti Facebook dan Twitter memberikan kesempatan yang dapat digunakan untuk menarik kesimpulan dan membuat prediksi tentang kegiatan yang terjadi di area tertentu dan pada waktu tertentu [20]. Pemanfaatkan cuitan di twitter misalnya untuk mendapatkan *insight* kemacetan di suatu area beserta penyebabnya, juga informasi lakalantas, bahaya atau keadaan darurat, pengalihan jalan, dan peristiwa penting.

BDA yang diterapkan ke Twitter bisa dimanfaatkan untuk analisis sentimen berdasarkan pembaruan status [21]. Dalam prosesnya, t*ools* yang digunakan adalah Mahout dari Hadoop dan naïve bayes sebagai algoritmanya. Berdasarkan hasil yang didapatkan, model BDA pada Gambar 8 yang diusulkan dapat mengukur opini orang di berbagai topik yang berkaitan dengan berbagai bidang. Dengan menggunakan *hash tag,* penulis dapat menyediakan metode otomatis sederhana untuk mengevaluasi pendapat orang.

Gambar 8 Arsitektur Analisis Sentimen pada Twitter

Aplikasi pengiriman pesan dari Facebook, salah satu media sosial raksasa, menggunakan Hadoop dengan Hive untuk penyimpanan dan analisis [22]. *Platform* media sosial ini sangat informatif melalui sudut pandang *crowdsourcing*. *Insight* dari Facebook dapat memberi peluang pada pengembang dan pemilik situs web untuk analisis *real-time* terkait aktivitas pengguna Facebook di seluruh dunia dengan memanfaatkan *plugin*, halaman facebook itu sendiri, dan iklan dalam facebook. Analitik ini dapat membantu mendapatkan wawasan tentang bagaimana pengguna facebook berinteraksi dengan konten yang ada, sehingga para pemanfaat kepentingan dapat mengoptimalkan layanan yang dimiliknya, untuk promosi melalui iklan yang sesuai misalnya.

Sayangnya *real-time* BDA ini kebanyakan hanya menghasilkan *insight* yang berguna dalam waktu singkat karena pergeseran data/informasi terkait suatu topik analisis tertentu. Kecenderungan pengguna media sosial hanya menyampaikan satu hal yang disukai saja pada unggahan sosial media tanpa menyampaikan hal yang tidak disukai, berlaku juga sebaliknya, membuat informasi yang disajikan cepat kadaluwarsa. Misalnya, apabila terjadi kemacetan di suatu area, maka pengguna akan mengunggah status/cuitan tentang kemacetan, namun apabila kemacetan sudah berakhir, pengguna cenderung tidak menyampaikan hal tersebut pada media sosialnya. Hal inilah yang membuat *insight* yang dihasilkan hanya bertahan beberapa jam saja, sekaligus menjadi tantangan ke depannya bagaimana BDA dapat dilakukan dengan cepat dan menghasilkan *insight* yang akurat.

5. *Big data analytics* dalam bidang perbankan.

Bidang perbankan saat ini membutuhkan peningkatan dalam hal kemampuan menghasilkan data, bertukar data, pemrosesan, analisis, pelaporan, dan penyimpanan yang aman [23]. Peluang baru telah muncul untuk teknologi informasi keuangan tidak hanya untuk mengurangi biaya, tetapi juga untuk memperoleh kemampuan baru untuk penggunaan aplikasi analitik dan pembelajaran mesin pada skala yang belum pernah terjadi sebelumnya yang dikenal sebagai tren *big data*.

Analitik *big data* sekarang sedang diterapkan di bidang perbankan guna membantu penyedia layanan perbankan memberikan pelayanan yang lebih baik kepada pelanggannya, baik internal maupun eksternal. Investasi dapat ditingkatkan dengan menggunakan informasi agregat dari berbagai sumber seperti peramalan keuangan, penentuan harga aset, dan manajemen portofolio. Selain itu, manajemen data keuangan yang lebih baik dapat disarankan dengan menggunakan BDA.

6. *Big data analytics* dalam bidang keamanan.

Keamanan merupakan bagian penting dalam kehidupan manusia. Ancaman keamanan adalah hal yang menjadi perhatian utama untuk dicegah. Ancaman yang dimaksud adalah terorisme, meningkatnya tingkat kejahatan, kerusuhan sipil, penembakan, bencana alam, dan keadaan darurat lainnya [24]. *Big data analytics* dapat membantu mengatasi masalah terkait keamanan tersebut.

Big data analytics dapat diimplementasikan pada dunia pendidikan di universitas, dimana teknologi *big data* dapat digunakan untuk memeriksa/memantau peserta didik dengan kesimpulan analisis prediksi atau prediksi dilakukan terhadap peserta didik apakah mereka sudah terbiasa dengan ideologi ortodoks yang dapat menyebabkan masalah yang sangat sensitif seperti terorisme. Dalam lingkungan universitas, tersedia sejumlah besar data pribadi dan akademik. Berbagai alat tersedia untuk pemantauan dan analisis data. Ide dasarnya adalah analisis perilaku peserta didik yang bertujuan untuk mengamati apakah dia menyimpang dari perilaku normal yang menyebabkan kegiatan ilegal atau tidak sah. Pemantauan dan prediksi menyediakan sistem peringatan dini. Peringatan dini ini dapat didukung dengan dengan informasi, bimbingan, saran, motivasi dan umpan balik yang didukung untuk mencegah perilaku menyimpang yang berujung terorisme. Sumber *big data* untuk analisis ini bisa berasal dari basis data tradisional, data pribadi, jejak digital web, aktivitas luar ruangan, video pengawasan, sensor parker, dan lainnya. *Platform* Hadoop dapat dipakai menangani *big data* ini.

7. *Big data analytics* dalam bidang bisnis.

Analisis perilaku pelanggan adalah pasar yang besar, menjanjikan, dan belum dijelajahi secara menyeluruh. Potensi besar yang dijanjikan untuk mengetahui pelanggan mana yang merupakan pembeli paling berharga menjadi sengan penting karena membantu keberlangsungan bisnis ke depanya. *Big data analytics* memiliki kemampuan untuk membawa organisasi bisnis pada tingkat yang lebih tinggi dengan menganalisis perilaku pelanggan dan mengubahnya menjadi wawasan yang berharga. KDD menjadi kunci utama dari proses menghasilkan data menjadi wawasan.

Big data analytics menawarkan banyak peluang untuk meningkatkan nilai bisnis dan produktivitas dan salah satu aplikasi utamanya adalah untuk meningkatkan kemampuan pengambilan keputusan yang lebih cepat, memahami kebutuhan pelanggan, mengembangkan strategi untuk meluncurkan produk dan layanan baru, menjelajahi pasar baru, meningkatkan perputaran persediaan, mengurangi keluhan pelanggan, dan meningkatkan produktivitas dan efisiensi staf [25], segmentasi pelanggan yang tepat berdasarkan transaksi sebelumnya dan informasi profil dapat dilakukan. Analisis pola pembelian dan penawaran produk yang dibuat khusus, analisis data tidak terstruktur dari media sosial, multi-media untuk memahami selera, preferensi, dan pola pelanggan, serta melakukan analisis sentiment dari produk yang dihasilkan juga bisa dilakukan dengan *big data analytics*. Lebih jauh lagi pemasaran tertarget dan analisis pesaing bisa dihasilkan dari *big data analytics*.

Tools yang dapat umum dipakai untuk menganalisis data adalah pohon keputusan, karena dapat digunakan secara efisien. Visualisasi data pelanggan dari hasil analisis memegang peran penting karena analitik pelanggan tidak lengkap tanpa visualisasi data. *Tools* untuk visualisasi data adalah bisa berupa SAS, Polymaps, Visual Analytics, atau Flot.

IV. KESIMPULAN

Dalam tulisan ini, pembahasan mengenai *big data* dilakukan. Topik ini dinilai telah mendapatkan banyak perhatian dan minat saat ini. Terdapat banyak sekali bahasan dengan topik *big data* dan yang dibahas dalam tulisan ini adalah topik-topik yang berkaitan dengan model analitiknya, *tools*, dan teknologi yang digunakan. Tulisan ini juga membahas penerapan BDA dalam berbagai bidang.

Model analitik pada *big data* melibatkan banyak proses dengan berbagai teknologi yang digunakan. Keterampilan dalam menangani *big data*, mengekstrak makna, dan mengembangkan wawasan adalah kemampuan yang dibutuhkan oleh seseorang dalam menerapkan BDA. Tidak hanya SDMnya, untuk mendapatkan wawasan yang dibutuhkan kesesuaian perangkat keras atau perangkat lunak analitik yang dipakai juga diperlukan dalam pengambilan keputusan. Hanya sebagian kecil saja dari *big data* yang sampai saat ini dimanfaatkan, padahal BDA adalah kunci dari strategi bisnis yang akan menjadikan keuntungan besar bagi sebuah organisasi. BDA dapat menjadi jawaban bagaimana mengatur biaya, waktu, dan strategi pengembangan dan pengoptimalan, dan pilihan pengambilan keputusan lainnya.

Namun, ada banyak tantangan dalam teknologi *big data analytics* baik dalam hal pemrosesan, penyimpanan, maupun hasil penerapannya di berbagai bidang. Lebih jauh lagi, permasalahan yang dibahas dalam tulisan ini akan membutuhkan solusi transformatif untuk ditangani dan menjadi peluang untuk pengembangan selanjutnya. Oleh karena itu, untuk mencapai manfaat yang dijanjikan dari BDA, penelitian lanjutan masih terbuka lebar untuk pengaplikasian BDA di berbagai bidang dengan memanfaatkan *tools*, teknologi, atau model yang sudah ada, atau bahkan menciptakan model analitik yang baru pada *big data*.

REFERENSI

[1] C. Vercellis, Data mining and optomization for decision making, vol. 1. 2009.

[2] 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.

[3] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," AI Mag., vol. 17, pp. 37–54, 1996.

[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] M. Cao, R. Chychyla, and T. Stewart, "Big Data Analytics in Financial Statement Audits," Account. Horizons, vol. 29, p. 150219103526005, 2015.

[6] T. H. Davenport, Big Data at Work: Dispelling the Myths, Uncovering the Opportunities. USA: Harvard Business Review Press, 2014.

[7] H. Chen and V. C. Storey, "Business Intelligence and Analytics: From Big Data to Big Impact," vol. 36, no. 4, pp. 1165–1188, 2012.

[8] J. Li and H. Liu, "Challenges of Feature Selection for Big Data Analytics," IEEE Intell. Syst., vol. 32, no. 2, pp. 9–15, 2017.

[9] R. Dollah and H. Aris, "A Big Data Analytics Model for Household Electricity Consumption Tracking and Monitoring," 2018 IEEE Conf. Big Data Anal., pp. 44–49, 2019.

[10] X. Wu et al., Top 10 algorithms in data mining. 2008.

[11] M. Khosrow-pour et al., "Encyclopedia of Information Science and Technology, First Edition," Encycl. Inf. Sci. Technol. First Ed., vol. IX, 2011.

[12] D. Laney, "3-D Data Management: Controlling Data Volume, Velocity, and Variety," META Gr. Res Note 6, vol. 6, 2001.

[13] B. P. Russom, "Big Data Analytics," 2011.

[14] P. Radhika, P. P. Kumar, S. L. Sailaja, and V. Gayatri, "Confrontation and oppurtunities of big data - A survey," Proc. 2017 Int. Conf. Big Data Anal. Comput. Intell. ICBDACI 2017, vol. 6859, no. 6 2, pp. 153–157, 2017.

[15] Karmasphere, "Deriving Intelligence from Big Data in Hadoop: A Big Data Analytics Primer," 2011.

[16] M. Merrouchi, M. Skittou, and T. Gadi, "Popular platforms for big data analytics : A survey," 2018 Int. Conf. Electron. Control. Optim. Comput. Sci., pp. 1–6, 2018.

[17] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare : promise and potential," pp. 1–10, 2014.

[18] J. Manyika et al., "Big data: The next frontier for innovation, competition, and productivity." 2011.

[19] L. Deng, J. Gao, and C. Vuppalapati, "Building a Big Data Analytics Service Framework for Mobile Advertising and Marketing," 2015 IEEE First Int. Conf. Big Data Comput. Serv. Appl., pp. 256–266, 2015.

[20] A. Bifet, "Mining Big Data in Real Time," no. March 2013, 2018.

[21] M. Kumar, "Analyzing Twitter sentiments through big data," 2016 3rd Int. Conf. Comput. Sustain. Glob. Dev., pp. 2628–2631, 2016.

[22] D. Borthakur, S. Rash, N. Spiegelberg, R. Schmidt, and J. Gray, "Apache Hadoop goes realtime at Facebook Apache Hadoop Goes Realtime at Facebook," no. May, 2014.

[23] A. Munar, E. Chiner, and I. Sales, "A Big Data Financial Information Management Architecture for Global Banking," 2014.

[24] Setiyono and S. H. Supangkat, "Big Data Analytics for Safe and Secure City," Proceeding -

2018 Int. Conf. ICT Smart Soc. Innov. Towar. Smart Soc. Soc. 5.0, ICISS 2018, pp. 1–5, 2018.

[25] J. Ram, C. Zhang, and A. Koronios, “The implications of Big Data analytics on Business Intelligence : A qualitative study in China,” Procedia - Procedia Comput. Sci., vol. 87, pp. 221–226, 2016.