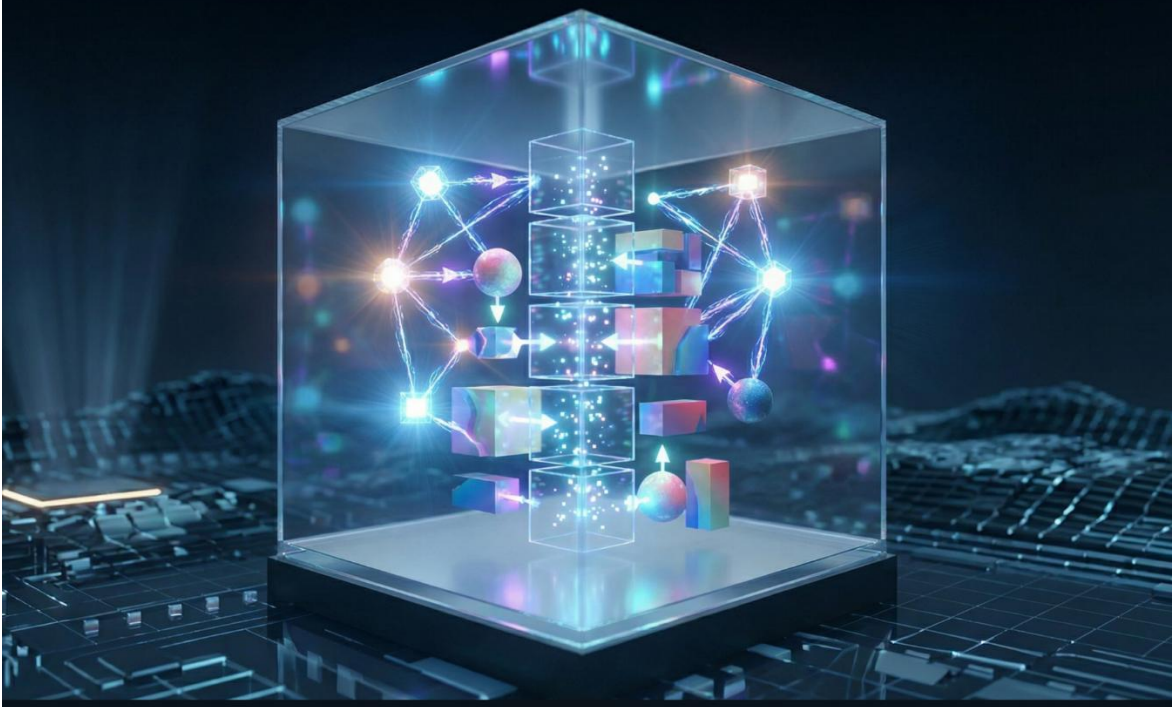
METODE KUANTITATIF ERA BIG DATA

Teori dan Implementasi



Tim Penulis:
Yuni Roza | Afiyati | Mohamad Yusuf | Firmansyah Apryadhi
Nila Natalia | Luthfia Fauzia Dewi Aryanti | Nungky Awang Chandra
Umniy Salamah | Fauzi Nur Iman | Rusdah | Lukman Hakim
Ida Farida | Dody

METODE KUANTITATIF ERA BIG DATA

Teori dan Implementasi

Yuni Roza
Afiyati
Mohamad Yusuf
Firmansyah Apryadhi
Nila Natalia
Luthfia Fauzia Dewi Aryanti
Nungky Awang Chandra
Umniy Salamah
Fauzi Nur Iman
Rusdah
Lukman Hakim
Ida Farida
Dody

Editor: Siti Maesaroh, S.Kom., M.TI.

sada
kurnia pustaka

METODE KUANTITATIF ERA BIG DATA

Teori dan Implementasi

Tim Penulis:
Yuni Roza
Afiyati
Mohamad Yusuf
Firmansyah Apryadhi
Nila Natalia
Luthfia Fauzia Dewi Aryanti
Nungky Awang Chandra
Umniy Salamah
Fauzi Nur Iman
Rusdah
Lukman Hakim
Ida Farida
Dody

Editor : Siti Maesaroh, S.Kom., M.TI.
Tata Letak : Lilis Khalisatul Karimah, S.H.
Desain Cover : Asep Nugraha, S.Hum.
Ukuran : UNESCO 15,5 x 23 cm
Halaman : viii, 224
ISBN : 978-634-7522-52-8
Terbit Pada : Mei 2026
Anggota IKAPI : No. 073/BANTEN/2023

PENERBIT PT SADA KURNIA PUSTAKA
Jl. Kramat, Panenjoan Kec. Carenang, Kab. Serang – Banten, 42195
Email : sadapenerbit@gmail.com
Website : sadapenerbit.com & repository.sadapenerbit.com
Telpon/WA : +62 838 1281 8431

KATA PENGANTAR

Puji syukur kami panjatkan ke hadirat Tuhan Yang Maha Esa atas segala rahmat dan hidayah-Nya, sehingga buku yang berjudul **"METODE KUANTITATIF ERA BIG DATA: Teori dan Implementasi"** ini dapat diselesaikan dengan baik. Buku ini hadir sebagai respons atas pesatnya perkembangan teknologi dan ledakan data (big data) yang mengubah paradigma analisis kuantitatif di berbagai bidang, baik ekonomi, sosial, kesehatan, maupun bisnis.

Perkembangan era big data tidak hanya membawa tantangan dalam hal volume, kecepatan, dan keragaman data, tetapi juga membuka peluang besar bagi peneliti dan praktisi untuk memperoleh wawasan yang lebih mendalam dan akurat. Metode kuantitatif konvensional perlu diperkaya dan diadaptasi dengan pendekatan komputasi modern, pembelajaran mesin (machine learning), serta teknik analisis data berskala besar. Buku ini disusun untuk menjembatani kesenjangan antara teori statistik klasik dan praktik analisis data di era digital.

Materi dalam buku ini mencakup konsep dasar metode kuantitatif, statistika inferensial, regresi, hingga teknik-teknik mutakhir seperti analisis big data, data mining, serta implementasinya menggunakan perangkat lunak populer seperti Python, R, dan SPSS. Setiap bab dilengkapi dengan contoh kasus nyata dan panduan praktis agar pembaca mampu mengaplikasikan langsung metode yang dipelajari.

Ucapan terima kasih kami sampaikan kepada semua pihak yang telah membantu dalam penyusunan buku ini, terutama keluarga, kolega, dan mahasiswa yang telah memberikan masukan berharga. Semoga buku ini dapat memberikan manfaat dan kontribusi nyata bagi pengembangan ilmu pengetahuan, khususnya dalam penerapan metode kuantitatif di era big data.

Akhir kata, selamat membaca dan semoga buku ini menjadi inspirasi bagi pembaca sekalian.

Tim Penulis

DAFTAR ISI

BAB 1
PENGANTAR ERA BIG DATA DAN TRANSFORMASI METODOLOGI KUANTITATIF

Yuni Roza, S.Kom., M.Kom.
Institut Teknologi Perusahaan Listrik Negara

Pendahuluan

Perkembangan teknologi *digital* dalam dua dekade terakhir telah menghasilkan peningkatan jumlah data secara eksponensial. Aktivitas manusia di internet, penggunaan perangkat *mobile* seperti pada saat menggunakan media sosial seperti menekan *button "like",* melakukan transaksi perbankan daring, mengirim pesan instan, hingga detak jantung yang terekam oleh jam tangan pintar semuanya menciptakan jejak *digital*. Jejak transaksi *digital*, sensor untuk penerapan *Internet of Things* (IoT), serta sistem informasi organisasi menghasilkan data dalam volume yang sangat besar setiap hari. Fenomena ini dikenal sebagai era *big data.*

Big data itu tidak hanyalah soal "ukuran yang besar", akan tetapi jika kita melihat lebih dalam, *big data* adalah tentang kompleksitas dan kecepatan, seperti menyerupai air terjun yang deras, tidak hanya bervolume raksasa, tetapi juga datang dengan kecepatan tinggi (*velocity*) dan bentuk yang sangat beragam (*variety*) mulai dari teks, gambar, video, hingga *log* sensor mesin yang rumit. Kehadiran *big data* telah memicu transformasi fundamental dalam metodologi penelitian kuantitatif, dimana pendekatan tradisional yang sebelumnya sangat bergantung pada survei dan eksperimen dengan sampel terbatas, kini beralih menuju pemanfaatan *dataset* yang jauh lebih masif dan kompleks. Evolusi dari analisis statistik klasik ke analitik data besar ini memungkinkan para peneliti untuk mengeksplorasi pola, tren, dan hubungan tersembunyi yang sebelumnya sulit dijangkau oleh teknik statistik konvensional. Transformasi ini memungkinkan peneliti untuk mengidentifikasi pola yang lebih luas, memprediksi tren masa depan, serta menghasilkan model analitik yang lebih akurat. Oleh karena itu, pemahaman terhadap *big data* dan transformasi metodologi kuantitatif menjadi penting bagi peneliti modern. Diagram arus evolusi *big data* (1990-an-masa depan):

Gambar 1.1: Evolusi Big Data
Sumber: Diolah Penulis.

1. **Pra-2000-an: Era Awal (*Early Era*)**
   a. Karakteristik: fokus pada penyimpanan data terstruktur yang rapi, seperti data keuangan atau inventaris.
   b. Teknologi: sistem manajemen basis data relasional (RDBMS), SQL, dan penggunaan media penyimpanan tradisional seperti *floppy disk.*
   c. Masalah: skalabilitas terbatas. Menambah kapasitas penyimpanan sangat mahal (vertikal).

2. **Awal 2000-an: Embrio *Big Data* (*Genesis*)**
   a. Karakteristik: munculnya web 2.0 dan ledakan data dari perusahaan internet besar. RDBMS mulai runtuh karena tidak sanggup menangani volume data yang tak terstruktur.
   b. Teknologi: publikasi *paper MapReduce* oleh *Google* (2004) dan lahirnya *Hadoop* (2006). Ini memperkenalkan konsep *Distributed Storage* & *Processing* (HDFS), yang memungkinkan pemrosesan data di ribuan komputer standar.

3. **2010-an: Ekosistem 5V (*Ecosystem*)**
   a. Karakteristik: era dimana data meledak dalam *Volume*, *Velocity*,

Variety, Veracity, dan *Value*. Data datang dari media sosial, video, dan sensor IoT (*variety*), mengalir secara *real-time* (*velocity*).

b. Teknologi: munculnya sistem NoSQL untuk data fleksibel. Lahirnya *apache spark* (2014) yang memperkenalkan *in-memory computing*, mempercepat pemrosesan ribuan kali lebih cepat dari *Hadoop*. *Real-time analytics* menjadi standar.

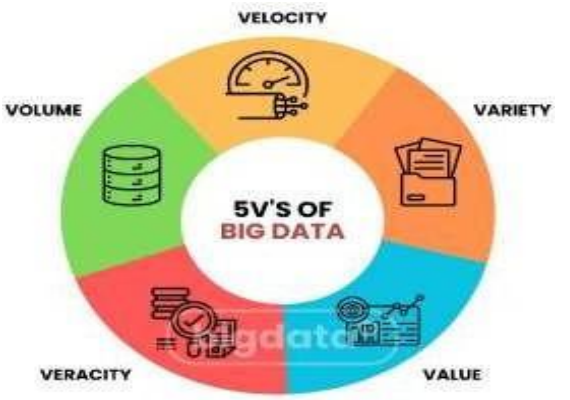
4. Masa Depan: Era Wawasan & AI (*Insight* & AI)

a. Karakteristik: tantangannya bukan lagi "menyimpan", melainkan "memahami". Data yang masif kini menjadi "bahan bakar" untuk AI. Wawasan harus ditemukan secara otomatis.

b. Teknologi: integrasi mendalam dengan *machine learning*, pengambilan keputusan berbasis AI (AI-*Driven Decisions*), pemrosesan data *streaming* yang terus-menerus (*ubiquitous data*), dan analitik prediktif. Kita beralih dari sekadar melihat "apa yang terjadi" menjadi "apa yang mungkin terjadi."

Konsep Big Data

Big data kumpulan data besar yang terus bertambah secara drastis dari waktu ke waktu. *Big data* adalah kumpulan data yang sangat besar dan rumit sehingga tidak ada teknologi manajemen data yang dapat menyimpan atau memprosesnya secara efektif (Varudharajulu & Ma, 2018).



Gambar 1.2: Karakteristik Big Data
Sumber: Diolah Penulis.

Daftar Pustaka

Chen, M., Mao, S., & Liu, Y. (2022). Big Data: A Survey. *Mobile Networks And Applications*.

Creswell, J. W., & Creswell, J. D. (2023). *Research Design: Qualitative, Quantitative, And Mixed Methods Approaches*. Sage Publications.

Hassani, H., Huang, X., & Silva, E. (2023). Big Data And Analytics: From Data To Knowledge. *Technological Forecasting And Social Change*.

Karich, I. P., & Joss, S. (2025). *Emergence And Evolution of Big Data Research. Metrics*.

Kitchin, R. (2021). *The Data Revolution: Big Data, Open Data, Data Infrastructures And Their Consequences*. Sage Publications.

Provost, F., & Fawcett, T. (2023). *Data Science For Business*. O'Reilly Media.

Shmueli, G., Bruce, P., Yahav, I., Patel, N., & Lichtendahl Jr., K. (2023). *Data Mining For Business Analytics: Concepts, Techniques, And Applications*. Wiley.

Tosi, D., Kokaj, R., & Roccetti, M. (2024). Fifteen Years of Big Data: A Systematic Literature Review. *Journal of Big Data*.

PROFIL PENULIS

Yuni Roza, S.Kom., M.Kom.

Dilahirkan dari keluarga sederhana di salah satu daerah kecil di Sumatera Barat, pinggir danau Singkarak. Menyelesaikan studi S1 dan S2 dengan *background* komputer di Universitas Putra Indonesia YPTK Padang. Penulis memiliki keterkaitan ilmu dalam bidang ilmu *controlling* berbasis IoT, *machine learning*, dan di bidang *web programming*. Dalam menjalan profesi sebagai dosen, penulis aktif sebagai peneliti di bidang keilmuannya tersebut baik secara internal maupun pembiayaan hibah DIKTI.

Selain meneliti, penulis juga aktif menulis artikel yang sudah terindeks sinta dan skala nasional. Terus meningkatkan kemampuan dalam bidang ilmu komputer dan kepemimpinan melalui pengalaman praktisi, pembelajaran berkelanjutan serta *sharing session* dengan yang lainnya.

Email Penulis: yuni.roza17@gmail.com.

BAB 2
PARADIGMA BARU DALAM ANALISIS DATA: DARI *SMALL DATA* KE *BIG DATA*

Dr. Ir. Afiyati, S.Si., M.T.
Universitas Mercu Buana

Definisi Era *Small Data*

Small data adalah kumpulan data berukuran kecil hingga sedang (biasanya dalam satuan GB), yang mudah diolah dengan perangkat lunak sederhana seperti Excel, SPSS, atau *database* SQL. Analisisnya berbasis hipotesis (*hypothesis-driven*), menggunakan sampel representatif untuk menarik kesimpulan. Kelebihan: kualitas tinggi dan mudah diverifikasi, biaya rendah, interpretasi cepat dan *actionable*.

Kekurangan: terbatas dalam generalisasi, kurang mampu menangkap pola kompleks di era *digital*, Contoh: survei pelanggan 1.000 orang atau laporan penjualan bulanan. *Big data* lahir karena ledakan data dari internet, *smartphone*, IoT, dan media sosial. Definisi sederhana: data yang terlalu besar, cepat, dan kompleks untuk diolah dengan metode tradisional. Paradigma baru ini menggeser fokus dari "apa yang kita tahu" menjadi "apa yang bisa kita temukan dari data massal".

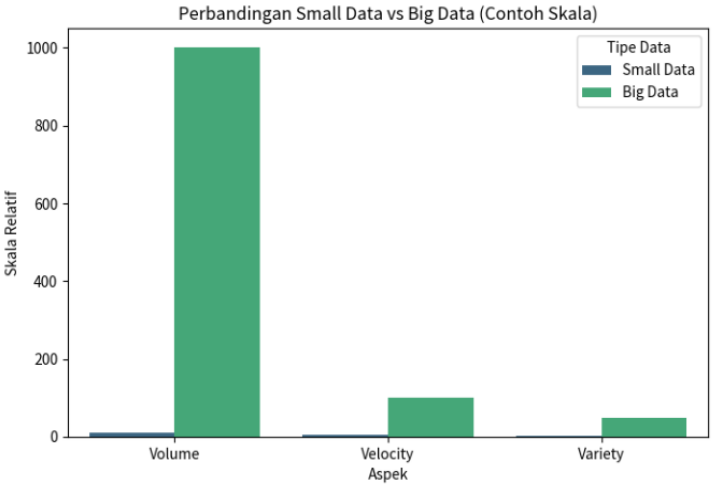
Gambar 2.1: Karakteristik *Big Data* 5V
Sumber: Diolah Penulis.

Karakteristik *Big Data*-5V

Big Data tidak hanya soal "banyak data", melainkan memiliki lima karakteristik inti yang dikenal sebagai 5V (*Volume, Velocity, Variety, Veracity, Value*). Karakteristik ini menjadi fondasi paradigma baru analisis data. Berikut penjelasan rinci:

1. **Volume**
   Ukuran data yang sangat besar, mencapai skala *zettabytes* (1 ZB = 1 miliar *terabyte*). Contoh: satu pesawat terbang menghasilkan 20–40 TB data per jam penerbangan. Volume ini membuat metode tradisional tidak lagi memadai.
2. ***Velocity***
   Kecepatan generasi dan pemrosesan data yang sangat tinggi (*real-time*). *Data streaming* dari sensor IoT, media sosial, dan transaksi *e-commerce* datang dalam hitungan milidetik. Analisis harus dilakukan secara cepat agar tetap relevan.
3. ***Variety***
   Keragaman format data: *structured* (tabel *database*), *semi-structured* (JSON, XML), dan *unstructured* (teks, gambar, video, audio). Ini menuntut *tools* yang fleksibel untuk mengolah semua jenis data sekaligus.
4. ***Veracity***
   Tingkat keandalan dan kualitas data. Data *big data* seringkali "kotor" (*noise*, duplikat, bias, atau tidak lengkap). *Veracity* menekankan pentingnya *data cleaning* dan *governance* agar *insight* yang dihasilkan akurat.
5. ***Value***
   Nilai bisnis yang dapat diekstrak dari data. Ini adalah "V" yang paling penting data hanya bernilai jika menghasilkan keputusan yang lebih baik, efisiensi operasional, atau inovasi baru.

Perbandingan *small data* vs *big data*:



Gambar 2.2: Grafik Perbandingan *Small Data* Vs *Big Data*
Sumber: Diolah Penulis.

Grafik batang di atas menunjukkan perbandingan skala relatif 3 aspek utama. *Small data* adalah data yang "manusiawi" mudah dipahami, diakses, dan dianalisis oleh individu atau tim kecil. Karena ukurannya terbatas, kualitas data biasanya tinggi dan analisisnya lebih fokus pada penjelasan mendalam (mengapa sesuatu terjadi). Kelebihan: biaya rendah dan cepat menghasilkan keputusan, mudah diverifikasi dan minim bias, cocok untuk bisnis kecil, riset akademik, atau analisis spesifik.

Kekurangan: terbatas dalam menemukan pola kompleks atau tren jangka panjang, kurang mampu menangani data *real-time* atau beragam sumber, *big data* muncul karena ledakan data dari internet, *smartphone*, IoT, dan media sosial. Ia ditandai oleh 5V (*Volume, Velocity, Variety, Veracity, Value*). Paradigma ini menggeser fokus dari "kita tahu apa" menjadi "apa yang bisa kita temukan dari data massal".

Kelebihan: mampu mengungkap insight tersembunyi, prediksi akurat, dan personalisasi massal (contoh: rekomendasi *netflix* atau *matching driver gojek*), *scalable* untuk pertumbuhan bisnis besar. Kekurangan: sulit dikelola (85% proyek *big data* gagal karena kompleksitas), masalah privasi, kualitas data, dan kebutuhan skill tinggi, biaya infrastruktur dan pemrosesan lebih mahal.

Daftar Pustaka

Anuradha, J. (2015). A Brief Introduction on Big Data 5Vs Characteristics And Hadoop Technology, *Procedia Computer Science*, 48, pp. 319–324. Available at: https://doi.org/10.1016/j.procs.2015.04.188.

Coursera (2025) *Big Data Vs. Small Data: What's the difference?* Available at: https://www.coursera.org/articles/big-data-vs-small-data (Accessed: 1 April 2026).

Faraway, J. And Augustin, N. (2017). *When Small Data Beats Big Data*. University of Bath. Available at: https://www.maths.bath.ac.uk/~jjf23/papers/smallvbig.pdf (Accessed: 1 April 2026).

Hadi, H.J., Shnain, A.H., Hadishaheed, S. and Ahmad, A.H. (2015). Big Data And Five V's Characteristics, *International Journal of Advances in Electronics And Computer Science*, 2(1), pp. 16–23. Available at: http://www.iraj.in/journal/journal_file/journal_pdf/12-105-142063747116-23.pdf.

Nyikana, W. (2023). The Logical Differentiation Between Small Data And Big Data, *South African Journal of Information Management*, 25(1). Available at: https://scielo.org.za/pdf/sajim/v25n1/31.pdf.

Santoso, J.T. (2022). *Analisis Big Data*. Semarang: Yayasan Prima Agus Teknik.

Statista (2025). *Volume of Data Created, Captured, Copied, And Consumed Worldwide From 2010 to 2029*. Available at: https://www.statista.com/statistics/871513/worldwide-data-created/ (Accessed: 1 April 2026).

UpGrad (2025). *Characteristics of big data: Types & 5V's*. Available at: https://www.upgrad.com/blog/characteristics-of-big-data/ (Accessed: 1 April 2026).

PROFIL PENULIS

Dr. Ir. Afiyati, S.Si., M.T.

Lahir di Pekanbaru, 16 Oktober 1969 penulis menyelesaikan studi S1 di Universitas Gadjah Mada prodi Ilmu Komputer pada tahun 1988-1994 dan S2 di prodi Magister Teknik Elektro Universitas Mercu Buana, kemudian melanjutkan S3 di Universitas Gadjah Mada prodi Doktoral Ilmu Komputer Sebagai penulis yang berpengalaman dalam bidang teknologi dan pendidikan, saya percaya bahwa Metode Kuantitatif Era *Big Data* adalah metode yang sangat menarik untuk dieksplorasi dan dipelajari oleh pemula maupun profesional dalam dunia data sains.

Buku Metode Kuantitatif Era *Big Data* menyajikan rekonstruksi paradigmatik metode penelitian kuantitatif di tengah ledakan data digital yang masif. Di era di mana volume data global diperkirakan mencapai lebih dari 180 *zettabyte* pada tahun 2025, pendekatan kuantitatif tradisional yang berbasis sampel kecil, hipotesis deduktif, dan analisis statistik konvensional menghadapi tantangan fundamental. Buku ini mengusulkan kerangka baru yang mengintegrasikan fondasi metodologi kuantitatif klasik dengan karakteristik dan teknologi *big data*.

Secara sistematis, buku membahas pergeseran epistemologis dari pendekatan *hypothesis-driven* menuju *data-driven research*, dari analisis berbasis sampel menuju analisis pada *full population dataset*, serta dari pemrosesan *batch* tradisional menuju pemrosesan *real-time* dan *distributed computing*. Pembaca diajak memahami lima karakteristik utama big data (5V: *Volume, Velocity, Variety, Veracity,* dan *Value*) beserta implikasinya terhadap desain penelitian, pengumpulan data, validitas, reliabilitas, dan generalisasi temuan.

Buku ini menguraikan secara mendalam teknik pengumpulan data di era *digital* (API *scraping*, sensor IoT, *streaming data*, dan data media sosial), pra-pemrosesan data besar (*data cleaning, feature*

engineering, dimensionality reduction), serta metode analisis kuantitatif lanjutan yang mencakup *data mining, machine learning, predictive modeling,* dan *prescriptive analytics*. Selain itu, dibahas pula integrasi *framework big data* seperti *Hadoop* dan *Apache Spark*, serta pemanfaatan bahasa pemrograman *Python* (*PySpark*) dalam konteks penelitian kuantitatif. Dengan menulis buku ini, saya berharap dapat menyebarkan pengetahuan tentang *big data* dan memberikan kontribusi positif bagi komunitas data sains secara luas. Saya berkomitmen untuk menyajikan materi yang akurat, relevan, dan terkini sesuai dengan perkembangan terbaru dalam dunia Data Sains.

Email Penulis: afiyati.reno@mercubuana.ac.id.

BAB 3
FONDASI MATEMATIKA DAN STATISTIK UNTUK BIG *DATA ANALYTICS*

Mohamad Yusuf, S.Kom., M.C.S.
Universitas Mercu Buana Jakarta

Pendahuluan

Di era *digital* sekarang ini, *big data* telah menjadi salah satu faktor utama yang mendorong perubahan di berbagai sektor seperti bisnis, kesehatan, pemerintahan, ilmu pengetahuan, dan kebijakan publik. Namun, hanya memiliki data dalam jumlah yang sangat besar tidak cukup untuk menciptakan nilai. Di sinilah peran matematika dan statistik menjadi penting sebagai "bahasa" universal yang membantu kita memahami, mengolah, dan mengambil makna dari volume data yang begitu besar.

Mengapa matematika dan statistik menjadi "bahasa" *big data*: *big data* secara tradisional diidentifikasi melalui empat dimensi utama yang dikenal sebagai 4V: *Volume*, yang menunjukkan besarnya jumlah data yang dapat mencapai *terabyte* hingga *zettabyte; Velocity*, yang merujuk pada kecepatan dalam menghasilkan dan memproses data, sering kali dalam skala *real-time* atau mendekatinya; *Variety*, yang melibatkan beragam format dan sumber data, mulai dari yang terstruktur seperti tabel *database*, semi-terstruktur seperti JSON atau XML, hingga yang tidak terstruktur seperti teks, gambar, sensor IoT, log sistem, dan video; serta *Veracity*, yang mencerminkan tingkat ketidakpastian, *noise*, inkonsistensi, dan ketidakakuratan yang melekat pada data. Keempat karakteristik ini menciptakan tantangan kompleks yang tidak dapat diatasi hanya dengan kekuatan komputasi biasa (*Data Basecamp Substack*, 2021). Di sinilah matematika dan statistik berperan sebagai "bahasa" universal *big data*. Matematika menyediakan kerangka formal untuk merepresentasikan data dalam bentuk vektor, matriks, tensor, graf, atau distribusi probabilitas, sehingga data yang beragam dapat dipetakan ke dalam ruang matematis yang terstruktur.

Sementara itu, statistik memberikan metode yang efektif untuk menangani ketidakpastian dan noise melalui inferensi statistik, pemodelan probabilistik, serta pendekatan statistik yang tangguh. Selain itu, kedua bidang ini mendukung penemuan pola tersembunyi di tengah volume dan variasi data yang besar, menggunakan teknik seperti pengurangan dimensi, pengelompokan, dan penambangan aturan asosiasi.

Algoritma matematis dan statistik yang efisien secara komputasi, seperti algoritma aproksimasi, *sketching*, dan *sampling*, menjamin

kecepatan pemrosesan yang praktis. Tanpa kerangka matematika-statistik ini, *big data* hanya akan menjadi kumpulan *bit* dan *byte* yang masif tanpa makna.

Paradigma Matematika: Klasik Vs *Big Data*

Matematika yang umumnya diajarkan di sekolah dan kuliah awal yang bisa disebut sebagai matematika klasik sangat berbeda dari matematika yang diperlukan dalam konteks *big data.* Matematika klasik biasanya berurusan dengan data berukuran kecil hingga sedang (puluhan hingga ribuan sampel), dengan fokus pada solusi eksak dan bukti yang sangat ketat.

Komputasi dilakukan secara sekuensial pada satu mesin menggunakan metode analitik yang sering menghasilkan solusi bentuk tertutup, seperti invers matriks penuh atau least squares tradisional. Sebaliknya, matematika *big data* menghadapi data dalam skala yang sangat besar (miliaran hingga triliunan observasi), sehingga prioritas beralih ke masalah skalabilitas, aproksimasi, dan *trade-off* antara akurasi dan kecepatan (Dr. Budi Raharjo, 2022; Prof. Agus Yodi Gunawan, 2024). Alih-alih mengejar presisi sempurna, pendekatan ini menerima hasil aproksimasi yang "cukup baik" asalkan dapat dihitung dengan cepat dan efisien. Proses komputasi juga menjadi paralel dan terdistribusi, memanfaatkan kerangka kerja seperti *Spark, Dask*, atau *cluster* GPU (P. Kardani dkk, 2024). Metode yang dominan menjadi iteratif, stokastik, dan berbasis optimasi numerik, seperti *Stochastic Gradient Descent* (SGD), *mini-batch gradient descent, Adam, sketching, randomized Singular Value Decomposition* (SVD), serta *distributed gradient descent*.

Singkatnya, dalam dunia *big data,* ketepatan yang sempurna sering kali dikorbankan demi meningkatkan skalabilitas dan kecepatan. Sebagai contoh, alih-alih menghitung invers matriks secara eksak yang memiliki kompleksitas $O(n^3)$ dan sulit untuk diskalakan, praktisi lebih memilih teknik aljabar linier acak atau aproksimasi *Nyström* yang memungkinkan pemrosesan di ratusan mesin dengan waktu yang jauh lebih singkat, meskipun hasilnya berupa aproksimasi berkualitas tinggi.

Aljabar Linear Untuk Representasi Data Besar

Di zaman *big data*, aljabar linear lebih dari sekadar pelajaran matematika dasar; ia berfungsi sebagai bahasa utama untuk mewakili, memproses, dan mengambil wawasan dari data yang sangat besar. Hampir semua algoritma *machine learning* dan analisis data modern mulai dari regresi hingga *deep learning* berdasarkan pada operasi aljabar linear yang efisien (Gilbert Strang, 2019; Marc Peter Deisenroth, 2020). Sub-bab ini akan menjelaskan secara rinci bagaimana data direpresentasikan sebagai vektor dan matriks, operasi dasar yang diperlukan, serta konsep eigenvalue, eigenvector, dan *Singular Value Decomposition* (SVD) serta aplikasinya yang sangat kuat di skala besar.

1. **Vektor dan Matriks Sebagai Representasi *Dataset* (n × p)**

   Dataset besar sering kali disimpan dalam bentuk matriks dua dimensi $X \in R^{nxp}$, dimana (Marc Peter Deisenroth, 2020):

   n = jumlah observasi (baris),

   p = jumlah fitur atau variabel (kolom).

   Setiap baris X_i (vektor l x p) mencerminkan satu sampel data lengkap, sedangkan setiap kolom $X_{.j}$ menunjukkan nilai dari satu fitur di seluruh sampel. Representasi ini sangat efisien karena memungkinkan dilakukannya operasi vektor-matriks secara paralel dan terdistribusi (Benjamin Draves, 2019). Sebagai contoh, *dataset* yang terdiri dari 1 juta pelanggan dengan 500 fitur (seperti usia, pendapatan, riwayat belanja, dll.) dapat diwakili dalam matriks berukuran 1.000.000 x 500.

Daftar Pustaka

Benjamin Draves. (2019, November 29). *Probabilistic Matrix Factorization*. Probabilistic Matrix Factorization.

Data Basecamp Substack. (2021, November 26). *Big Data-Definition And 4 V's*. Https://Databasecamp.de/En/Data/Big-Data-Basics.

Dr. Budi Raharjo. (2022). *Ilmu Big Data dan Mesin Cerdas*. Semarang: Yayasan Prima Agus Teknik.

GeekForGeek. (2025, July 23). *Vector Norms*. Https://Www.Geeksforgeeks.Org/Maths/Vector-Norms/.

Geekforgeek. (2025, December 11). *Transpose of a Matrix*. Https://Www.Geeksforgeeks.Org/Maths/Transpose-of-a-Matrix/.

Gilbert Strang. (2019). *Linear Algebra And Learning from Data*.

Junaid Qazi, P. (2022, January 30). *A36: K-Nearest Neighbors (KNN) — Understand With Hands-on Code!* Https://Junaidsqazi.Medium.Com/.

Khosravy, M., Nitta, N., Nakamura, K., & Babaguchi, N. (2020). Compressive Sensing Theoretical Foundations In A Nutshell. In *Compressive Sensing In Healthcare* (pp. 1–24). Elsevier. https://doi.org/10.1016/B978-0-12-821247-9.00006-8

Marc Peter Deisenroth, A. A. F. C. S. O. (2020). *Mathematics For Machine Learning*.

Markelic. (2025, November 15). *A Visual Tutorial For Matrix Multiplication*. Https://Markelic.de/a-Visual-Tutorial-for-Matrix-Multiplication/.

P. Kardani dkk. (2024). Perbandingan Kinerja Infrastruktur Paralel Dalam Pengolahan Data Besar Menggunakan Apache Spark. *JINACS (Jurnal Ilmiah Nasional Dan Komputer Sains), Universitas Negeri Surabaya*.

Paul Dawkin. (2026). *Section 11.3 : Dot Product*. Https://Tutorial.Math.Lamar.Edu/Classes/Calcii/Dotproduct.Asp.

Prof. Agus Yodi Gunawan. (2024). *Pemodelan Matematika: Simplifikasi Dunia Nyata*. Https://Fgb.Itb.Ac.Id/Wp-

Content/Uploads/Sites/26/2024/09/Ebook-Prof.-Agus-Yodi-Gunawan-Pemodelan-Matematika-Simplifikasi-Dunia-Nyata.Pdf.

Team AlgoDaily. (2022). *What Is The Manhattan Distance?* Https://Algodaily.Com/Lessons/What-Is-the-Manhattan-Distance.

PROFIL PENULIS

Mohamad Yusuf, S.Kom., M.C.S.

Dilahirkan di Jakarta pada tanggal 7 September 1976. Pendidikan S1 diselesaikan pada tahun 2001 di Universitas Budi Luhur Jurusan Teknik Informatika. Selanjutnya, pendidikan S2 di *Preston University* Islamabad Pakistan dan saat ini sedang menyelesaikan S3 di Universiti Malaysia Kelantan pada bidang *Data Science.* Penulis adalah Dosen Tetap Universitas Mercu Buana Jakarta.

Buku populer yang telah dihasilkan adalah Pemrograman *Mobile* dengan *Flutter*, Pemrograman *Java*, Aplikasi *Mobile* Teori dan Praktek, Teknik 2D/3D *Blender* Untuk Pemula hingga Ahli, Bahasa Pemrograman *Python* dan Pembelajaran Mesin dan Kecerdasan Buatan Teori dan Aplikasi Praktis. Selain itu juga telah menulis buku tentang Tata Kelola Teknologi Informasi. Untuk menghubungi penulis di email: mhd.yusuf@mercubuana.ac.id

BAB 4
PROBABILITAS DAN INFERENSI STATISTIK DALAM KONTEKS *BIG DATA*

Firmansyah Apryadhi, S.Kom., M.TI.
Institut Teknologi Perusahaan Listrik Negara

Pendahuluan

Perkembangan teknologi informasi dan komunikasi telah menghasilkan ledakan data dalam jumlah yang sangat besar, yang dikenal sebagai *big data.* Data tidak lagi hanya berasal dari survei atau eksperimen terkontrol, tetapi juga dari media sosial, sensor IoT, transaksi *digital*, dan berbagai *platform* daring lainnya. Fenomena ini mengubah secara fundamental cara data dikumpulkan, disimpan, dan dianalisis. Dalam konteks ini, probabilitas dan inferensi statistik tetap menjadi fondasi utama untuk mengekstraksi informasi dan membuat keputusan berbasis data.

Statistika secara umum didefinisikan sebagai ilmu yang berkaitan dengan pengumpulan, pengolahan, analisis, interpretasi, dan penyajian data. Probabilitas berperan sebagai dasar teoritis yang memungkinkan statistika mengukur ketidakpastian dan membuat prediksi mengenai kejadian yang belum terjadi. Menurut Liu (2017), era *big data* tidak menghilangkan peran statistika, tetapi justru memperluas ruang lingkup dan kompleksitas analisis data karena volume, variasi, dan kecepatan data yang meningkat secara drastis. Probabilitas menyediakan kerangka matematis untuk memodelkan ketidakpastian dan variasi dalam data. Dengan probabilitas, peneliti dapat menghitung peluang terjadinya suatu peristiwa, mengevaluasi risiko, serta membangun model prediktif.

Tanpa konsep probabilitas, analisis data besar hanya akan menghasilkan kumpulan angka tanpa makna inferensial. Di sisi lain, inferensi statistik berfungsi untuk menarik kesimpulan mengenai populasi berdasarkan data yang diamati. Dalam konteks *big data,* peran inferensi menjadi semakin penting karena data yang besar tidak selalu berarti data yang representatif. Chen (2023) menekankan bahwa meskipun *big data* menyediakan jumlah observasi yang sangat besar, data tersebut seringkali bersifat bising, tidak terstruktur, dan mengandung bias, sehingga tetap memerlukan metode inferensi statistik untuk menghasilkan kesimpulan yang valid.

Big data juga menghadirkan paradoks dalam statistika: ukuran sampel yang sangat besar dapat meningkatkan presisi estimasi, tetapi pada saat yang sama meningkatkan risiko *overfitting*, bias seleksi, dan akumulasi *noise*. *National Science Review* menyatakan bahwa

heterogenitas dan dimensi tinggi dalam *big data* membuat proses inferensi statistik menjadi lebih kompleks dan memerlukan metode komputasi serta regularisasi yang lebih canggih dibandingkan pendekatan statistik klasik.

Dengan demikian, probabilitas dan inferensi statistik tidak hanya tetap relevan dalam era *big data*, tetapi justru menjadi semakin penting. Probabilitas memberikan dasar teoritis untuk memahami ketidakpastian dalam data besar, sementara inferensi statistik memungkinkan peneliti dan praktisi *data science* untuk menarik kesimpulan yang dapat digeneralisasi dan digunakan dalam pengambilan keputusan strategis. Dalam praktiknya, integrasi antara statistika klasik dan teknik komputasi modern seperti *machine learning* dan *data mining* telah melahirkan bidang baru yang dikenal sebagai *statistical learning.*

Bidang ini menggabungkan teori probabilitas, inferensi statistik, dan algoritma komputasi untuk menganalisis data dalam skala besar secara efisien dan akurat. Oleh karena itu, pemahaman yang kuat mengenai probabilitas dan inferensi statistik merupakan prasyarat utama bagi siapa pun yang ingin bekerja dalam bidang *data science,* kecerdasan buatan, maupun analitik *big data*. Tanpa fondasi tersebut, analisis *big data* berisiko menghasilkan kesimpulan yang menyesatkan, meskipun didukung oleh volume data yang sangat besar.

Dasar-Dasar Probabilitas

1. Pengertian Probabilitas

Probabilitas merupakan cabang matematika yang mempelajari kemungkinan terjadinya suatu peristiwa. Konsep probabilitas digunakan untuk mengukur tingkat ketidakpastian dalam berbagai fenomena yang bersifat acak. Dalam kehidupan sehari-hari maupun dalam penelitian ilmiah, probabilitas menjadi alat penting untuk memprediksi kejadian di masa depan berdasarkan informasi yang tersedia. Menurut Sheldon M. Ross, probabilitas adalah suatu ukuran numerik yang menggambarkan kemungkinan terjadinya suatu peristiwa dalam suatu percobaan acak (Ross, 2014).

Dengan kata lain, probabilitas memberikan nilai antara 0 dan 1 yang menunjukkan seberapa besar kemungkinan suatu kejadian

terjadi. Sementara itu, Walpole, Myers, Myers, dan Ye menyatakan bahwa probabilitas merupakan ukuran kuantitatif dari peluang suatu kejadian yang dapat digunakan untuk membuat keputusan dalam kondisi ketidakpastian (Walpole *et al.,* 2012). Oleh karena itu, probabilitas menjadi dasar penting dalam analisis statistik, pengambilan keputusan, serta berbagai bidang seperti ekonomi, teknik, dan ilmu data.

2. Konsep Percobaan Acak

Percobaan acak adalah suatu proses atau eksperimen yang menghasilkan hasil yang tidak dapat dipastikan sebelumnya, meskipun dilakukan dalam kondisi yang sama. Contoh percobaan acak antara lain melempar koin, melempar dadu, atau memilih sampel dari suatu populasi.

Menurut Devore, percobaan acak adalah proses yang menghasilkan satu dari beberapa kemungkinan hasil yang diketahui sebelumnya, tetapi hasil yang tepat tidak dapat diprediksi secara pasti sebelum percobaan dilakukan (Devore, 2016). Dalam konteks probabilitas, setiap hasil dari percobaan acak disebut *outcome* atau hasil percobaan.

3. Ruang Sampel dan Kejadian

Dalam teori probabilitas, konsep penting yang harus dipahami adalah ruang sampel dan kejadian.

a. Ruang Sampel (*Sample Space*), ruang sampel adalah himpunan semua kemungkinan hasil yang dapat terjadi dari suatu percobaan acak. Ruang sampel biasanya dilambangkan dengan huruf S. Sebagai contoh, jika sebuah dadu dilempar sekali, maka ruang sampelnya adalah: S = {1, 2, 3, 4, 5, 6}
b. Kejadian (*Event*), kejadian adalah himpunan bagian dari ruang sampel yang terdiri dari satu atau lebih hasil percobaan. Misalnya, kejadian munculnya angka genap pada pelemparan dadu adalah: A = {2, 4, 6}. Menurut Grimmett dan Stirzaker, kejadian dalam probabilitas merupakan himpunan hasil dari ruang sampel yang menjadi fokus pengamatan dalam suatu eksperimen acak (Grimmett & Stirzaker, 2001).

Daftar Pustaka

Bishop, C. M. (2006). *Pattern Recognition And Machine Learning*. Springer.

Bruce, P., & Bruce, A. (2017). *Practical Statistics for Data Scientists*. O'Reilly Media.

Casella, G., & Berger, R. L. (2002). *Statistical Inference (2nd ed.).* Duxbury.

Casella, G., & Berger, R. L. (2002). *Statistical Inference.* Duxbury Press.

Chen, L.-P. (2023). *Statistical Inference And Machine Learning For Big Data.* Biometrics, 79(4), 4013–4025.

DeGroot, M. H., & Schervish, M. J. (2012). *Probability And Statistics (4th ed.).* Pearson.

Devore, J. L. (2016). *Probability And Statistics For Engineering And The Sciences.* Cengage Learning.

Feller, W. (1968). *An Introduction To Probability Theory And Its Applications.* Wiley.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis (3rd ed.).* CRC Press.

Grimmett, G., & Stirzaker, D. (2001). *Probability And Random Processes.* Oxford University Press.

Hand, D. J. (2018). *Statistical Challenges of Big Data*. Annual Review of Statistics and Its Application.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning.* Springer.

Hogg, R. V., McKean, J., & Craig, A. T. (2019). *Introduction to Mathematical Statistics.* Pearson.

Laney, D. (2001). 3D Data Management: *Controlling Data Volume, Velocity, And Variety*. META Group Research Note.

Liu, G. (2017). Statistics In The Age of Big Data: Opportunities And Challenges. *Proceedings of The 3rd International Symposium on Social Science.*

Montgomery, D. C., & Runger, G. C. (2014). *Applied Statistics And Probability for Engineers*. Wiley.

Murphy, K. P. (2022). *Probabilistic Machine Learning: An Introduction*. MIT Press.

National Science Review. (2014). *Challenges Of Big Data Analysis.*

O'Neil, C. (2016). *Weapons of Math Destruction*. Crown Publishing.

Pearson, K. (1895). *Contributions To The Mathematical Theory of Evolution.* Philosophical Transactions of The Royal Society.

Qiu, P. (2016). *Big Data? More Challenges! Technometrics*, 58(3), 283–284.

Ross, S. M. (2014). *Introduction to Probability Models*. Academic Press.

Ross, S. M. (2019). *Introduction to Probability Models (12th ed.).* Academic Press.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information*. Graphics Press.

Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2012). *Probability And Statistics For Engineers And Scientists*. Pearson.

Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2017). *Probability And Statistics For Engineers And Scientists (9th ed.).* Pearson.

Wickham, H., & Grolemund, G. (2017). *R For Data Science*. O'Reilly Media.

Wu, H. (2024). *Statistics Evolution And Revolution To Meet Data Science Challenges.* Statistics In Biosciences.

PROFIL PENULIS

Firmansyah Apryadhi, S.Kom., M.TI.

Minat penulis terhadap dunia ilmu komputer mulai tumbuh sejak tahun 2004. Ketertarikan tersebut mendorong penulis untuk menempuh pendidikan tinggi di Universitas Bina Nusantara, memilih Jurusan Ilmu Komputer dengan konsentrasi pada bidang Teknik Informatika. Penulis berhasil menyelesaikan pendidikan jenjang Sarjana (S1) dan lulus pada tahun 2008. Guna memperdalam pemahaman dan kompetensinya di bidang Teknologi Informasi, penulis melanjutkan studi ke jenjang Magister (S2) di Universitas Indonesia, dan pada tahun 2009 berhasil meraih gelar Magister Teknologi Informasi, dengan peminatan pada Teknologi Informasi.

Saat ini, penulis aktif sebagai dosen tetap di lingkungan pendidikan tinggi, tepatnya pada Program Studi Sistem Informasi di bawah naungan LLDIKTI Wilayah III, bertugas di Institut Teknologi PLN. Dalam kapasitasnya sebagai tenaga pengajar, penulis mengampu berbagai mata kuliah yang berfokus pada pengembangan kompetensi teknis dan manajerial mahasiswa, antara lain: Rekayasa Perangkat Lunak, Konsep Basis Data, Pengantar Kecerdasan Buatan, Manajemen dan Akuisisi Sistem Informasi, Proyek Teknologi Informasi, Kecerdasan Buatan, Pengantar Bisnis dan Manajemen, dan *Technopreneurship*.

Selain mengajar, penulis juga aktif melakukan penelitian dan publikasi ilmiah. Karya-karya ilmiah penulis dapat ditemukan dan diakses melalui portal akademik seperti SINTA (*Science and Technology Index*) dan *Google Scholar*. Dengan latar belakang akademik yang kuat dan pengalaman mengajar yang luas, penulis terus berkomitmen untuk berkontribusi dalam pengembangan ilmu pengetahuan dan teknologi, serta membimbing generasi muda dalam menghadapi tantangan dunia *digital* yang semakin kompleks.

Email Penulis: penulisbayangan2@google.com.

BAB 5
ARSITEKTUR SISTEM *BIG DATA, HADOOP, SPARK* DAN *CLOUD COMPUTING*

Nila Natalia, M.Kom.
Politeknik Sukabumi

Arsitektur Sistem *Big Data*

Merancang sebuah sistem yang mampu menangani *Big Data* kesehatan memerlukan sebuah cetak biru arsitektural yang secara fundamental berbeda dari sistem basis data relasional tradisional. Tujuannya bukan lagi sekadar menyimpan dan mengambil data, tetapi untuk membangun sebuah alur kerja (*pipeline*) yang mampu menyerap, memproses, dan menganalisis aliran data yang masif dan heterogen secara andal dan efisien (Kuo *et al.,* 2022).

Arsitektur *Big Data* yang efektif harus mampu menyeimbangkan dua kebutuhan yang seringkali bertentangan: kebutuhan akan analisis data historis yang komprehensif dan akurat (pemrosesan *batch*) dengan kebutuhan akan respons instan terhadap data yang baru masuk (pemrosesan *streaming* atau *real-time*). Untuk mengatasi dualisme ini, arsitektur seperti Lambda dan Kappa telah dikembangkan.

Arsitektur Lambda, yang dipopulerkan oleh Nathan Marz, secara eksplisit memisahkan alur data menjadi dua jalur: "lapisan *batch*" (*batch layer*) dan "lapisan kecepatan" (*speed layer*) (Kaur & Kumar, 2021). Lapisan *batch* secara periodik menganalisis keseluruhan set data historis untuk menghasilkan pandangan yang paling akurat dan lengkap. Sementara itu, lapisan kecepatan secara terus-menerus memproses data yang baru masuk untuk memberikan pembaruan dan analisis secara *real-time*, meskipun dengan tingkat akurasi yang mungkin sedikit lebih rendah. Hasil dari kedua lapisan ini kemudian digabungkan di "lapisan penyajian" (*serving layer*) untuk memberikan jawaban yang komprehensif kepada pengguna.

Komponen inti dari setiap arsitektur *Big Data* dapat dipecah menjadi tiga lapisan fungsional utama. Lapisan penyerapan (*ingestion layer*) adalah gerbang depan sistem, bertanggung jawab untuk mengumpulkan data dari berbagai sumber, seperti RME, perangkat IoT, dan sistem laboratorium, dan memasukkannya ke dalam sistem (Ghasemi *et al.,* 2023). Lapisan penyimpanan (*storage layer*) harus mampu menampung volume data yang sangat besar dalam berbagai format.

Sistem penyimpanan terdistribusi seperti HDFS (*Hadoop Distributed File System*) menjadi pilihan umum karena kemampuannya

untuk menyimpan data di banyak mesin secara redundan dan berbiaya rendah (Ghimire & Thapa, 2022). Terakhir, lapisan pemrosesan (*processing layer*), yang merupakan otak dari sistem, adalah tempat data dianalisis menggunakan kerangka kerja seperti *Hadoop MapReduce* atau *Apache Spark*. Tantangan utama yang dihadapi oleh arsitektur ini adalah skalabilitas untuk menangani karakteristik unik data medis.

Volume data genomik dan citra medis yang terus meledak menuntut sistem penyimpanan dan pemrosesan yang dapat diperluas dengan mudah. *Velocity* dari data *streaming* monitor pasien menuntut lapisan kecepatan yang berlatensi sangat rendah untuk deteksi dini kondisi kritis. *Variety* data, yang mencakup teks, angka, gambar, dan sinyal, menuntut arsitektur yang fleksibel dan dapat mengintegrasikan serta menganalisis berbagai format data ini secara bersamaan (Haleem *et al.,* 2022).

Merancang arsitektur yang mampu mengatasi ketiga tantangan ini secara simultan adalah kunci untuk membangun *platform* analitik kesehatan yang benar-benar transformatif. Analogi**:** bayangkan Anda mengelola sebuah perpustakaan nasional yang sangat besar yang menerima ribuan buku dan koran baru setiap jam. Arsitektur Lambda adalah seperti memiliki dua tim pustakawan yang berbeda. Tim pertama (lapisan *batch*) bekerja semalaman untuk secara teliti membaca, mengkatalogkan, dan mengindeks setiap buku yang pernah ada di perpustakaan untuk membuat ensiklopedia yang paling lengkap dan akurat.

Tim kedua (lapisan kecepatan) berdiri di pintu depan, dengan cepat memindai setiap koran yang baru masuk dan menulis ringkasan berita utama di papan pengumuman. Pengunjung perpustakaan (pengguna) dapat membaca ringkasan cepat di papan pengumuman (data *real-time*) sambil menunggu pustakawan mengambilkan ensiklopedia yang paling akurat (data *batch*) dari ruang arsip untuk jawaban yang lebih mendalam.

Konsep Dasar Arsitektur Lambda dan Kappa Dalam Data Kesehatan

Arsitektur Lambda dirancang untuk sistem yang membutuhkan toleransi kesalahan yang tinggi dan keakuratan data historis yang

absolut. Dalam konteks kesehatan, lapisan *batch* dapat digunakan untuk melatih model prediktif penyakit kronis dengan menganalisis jutaan rekam medis historis. Proses ini mungkin memakan waktu berjam-jam atau berhari-hari, tetapi menghasilkan model yang sangat akurat. Di sisi lain, lapisan kecepatan dapat menganalisis data *streaming* dari perangkat *wearable* seorang pasien untuk mendeteksi anomali detak jantung secara instan dan mengirimkan peringatan (Kaur & Kumar, 2021).

Lapisan penyajian kemudian dapat menampilkan peringatan *real-time* ini bersama dengan profil risiko kronis jangka panjang pasien yang dihitung oleh lapisan *batch*. Arsitektur Kappa muncul sebagai penyederhanaan dari Lambda. Pendukung arsitektur Kappa berpendapat bahwa dengan kemajuan teknologi pemrosesan *streaming* seperti *Apache Spark* dan *Apache Flink*, kini dimungkinkan untuk melakukan semua jenis analisis (baik *real-time* maupun *batch*) hanya dengan menggunakan satu alur pemrosesan berbasis *stream* (Ayyadevara, 2022).

Dalam arsitektur ini, tidak ada lagi lapisan *batch*. Jika analisis ulang seluruh data historis diperlukan, sistem hanya akan memutar ulang (*replay*) seluruh aliran data historis melalui mesin pemroses *stream* yang sama. Dalam konteks kesehatan, ini berarti sistem yang sama yang mendeteksi anomali detak jantung juga dapat digunakan untuk melatih model prediktif dengan memutar ulang data detak jantung historis selama bertahun-tahun. Arsitektur Kappa lebih sederhana untuk dikelola karena hanya ada satu basis kode, tetapi sangat bergantung pada kematangan dan kecepatan teknologi pemrosesan *stream*.

Komponen Utama: *Ingestion*, *Storage*, dan *Processing Layer*

Lapisan penyerapan (*ingestion layer*) menggunakan alat seperti *Apache Kafka* atau *Rabbit*MQ untuk membuat "saluran" data yang andal dan dapat diskalakan. Alat-alat ini berfungsi sebagai antrean pesan, memungkinkan berbagai sistem sumber untuk mengirimkan data tanpa membebani sistem pemrosesan secara langsung dan memastikan tidak ada data yang hilang bahkan jika sistem

Daftar Pustaka

Ahmed, S., & Singh, P. K. (2021). Big Data In Healthcare: A Comprehensive Review. *Health And Technology*, 11(3), 479-491. https://doi.org/10.1007/s12553-021-00543-y.

Al-Ghofari, F., & Hidayanto, A. N. (2023). Analisis Kesiapan Implementasi Platform SATUSEHAT Berdasarkan Persepsi Tenaga Kesehatan Di Indonesia. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 7(5), 2345-2354.

Ayyadevara, V. K. (2022). *Modern Big Data Architectures*. Apress.

Ghasemi, M., Z. H., & A. A. (2023). A Systematic Review on Big Data Ingestion Techniques And Frameworks. *Journal of Big Data*, 10(1), 1-25.

Ghimire, B., & Thapa, C. (2022). A Review of Hadoop Ecosystem For Big Data Processing. *Journal of Big Data*, 9(1), 1-19.

Griebel, L., M. S., & H. S. (2023). Cloud Computing In Healthcare: A Review of The Literature And Research Challenges. *Journal of Medical Internet Research*, 25, e45942.

Haleem, A., Javaid, M., Singh, R., & Suman, R. (2022). Medical 4.0 Technologies For Healthcare: Features, Capabilities, And Applications. *Internet of Things And Cyber-Physical Systems*. https://doi.org/10.1016/j.iotcps.2022.04.001.

Javed, A. R., Shahzad, F., ur Rehman, S., Zikria, Y. B., Razzak, I., Jalil, Z., & Xu, G. (2022). Future Smart Cities: Requirements, Security Challenges, And Future Directions. *Future Generation.*

Computer Systems, 129, 25-39. https://doi.org/10.1016/j.future.2021.10.027.

Kaur, P., & Kumar, R. (2021). A Review on Lambda Architecture For Big Data Processing. *Journal of King Saud University-Computer and Information Sciences*, 33(10), 1215-1225.

Kementerian Kesehatan Republik Indonesia. (2022). *Blueprint Strategi Transformasi Digital Kesehatan 2024*. Kementerian Kesehatan RI.

Kuo, T. T., Kim, H. E., & Ohno-Machado, L. (2022). Blockchain

Distributed Ledger Technologies For Biomedical And Health Care Applications. *Journal of The American Medical Informatics Association*, 24(6), 1211-1220.

Li, C., Wang, J., Wang, S., & Zhang, Y. (2024). A Review of IoT Applications In Healthcare. *Neurocomputing*, 565, 127017. https://doi.org/10.1016/j.neucom.2023.127017.

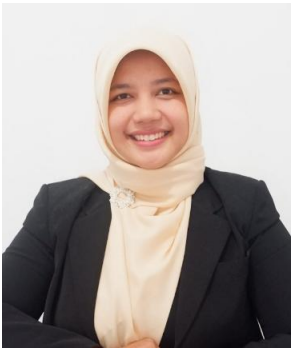
Li, X., Wang, Y., & Chen, Y. (2023). A Big Data-Driven Approach For Chronic Disease Prediction And Management. *IEEE Journal of Biomedical and Health Informatics*, 27(4), 1835-1846.

Shao, M., Fan, J., Huang, Z., & Chen, M. (2022). The Impact of Information And Communication Technologies (ICTs) on Health Outcomes: A Mediating Effect Analysis Based on Cross-National Panel Data. *Journal of Environmental And Public Health*, 2022. https://doi.org/10.1155/2022/2225723.

Stoumpos, A., Kitsios, F., & Talias, M. (2023). Digital Transformation in Healthcare: Technology Acceptance And Its Applications. *International Journal of Environmental Research and Public Health*, 20. https://doi.org/10.3390/ijerph20043407.

Yadav, A., S. S., & A. G. (2022). A Review of Big Data Analytics In Genomics. *Briefings in Bioinformatics*, 23(3), bbac093.

PROFIL PENULIS



Nila Natalia, S.T., M.Kom.

Lahir di Karang Anyar pada 5 Desember 1989 yang disapa Mrs. Nila, adalah dosen, praktisi dan motivator. Alumnus 2010 Teknik Informatika LP3I Bandung, 2012 STMIK JABAR Bandung, 2018 STMIK Likmi Bandung, Sejak 2019, Politeknik Sukabumi menetapkan beliau mengajar di jenjang diploma tiga. Penulis ditempatkan pada prodi Teknik Komputer Politeknik Sukabumi, 2024.

Menempuh pendidikan Doktor di Universiti Malaysia Kelantan, Selain itu beliau juga sebagai direktur utama CV. Inovasi Sejahtera. Sampai buku ini diterbitkan, aktif melaksanakan kolaborasi penelitian, pengabdian dan Ikut serta dalam keanggotaan organisasi profesi seperti APTIKOM dan Kopertip Indonesia.

Email Penulis: nilanataliainovasi@gmail.com.

BAB 6
DATABASE NOSQL DAN SISTEM PENYIMPANAN TERDISTRIBUSI

Luthfia Fauzia Dewi Aryanti, S.Kom., M.TI.
Universitas Raharja Tangerang

Pendahuluan

Perkembangan teknologi informasi yang sangat pesat telah menghasilkan volume data yang luar biasa besar, dengan kecepatan dan keberagaman jenis yang tidak pernah terjadi sebelumnya. Fenomena ini dikenal sebagai *big data*, yang ditandai oleh tiga karakteristik utama: *volume, velocity*, dan *variety*. Dalam konteks ini, sistem manajemen basis data relasional (*Relational Database Management System*/RDBMS) yang telah digunakan selama beberapa dekade mulai menunjukkan keterbatasan, terutama ketika harus menangani data dalam skala besar, tidak terstruktur, atau tersebar secara geografis.

Untuk menjawab tantangan tersebut, lahirlah paradigma baru dalam pengelolaan data, yaitu *Database NoSQL* (*Not Only SQL*) dan Sistem Penyimpanan Terdistribus**i**. *Database NoSQL* menawarkan pendekatan *non-relasional* yang fleksibel, mampu menangani data dalam berbagai format seperti dokumen, *key-value*, kolom, atau graf. Teknologi ini tidak hanya mengatasi keterbatasan skema kaku dalam RDBMS, tetapi juga memungkinkan skalabilitas horizontal, di mana sistem dapat diperluas hanya dengan menambahkan *server* baru.

Sementara itu, sistem penyimpanan terdistribusi memungkinkan data disimpan di berbagai lokasi atau node dalam jaringan, dengan mekanisme replikasi dan *fault tolerance* yang tinggi. Sistem ini mendukung ketersediaan data yang tinggi (*high availability*) dan toleransi terhadap kegagalan (*fault tolerance*), yang sangat krusial bagi sistem modern yang beroperasi secara *real-time* dan terus-menerus.

Penerapan NoSQL dan sistem penyimpanan terdistribusi tidak hanya terbatas pada perusahaan teknologi raksasa seperti *Google*, *Amazon*, dan *Facebook*, tetapi juga telah merambah berbagai sektor industri, termasuk *e-commerce*, keuangan, kesehatan, transportasi, hingga pemerintahan. Oleh karena itu, pemahaman yang mendalam mengenai kedua konsep ini sangat penting bagi para profesional TI, pengembang perangkat lunak, hingga pengambil keputusan teknologi dalam organisasi (Sadalage & Fowler, 2013).

Bab ini akan mengulas secara komprehensif konsep dasar, arsitektur, jenis-jenis, kelebihan, tantangan, serta studi kasus

implementasi Database NoSQL dan Sistem Penyimpanan Terdistribusi, sehingga pembaca memiliki landasan teoritis dan praktis yang kuat untuk memahami dan mengimplementasikan solusi teknologi ini dalam berbagai skenario dunia nyata (Sadalage & Fowler, 2013).

Latar Belakang Munculnya NoSQL

Perkembangan sistem informasi modern menuntut pengelolaan data dalam volume sangat besar, format yang sangat beragam, serta kecepatan pemrosesan yang tinggi. Dalam banyak kasus, sistem tradisional berbasis *Relational Database Management System* (RDBMS) mengalami kesulitan dalam memenuhi kebutuhan ini.

RDBMS didesain untuk bekerja optimal dengan struktur data yang terorganisir secara ketat dalam tabel-tabel yang memiliki skema tetap. Ketika digunakan untuk menangani data tidak terstruktur, atau data yang tumbuh sangat cepat secara horizontal (skala pengguna atau data meningkat pesat), sistem ini kerap menemui kendala dalam performa, ketersediaan, dan skalabilitas.

Kondisi inilah yang melatarbelakangi munculnya *database NoSQL* (*Not Only SQL*), yaitu sistem manajemen basis data *non-relasional* yang dirancang untuk menangani kebutuhan penyimpanan dan pengolahan data skala besar secara fleksibel, terdistribusi, dan efisien. Istilah "NoSQL" pertama kali diperkenalkan sekitar tahun 2009 oleh Johan Oskarsson, namun konsepnya telah berkembang jauh lebih awal seiring dengan kebutuhan perusahaan teknologi seperti *Google*, *Amazon*, dan *Facebook* dalam mengelola data berskala masif.

Definisi dan Karakteristik Utama

Secara umum, NoSQL merujuk pada kelas sistem manajemen basis data yang tidak sepenuhnya bergantung pada struktur relasional dan bahasa SQL untuk operasi datanya. Meskipun istilah "NoSQL" seolah menolak SQL, dalam praktiknya banyak sistem NoSQL modern yang tetap mendukung semacam bahasa *query*, meski bukan SQL murni. Oleh karena itu, istilah tersebut kini lebih dimaknai sebagai "*Not Only* SQL", yaitu mencakup pendekatan alternatif selain SQL (Sadalage & Fowler, 2013). Ciri khas utama NoSQL meliputi:

1. **Skalabilitas Horizontal**
   Sistem NoSQL dirancang untuk mudah diskalakan dengan menambahkan lebih banyak *server* (*node*), bukan dengan meningkatkan kapasitas *server* tunggal (skala vertikal). Hal ini memungkinkan pertumbuhan sistem secara elastis sesuai kebutuhan.
2. **Model Data Fleksibel**
   Tidak memiliki skema tetap seperti tabel di RDBMS. Artinya, struktur data dapat berubah-ubah antar entitas atau dokumen dalam satu koleksi yang sama. Ini sangat berguna untuk data tidak terstruktur atau semi-terstruktur, seperti data sensor, *log* aplikasi, *metadata*, JSON, dan sebagainya.
3. **Kinerja Tinggi Untuk Beban Kerja Tertentu**
   Sistem NoSQL dioptimalkan untuk performa baca/tulis tinggi, terutama dalam skenario data berukuran besar dan akses cepat secara paralel (Sadalage & Fowler, 2013).
4. **Arsitektur Terdistribusi**
   Sebagian besar sistem NoSQL mendukung penyimpanan dan pemrosesan terdistribusi, dengan kemampuan replikasi otomatis, *sharding* (pembagian data), serta *fault-tolerance* (ketahanan terhadap kegagalan node).
5. **Prinsip BASE (*Basically Available, Soft State, Eventually Consistent*)**
   Tidak seperti RDBMS yang menerapkan prinsip ACID, NoSQL cenderung mengorbankan konsistensi jangka pendek demi ketersediaan dan toleransi partisi jaringan.
6. **Evolusi dari ACID ke BASE**
   Sebagai perbandingan, RDBMS mengedepankan prinsip ACID*:*
   a. *Atomicity*: transaksi dijalankan secara utuh atau tidak sama sekali.
   b. *Consistency*: data selalu dalam keadaan valid sesuai aturan.
   c. *Isolation*: transaksi berjalan independen satu sama lain.
   d. *Durability*: data tersimpan secara permanen setelah transaksi selesai.

NoSQL menggunakan pendekatan BASE*, basically available*: sistem selalu tersedia untuk menerima permintaan, *soft state*: data

Daftar Pustaka

A. Brewer. (2012). *CAP Twelve Years Later: How The 'Rules' Have Changed, Computer*, Vol. 45, No. 2, pp. 23–29.

Chang et al., (2006). Bigtable: A Distributed Storage System for Structured Data, *In Proc*. OSDI.

Corbett et al., (2012). Spanner: Google's Globally Distributed Database, *In Proc.* OSDI.

DeCandia et al., (2007). Dynamo: Amazon's Highly Available Key-Value Store, *In Proc*. SOSP.

Ghemawat, H. Gobioff, And S.-T. Leung, (2003). The Google File System, *In Proc.* SOSP,

Harrison. (2015). *Next Generation Databases: NoSQL And Big Data.* New York, NY, USA: Apress.

J. Sadalage And M. Fowler (2013). *NoSQL Distilled: A Brief Guide To The Emerging World of Polyglot Persistence*. Boston, MA, USA: Addison-Wesley.

Kleppmann. (2017). *Designing Data-Intensive Applications*. Sebastopol, CA, USA: O'Reilly Media.

Marz and J. Warren. (2015). *Big Data: Principles And Best Practices of Scalable Real-Time Data Systems*. Shelter Island, NY, USA: Manning Publications.

Redmond and J. R. Wilson. (2012). *Seven Databases In Seven Weeks.* Dallas, TX, USA: Pragmatic Bookshelf.

S. Tanenbaum And M. Van Steen. (2007). *Distributed Systems: Principles And Paradigms, (2nd ed).* Upper Saddle River, NJ, USA: Prentice Hall.

Silberschatz, H. F. Korth, And S. Sudarshan. (2019). *Database System Concepts, (7th ed).* New York, NY, USA: McGraw-Hill.

White. (2015). *Hadoop: The Definitive Guide, (4th ed).* Sebastopol, CA, USA: O'Reilly Media.

PROFIL PENULIS

Luthfia Fauzia Dewi Aryanti, S.Kom., M.TI.

Ketertarikan penulis terhadap perkembangan teknologi informasi dimulai saat sejak duduk dibangku SMA pada tahun 2009. Hal tersebut yang mendorong penulis memilih mengambil jurusan ilmu komputer pada Universitas Pendidikan Indonesia (UPI) Bandung dan penulis berhasil lulus pada tahun 2013. Setelah lulus penulis mencoba untuk menjadi seorang *programmer* pada perusahaan swasta. Karena kecintaan penulis terhadap teknologi, akhirnya penulis memutuskan untuk melanjutkan pendidikan di Universitas Raharja Tangerang pada tahun 2015 dan berhasil lulus dengan predikat *cumlaude* pada tahun 2017.

Setelah lulus S2 penulis mengabdikan diri sebagai dosen pada bidang teknologi informasi di Universitas Raharja sampai saat ini. Selain aktif mengajar dan meneliti, saat ini penulis juga menjalani profesi sebagai analis sistem pada perusahaan swasta. Kedua profesi tersebut dijalani sehari-hari dengan penuh semangat, dengan harapan dapat memberikan kontribusi positif dalam perkembangan teknologi baik di dalam dunia kampus ataupun didalam perusahaan tempat bekerja saat ini. Dengan semakin berkembangnya dunia teknologi semoga semakin banyak membantu manusia untuk mempermudah kehidupannya.

Email Penulis: fiafauzia@raharja.info

BAB 7
VISUALISASI DATA UNTUK DATASET BERSKALA BESAR

Dr. Nungky Awang Chandra, S.Si., M.TI.
Universitas Mercu Buana

Pendahuluan

Perkembangan *big data* telah mengubah cara organisasi, peneliti, pemerintah, dan industri memahami fenomena yang kompleks. Data tidak lagi hadir dalam ukuran kecil dan terstruktur rapi, melainkan dalam volume sangat besar, kecepatan tinggi, serta ragam format yang beragam. Dalam konteks ini, visualisasi data tidak cukup dipahami sebagai "membuat grafik", tetapi sebagai proses analitis untuk menerjemahkan data besar menjadi pola, tren, anomali, dan wawasan yang dapat dipahami manusia. Layanan analitik modern seperti Apache Spark menekankan kemampuan analitik skala besar, termasuk *exploratory data analysis* pada skala sangat besar, sementara *platform* seperti *Azure Data Explorer* dan *Microsoft Fabric* dirancang untuk analisis interaktif atas volume data yang tinggi.

Pada *dataset* berskala besar, tantangan utama visualisasi bukan hanya bagaimana menampilkan data, tetapi bagaimana menjaga keseimbangan antara ketepatan, kecepatan respons, keterbacaan visual, dan beban komputasi. Literatur *visual analytics* menunjukkan bahwa visualisasi *big data* menghadapi persoalan mendasar seperti *visual clutter, latency*, kebutuhan komputasi tinggi, dan keterbatasan persepsi manusia dalam menangkap terlalu banyak elemen sekaligus. Karena itu, pendekatan seperti *sampling, aggregation, binning, filtering, progressive visualization,* dan pemanfaatan mesin komputasi terdistribusi menjadi sangat penting.

Bab ini membahas landasan teoretis dan implementasi praktis visualisasi data untuk dataset berskala besar. Uraian dimulai dari konsep dasar, karakteristik *big data*, tantangan visualisasi, teknik reduksi data, arsitektur implementasi, pemilihan jenis visual, hingga contoh penerapan pada *dashboard* analitik modern. Fokus utamanya adalah bagaimana metode kuantitatif tetap relevan di era *big data* melalui dukungan visualisasi yang tepat, efisien, dan dapat dipertanggungjawabkan secara analitis.

Konsep Dasar Visualisasi Data Dalam Era *Big Data*

Visualisasi data adalah representasi grafis dari data untuk mendukung pemahaman, komunikasi, dan pengambilan keputusan. Dalam pendekatan tradisional, visualisasi sering digunakan untuk

merangkum *dataset* yang relatif kecil, misalnya melalui diagram batang, diagram garis, *histogram, scatter plot*, atau *pie chart.* Namun pada era *big data,* visualisasi berkembang menjadi bagian dari *visual analytics*, yaitu integrasi antara komputasi, interaksi manusia, dan representasi visual untuk mengekstraksi pengetahuan dari data berskala besar.

Literatur ACM tentang *big data visualization* menegaskan bahwa visualisasi bukan hanya sarana presentasi hasil akhir, tetapi juga medium eksplorasi analitik. Dalam lingkungan *big data*, visualisasi memiliki tiga fungsi utama. Pertama, fungsi eksploratif, yaitu membantu analis menemukan pola yang belum diketahui sebelumnya. Kedua, fungsi eksplanatif, yaitu menyampaikan hasil analisis secara jelas kepada pembuat keputusan. Ketiga, fungsi *monitoring*, yaitu memantau perubahan data secara *real time* atau *near-real time* melalui *dashboard.*

Platform analitik modern mendukung ketiga fungsi ini dengan menggabungkan pemrosesan skala besar dan layer visual interaktif. *Spark*, misalnya, mendukung SQL *analytics, batch*/*stream processing*, dan EDA pada data yang sangat besar, sedangkan *Power* BI/*Fabric* menyediakan *semantic model* dan mode akses seperti *Direct Lake* atau *DirectQuery* untuk analisis interaktif. Secara teoretis, visualisasi *big data* berada pada titik temu antara statistik, ilmu komputer, *human-computer interaction,* dan desain informasi. Statistik menyediakan metode kuantitatif untuk merangkum data; ilmu komputer menyediakan infrastruktur pengolahan; desain informasi memastikan keterbacaan; sedangkan interaksi manusia memastikan pengguna dapat mengeksplorasi data secara bertahap tanpa kehilangan konteks. Karena itu, keberhasilan visualisasi *big data* tidak hanya diukur dari keindahan tampilan, tetapi juga dari akurasi representasi, efisiensi query, dan kemampuan pengguna memahami insight yang ditampilkan.

Karakteristik *Dataset* Berskala Besar

Dataset berskala besar umumnya dijelaskan melalui konsep 5V, yaitu *Volume, Velocity, Variety, Veracity,* dan *Value*. *Volume* merujuk pada ukuran data yang sangat besar, dari *gigabyte* hingga *petabyte*. *Hadoop*

HDFS, misalnya, dirancang untuk *file* berukuran *gigabyte* sampai terabyte dan untuk skala ratusan *node*, menegaskan bahwa infrastruktur *big data* memang dibangun untuk data dalam ukuran masif. *Velocity* mengacu pada kecepatan data dihasilkan dan diproses.

Pada sistem modern seperti IoT, media sosial, *log* aplikasi, dan transaksi *digital*, data dapat mengalir terus-menerus sehingga visualisasi harus mendukung pemutakhiran cepat. *Azure Data Explorer* secara eksplisit diposisikan untuk analisis *real-time* atas volume besar data *streaming* dari aplikasi, situs web, dan perangkat IoT. *Variety* menjelaskan keberagaman format data, mulai dari tabel relasional, JSON, *log* teks, citra, video, sinyal sensor, hingga *graph data*. Keragaman ini membuat visualisasi *big data* tidak bisa hanya mengandalkan satu jenis grafik.

Sering kali diperlukan kombinasi visual, misalnya *time series* untuk aliran waktu, *heatmap* untuk kepadatan, geospatial map untuk lokasi, *network graph* untuk relasi, dan *box plot* untuk distribusi. *Veracity* berkaitan dengan kualitas, keandalan, dan konsistensi data. Pada skala besar, masalah *missing values*, duplikasi, *noise, outlier*, dan ketidaksesuaian semantik menjadi lebih kompleks. Karena itu, visualisasi perlu didukung *preprocessing* yang baik agar tidak menyesatkan interpretasi. *Value* adalah nilai yang dapat dihasilkan dari data. *big data* tidak otomatis bermanfaat; nilainya muncul jika data dapat dipahami dan ditransformasikan menjadi keputusan. Di sinilah visualisasi berfungsi sebagai jembatan antara data mentah dan *insight* yang bernilai.

Tantangan Visualisasi untuk Data Skala Besar

1. ***Overplotting* dan *Visual Clutter***
   Saat jutaan titik data langsung ditampilkan, elemen visual akan saling bertumpuk dan kehilangan makna. *Scatter plot* tradisional, misalnya, menjadi gelap dan padat sehingga pola distribusi sulit dikenali. Kondisi ini disebut *overplotting* atau *visual clutter*. Tantangan ini merupakan salah satu tema utama dalam kajian visualisasi *big data.*
2. **Keterbatasan Performa Interaktif**
   Visualisasi efektif biasanya menuntut interaksi cepat: *filter, zoom, drill-down, brush,* dan *cross-highlight*. Namun pada data besar, tiap

Daftar Pustaka

Apache Spark. (2025). *Apache Spark: Unified Engine For Large-Scale Data Analytics*. Available At: https://spark.apache.org/.

Heer, J., Shneiderman, B., Perer, A. And Others. (2023). Progressive Visual Analytics: Making Sense of Large-Scale Data Over Time, *IEEE Transactions on Visualization And Computer Graphics*, 29(1), pp. 1–15.

Kandel, S., Heer, J., Plaisant, C., Kennedy, J., van Ham, F., Riche, N.H., Weaver, C. And Lee, B. (2012). Research Directions In Data Wrangling: Visualizations And Transformations For Usable And Credible Data, *Information Visualization*, 10(4), pp. 271–288.

Keim, D.A., Kohlhammer, J., Ellis, G. And Mansmann, F. (2010). *Mastering The Information Age: Solving Problems With Visual Analytics*. Goslar: Eurographics Association.

Microsoft Learn. (2025). *Azure Data Explorer Documentation*. Available At: https://learn.microsoft.com/en-us/azure/data-explorer/.

Microsoft Learn. (2025). *Direct Lake Overview-Microsoft Fabric*. Available At: https://learn.microsoft.com/en-us/fabric/fundamentals/direct-lake-overview.

Munzner, T. (2014). *Visualization Analysis And Design*. Boca Raton: CRC Press.

Tableau. (2025). *Tableau and Big Data: An Overview*. Available at: https://www.tableau.com/learn/whitepapers/tableau-big-data-overview.

PROFIL PENULIS

Dr. Nungky Awang Chandra, M.TI., S.Si.

Nungky Awang Chandra, lahir di Semarang 1973. Penulis selain dosen teknik informatika fasilkom universitas mercubuana, juga berprofesi sebagai auditor sistem manajemen keamanan informasi ISO 27001, ISO22301, ISO 27701, ISO 20000, ISO 42001 yang teregister di BSSN. Penulis merupakan lulusan pendidikan Sarjana S1 jurusan Fisika Komputasi Institut Teknologi Bandung pada tahun 1998. Kemudian pada tahun 2007 melanjutkan pendidikan master di bidang Magister Teknologi Informasi Universitas Indonesia, menyelesaikan studinya pada tahun 2009.

Pada tahun 2022 penulis juga menyelesaikan studi S3 di Universitas Indonesia dengan disertasi dan publikasi jurnal bereputasi internasional tentang keamanan siber dan manajemen risiko keamanan siber. Selain itu pada tahun 2023 penulis juga menyelesaikan studi *postgraduate cyber security* di *Massachusetts Institute of Technology* (MIT). Penulis memiliki kepakaran di bidang keamanan siber, pengembangan aplikasi, *drone*. Guna mewujudkan karir sebagai dosen profesional, penulis pun aktif sebagai peneliti di bidang kepakarannya tersebut.

Beberapa penelitian yang telah dilakukan didanai oleh internal perguruan tinggi dan juga Kemenristek DIKTI. Selain peneliti, penulis juga aktif menulis buku dengan harapan dapat memberikan kontribusi positif bagi bangsa dan negara yang sangat tercinta ini. Adapun untuk koresponden dengan penulis dapat email ke penulis dengan email penulis: nungkyac707@gmail.com.

BAB 8
TIME SERIES ANALYSIS PADA DATA BERSKALA MASIF

Umniy Salamah, S.T., MMSI.
Universitas Mercu Buana

Pendahuluan

Perkembangan teknologi *digital* telah menghasilkan data dalam jumlah sangat besar dari berbagai sumber seperti IoT, transaksi, media sosial, dan *cloud*. Fenomena ini dikenal sebagai *big data* yang ditandai oleh volume, kecepatan, dan keberagaman data (Mayer-Schönberger and Cukier, 2014).

Dalam konteks ini, data deret waktu (*time series*) menjadi dominan karena sebagian besar data modern memiliki dimensi waktu. Analisis *time series* pada skala masif merupakan pengembangan metode klasik yang harus mampu menangani kompleksitas data dari sisi penyimpanan, pemrosesan, dan analisis (Hyndman and Athanasopoulos, 2013).

Berbeda dengan pendekatan tradisional, analisis modern menekankan pada skalabilitas, efisiensi komputasi, dan pemrosesan *real-time*. Oleh karena itu, diperlukan integrasi antara metode statistik, *machine learning*, dan komputasi terdistribusi.

Karakteristik *Data Time Series* Berskala Masif

Data deret waktu (*time series*) dalam skala masif memiliki karakteristik yang jauh lebih kompleks dibandingkan data konvensional. Kompleksitas ini tidak hanya disebabkan oleh ukuran data yang besar, tetapi juga oleh kecepatan pertumbuhan, keberagaman sumber, serta kualitas data yang bervariasi. Dalam kajian big data, karakteristik tersebut umumnya dijelaskan melalui pendekatan 5V, yaitu *Volume, Velocity, Variety, Veracity*, dan *Value* (Usama, Liu and Chen, 2017). Dalam konteks data deret waktu, kelima aspek ini berperan penting dalam menentukan metode pengolahan, penyimpanan, serta teknik analisis yang digunakan.

1. ***Volume***

   Volume mengacu pada jumlah data yang sangat besar yang dihasilkan dalam sistem modern. Pada data deret waktu, volume yang besar umumnya disebabkan oleh frekuensi pencatatan data yang tinggi serta banyaknya sumber data yang beroperasi secara simultan. Sebagai contoh, sistem berbasis *Internet of Things* (IoT) dapat menghasilkan jutaan titik data dalam waktu singkat, sementara *platform digital* seperti *e-commerce* atau media sosial mencatat aktivitas pengguna secara terus-menerus.

Besarnya volume data ini menimbulkan tantangan signifikan dalam penyimpanan dan pemrosesan (Prayitno, Perdana and Nasyuha, 2025). Sistem tradisional yang bergantung pada satu mesin tidak lagi mampu menangani data dalam skala tersebut secara efisien. Oleh karena itu, diperlukan pendekatan berbasis komputasi terdistribusi yang memungkinkan data diproses secara paralel pada banyak *node*.

Dengan demikian, volume tidak hanya menjadi indikator ukuran data, tetapi juga menentukan kebutuhan infrastruktur yang digunakan dalam analisis data deret waktu berskala masif.

2. *Velocity*

Velocity menggambarkan kecepatan data dihasilkan, dikirim, dan diproses. Dalam konteks data deret waktu modern, data tidak hanya dikumpulkan secara berkala, tetapi mengalir secara kontinu dalam bentuk *streaming* (White, 2015). Hal ini terutama terlihat pada sistem *real-time* seperti pasar saham, sensor kendaraan otonom, maupun sistem monitoring industri.

Kecepatan aliran data ini menuntut sistem untuk mampu melakukan pemrosesan secara cepat, bahkan dalam waktu yang hampir bersamaan dengan saat data dihasilkan. Pendekatan *batch* processing yang bersifat *offline* menjadi kurang relevan dalam banyak kasus, sehingga digantikan oleh metode *stream processing*. Selain itu, model analisis juga perlu dirancang agar mampu beradaptasi secara dinamis terhadap data yang terus berubah. Dengan demikian, *velocity* menjadi faktor kunci yang mendorong penggunaan teknologi *real-time analytics* dalam analisis time series berskala masif.

3. *Variety*

Variety merujuk pada keberagaman jenis dan format data yang dihasilkan. Dalam sistem modern, data deret waktu tidak hanya berbentuk numerik terstruktur, tetapi juga mencakup data semi-terstruktur seperti log sistem, serta data tidak terstruktur seperti gambar, audio, dan video. Keberagaman ini muncul karena data berasal dari berbagai sumber yang heterogen, seperti sensor, aplikasi *digital*, perangkat *mobile*, dan sistem *enterprise*.

Keberagaman format data ini menimbulkan tantangan dalam integrasi dan pengolahan data. Data yang berbeda format memerlukan teknik *preprocessing* yang berbeda pula sebelum dapat dianalisis secara bersama-sama. Selain itu, proses ekstraksi fitur menjadi lebih kompleks karena harus mampu mengubah berbagai jenis data menjadi representasi yang dapat digunakan oleh model analitik. Oleh karena itu, *variety* menuntut adanya sistem pengolahan data yang fleksibel dan mampu menangani berbagai format data secara efisien.

4. *Veracity*

Veracity berkaitan dengan kualitas dan keandalan data. Pada data deret waktu berskala masif, kualitas data sering kali tidak konsisten karena berbagai faktor, seperti gangguan sensor, kesalahan pencatatan, maupun keterlambatan transmisi data. Hal ini dapat menyebabkan munculnya data yang hilang (*missing values*), data yang mengandung noise, maupun nilai ekstrim (*outlier*) yang tidak merepresentasikan kondisi sebenarnya (Han, Kamber and Pei, 2012).

Kualitas data yang rendah dapat berdampak langsung pada performa model analisis, karena model akan belajar dari data yang tersedia. Jika data mengandung banyak kesalahan, maka hasil prediksi yang dihasilkan juga akan menjadi tidak akurat. Oleh karena itu, proses *preprocessing* menjadi sangat penting dalam memastikan kualitas data sebelum dilakukan analisis. Teknik seperti imputasi data, deteksi *outlier*, dan *filtering noise* merupakan langkah yang umum digunakan untuk meningkatkan *veracity data.*

5. *Value*

Value merupakan aspek yang paling penting dalam analisis data, yaitu sejauh mana data dapat memberikan manfaat atau *insight* yang berguna. Dalam konteks data deret waktu, nilai data biasanya diperoleh melalui proses analisis yang menghasilkan informasi seperti prediksi masa depan, deteksi anomali, atau identifikasi pola tertentu. Namun, tidak semua data memiliki nilai yang sama.

Dalam banyak kasus, sebagian besar data yang dikumpulkan mungkin tidak relevan atau tidak memberikan kontribusi

Daftar Pustaka

Box, G.E.P. *et al.* (2015). *Time Series Analysis: Forecasting And Control*. *5th ed.* New Jersey: Wiley.

Breiman, L. (2001). Random Forests, *Machine Learning*, 45(1), pp. 5–32. Available At: https://doi.org/10.1201/9780429469275-8.

Fatmawati, D. *et al.* (2023). Klasifikasi Tingkat Kepuasan Penggunaan Layanan Teknologi Informasi Menggunakan Decision Tree, *KLIK: Kajian Ilmiah Informatika dan Komputer*, 3(6), pp. 1056–1062. Available At: https://doi.org/10.30865/klik.v3i6.803.

Goodfellow, I., Bengio, Y. And Courville, A. (2016). *Deep Learning (Adaptive Computation And Machine Learning Series)*. The MIT Press.

Hadi, I. *et al.* (2025). A Comparative Study of Machine Learning With Statistical Feature Selection For Risk Detection of Diabetic, *Jurnal Ilmiah FIFO*, 17(2), pp. 102–118. Available at: https://doi.org/10.22441/fifo.2025.v17i2.001.

Han, J., Kamber, M. and Pei, J. (2012). *Data Mining Concepts And Techniques*. Elsevier Inc.

Hyndman, R.J. And Athanasopoulos, G. (2013). *Forecasting: Principles And Practice*. *3rd ed.* OTexts.

Jerome H. Friedman (2001). Greedy Function Approximation: A Gradient Boosting Machine, *Project Euclid*, 29(5). Available At: https://doi.org/10.1214/aos/1013203451.

Jumaryadi, Y. *et al.* (2025). Machine Learning Approaches to Sentiment Analysis of Mental Health Discussions on Platform X," *PIKSEL Penelitian Ilmu Komputer Sistem Embedded and Logic*, 13(2), pp. 235–246. Available at: https://doi.org/10.33558/piksel.v13i2.11350.

Mayer-Schönberger, V. and Cukier, K. (2014). *Big Data: A Revolution That Will Transform How We Live, Work, And Think*. Boston: Harper Business.

Prayitno, E., Perdana, I.J. and Nasyuha, A.H. (2025). Implementasi Data Mining dan Machine Learning Untuk Segmentasi Pelanggan:

Pendekatan Hybrid Menggunakan Big Data, *Jurnal Ilmiah FIFO*, 17(1), p. 57. Available At: https://doi.org/10.22441/fifo.2025.v17i1.007.

Priambodo, B. *et al.* (2019). Predicting GDP of Indonesia Using K-Nearest Neighbour Regression, *In Journal of Physics: Conference Series*. Institute of Physics Publishing. Available At: https://doi.org/10.1088/1742-6596/1339/1/012040.

Salamah, U. (2022). Prediksi Rating Film Menggunakan Bayesian Regressor dan Gradient Boosting Regressor, *JSAI (Journal Scientific and Applied Informatics)*, 5(3), pp. 203–208. Available At: https://doi.org/10.36085/jsai.v5i3.3614.

Salamah, U. *et al.* (2026). Multimodal Transfer Learning for Anti-Inflammatory Medicinal Plant Leaf Classification Using ResNet50," *PIKSEL: Penelitian Ilmu Komputer Sistem Embedded and Logic*, 14(1), pp. 131–140. Available At: https://doi.org/10.33558/piksel.v14i1.12279.

Usama, M., Liu, M. and Chen, M. (2017). Job Schedulers For Big Data Processing In Hadoop Environment: Testing Real-Life Schedulers Using Benchmark Programs, *Digital Communications and Networks*, 3(4), pp. 260–273. Available at: https://doi.org/10.1016/j.dcan.2017.07.008.

Vaswani, A. *et al.* (2017). Attention Is All You Need," In *31st Conference on Neural Information Processing Systems (NIPS 2017)*. California.

White, T. (2015) *Hadoop: The Definitive Guide*. *4th ed*. O'Reilly Media.

PROFIL PENULIS

Umniy Salamah, S.T., MMSI.

Umniy Salamah, S.T., MMSI. Dilahirkan di Jakarta pada tanggal 6 September 1981. Pendidikan Sarjana (S1) diselesaikan pada tahun 2005 di Universitas Gunadarma, Program Studi Teknik Informatika. Selanjutnya, pendidikan Magister (S2) diselesaikan pada tahun 2011 di Universitas Gunadarma, Program Studi Magister Manajemen Sistem Informasi. Semasa kuliah S1 penulis aktif menjadi asisten Laboratorium Sistem Informasi di Universitas Gunadarma. Setelah lulus S1 penulis pernah bekerja di perusahaan konsultan IT di Jakarta sebagai Sistem Analis hingga tahun 2008.

Sejak tahun 2009 hingga saat ini penulis merupakan Dosen Tetap pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana. Dalam menjalankan tugas tridharma perguruan tinggi, penulis aktif dalam kegiatan pengajaran, penelitian, dan pengabdian kepada masyarakat, khususnya pada bidang Teknologi Informasi berbasis Web, pengembangan perangkat lunak dan *Machine Learning.* Penulis dapat dihubungi melalui email: umniy.salamah@mercubuana.ac.id.

BAB 9
SUPERVISED LEARNING: *CLASSIFICATION* DAN *REGRESSION* PADA *BIG DATA*

Ir. Fauzi Nur Iman, S.Kom., M.Kom.
Universitas Mercu Buana

Pendahuluan

Perkembangan teknologi *digital* dalam beberapa dekade terakhir telah menghasilkan data dalam jumlah yang sangat besar dari berbagai sumber. Fenomena ini dikenal sebagai *big data*, yang merujuk pada kumpulan data yang sangat besar, kompleks, dan beragam, sehingga sulit diproses dan dianalisis menggunakan metode konvensional. Data ini terus berkembang baik dari segi jumlah maupun variasinya, serta berasal dari berbagai sumber seperti media sosial, perangkat IoT, data transaksi, data mesin, informasi geospasial, dan *dataset* publik (Shahnawaz & Kumar, 2025).

Konsep *big data* awalnya diperkenalkan oleh Gartner (Janvrin & Weidenmier Watson, 2017) melalui pendekatan 3V, yaitu: Volume merupakan jumlah data yang sangat besar (*terabyte* hingga *petabyte*), *Velocity*: kecepatan data dihasilkan dan diproses secara *real-time, Variety*: keberagaman jenis data (teks, gambar, video, sensor, dll.). Seiring perkembangan teknologi dan kompleksitas data, konsep ini diperluas menjadi 7V, dengan penambahan empat karakteristik baru (González García & Álvarez-Fernández, 2022): *Veracity*: tingkat keakuratan dan kepercayaan data, *Value*: nilai atau manfaat yang dapat diekstraksi dari data, *Visualization*: kemampuan menyajikan data dalam bentuk visual yang mudah dipahami, *Variability*: inkonsistensi atau perubahan pola data dari waktu ke waktu.

Dengan demikian, paradigma *big data* tidak hanya berfokus pada ukuran dan kecepatan data, tetapi juga pada kualitas, nilai analisis, serta kemampuan interpretasi data dalam pengambilan keputusan. Untuk mengelola dan menganalisis data dalam skala besar tersebut, diperlukan pendekatan yang mampu bekerja secara otomatis dan efisien. *Machine Learning* adalah bagian dari kecerdasan buatan *Artificial Intelligence* (Junaidi *et al.,* 2024). Pendekatan ini sangat relevan dalam konteks *big data* karena mampu menangani kompleksitas dan volume data yang tinggi.

Machine learning memberikan kemampuan kepada komputer untuk belajar tanpa harus diprogram secara eksplisit untuk tugas tertentu (Lindholm *et al*., 2019). Salah satu metode utama dalam *machine learning* adalah *supervised learning*, yaitu metode pembelajaran yang menggunakan data berlabel sebagai dasar

pembentukan (Retnoningsih & Pramudita, 2020). Dalam *supervised learning,* terdapat dua jenis permasalahan utama, yaitu *classification* dan *regression* (Dinata & Hasdyna, 2025; Junaidi *et al.,* 2024). Bab ini akan membahas kedua pendekatan tersebut secara komprehensif, mulai dari konsep dasar hingga penerapannya dalam *big data.*

Konsep Dasar *Supervised Learning*

1. Pengertian *Supervised Learning*

Supervised learning merupakan teknik dalam machine learning yang menggunakan data berlabel untuk melatih model agar dapat mempelajari hubungan antara *input* dan *output* (Nurhalizah *et al.,* 2024). Dalam metode ini, setiap data latih memiliki pasangan berupa fitur (*input*) dan label (*output*), sehingga model dapat belajar dari contoh yang sudah diketahui hasilnya. Tujuan utama dari *supervised learning* adalah melatih model agar dapat memprediksi hasil dari data baru. Data terdiri dari pasangan *input* (fitur) dan *output* (label), dan model belajar menemukan hubungan di antara keduanya agar mampu melakukan prediksi dengan baik (Dinata & Hasdyna, 2025) Hal ini sangat penting dalam aplikasi nyata, seperti prediksi penjualan, diagnosis penyakit, dan sistem rekomendasi.

2. Komponen Utama

Dalam *supervised learning*, terdapat beberapa komponen penting yang saling berkaitan. Dataset merupakan komponen utama yang terdiri dari data latih dan data uji. Data latih digunakan untuk membangun model, sedangkan data uji digunakan untuk mengevaluasi performa model setelah proses pelatihan selesai. Selain itu, fitur (*features*) merupakan variabel input yang digunakan untuk melakukan prediksi, sedangkan label (target) adalah nilai yang ingin diprediksi oleh model. Model itu sendiri merupakan algoritma yang digunakan untuk mempelajari pola dari data. Pemilihan fitur dan model yang tepat sangat mempengaruhi kualitas hasil prediksi.

3. Tahapan Proses *Supervised Learning*

Proses *supervised learning* umumnya terdiri dari beberapa tahapan yang sistematis. Tahap pertama adalah pengumpulan data, yang dapat berasal dari berbagai sumber. Selanjutnya, data tersebut diproses melalui tahap *preprocessing* untuk membersihkan dan mempersiapkan data agar sesuai untuk pelatihan model. Tahap berikutnya adalah pelatihan model, di mana algoritma digunakan untuk mempelajari pola dari data latih. Setelah itu, model dievaluasi menggunakan data uji untuk mengukur performanya. Dalam konteks *big data*, tahapan ini menjadi lebih kompleks karena melibatkan data dalam jumlah besar dan beragam (García *et al*., 2016; Guan *et al.,* 2017; Luengo *et al.,* 2019).

Classification Dalam *Big Data*

1. Pengertian *Classification*

Classification adalah teknik dalam *supervised learning* yang digunakan untuk mengelompokkan data ke dalam kategori atau kelas tertentu berdasarkan fitur yang dimiliki. *Output* dari *classification* bersifat diskrit (Dinata & Hasdyna, 2025), misalnya "positif" atau "negatif", "*spam*" atau "tidak *spam*". Metode *classification* banyak digunakan dalam berbagai bidang karena kemampuannya dalam menangani berbagai jenis permasalahan. Contohnya adalah deteksi email spam, klasifikasi penyakit, serta analisis sentimen terhadap suatu produk atau layanan.

2. Algoritma *Classification*

Berbagai algoritma telah dikembangkan untuk menyelesaikan permasalahan *classification*. Beberapa algoritma umum dalam klasifikasi supervised learning meliputi (Nurhalizah *et al.,* 2024):

a. *K-Nearest Neighbor* (KNN): metode sederhana yang menentukan kelas data baru berdasarkan mayoritas kelas dari tetangga terdekat.
b. *Naïve Bayes*: teknik berbasis probabilitas yang efektif untuk pengambilan keputusan dalam klasifikasi.

Daftar Pustaka

Dinata, R. K., & Hasdyna, N. (2025). *Supervised Learning: Strategi Prediksi dan Klasifikasi Data*. Serasi Media Teknologi.

Dritsas, E., & Trigka, M. (2025). Exploring The Intersection of Machine Learning And Big Data: A Survey. *In Machine Learning and Knowledge Extraction (Vol. 7, Number 1). Multidisciplinary Digital Publishing Institute (MDPI).* https://doi.org/10.3390/make7010013.

Feng, W., Sun, J., Zhang, L., Cao, C., & Yang, Q. (2016*). 2016 IEEE 35th International Performance Computing and Communications Conference. 2016 IEEE 35th International Performance Computing and Communications Conference (IPCCC)*. https://doi.org/10.1109/PCCC.2016.7820655.

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big Data Preprocessing: Methods And Prospects. *Big Data Analytics*, *1*(1). https://doi.org/10.1186/s41044-016-0014-0.

González García, C., & Álvarez-Fernández, E. (2022). What Is (Not) Big Data Based on Its 7Vs Challenges: A Survey. *Big Data And Cognitive Computing*, *6*(4). https://doi.org/10.3390/bdcc6040158.

Guan, Z., Ji, T., Qian, X., Ma, Y., & Hong, X. (2017). A Survey on Big Data Pre-Processing. *2017 5th Intl Conf on Applied Computing And Information Technology/4th Intl Conf on Computational Science/Intelligence and Applied Informatics/2nd Intl Conf on Big Data, Cloud Computing, Data Science (ACIT-CSII-BCD)*, 241–247.

Janvrin, D. J., & Weidenmier Watson, M. (2017). Big Data: A New Twist To Accounting. *Journal of Accounting Education*, *38*, 3–8. https://doi.org/10.1016/j.jaccedu.2016.12.009.

Jomthanachai, S., Wong, W. P., & Khaw, K. W. (2022). An Application of Machine Learning Regression To Feature Selection: A Study of Logistics Performance And Economic Attribute. *Neural Computing And Applications*, *34*(18), 15781–15805. https://doi.org/10.1007/s00521-022-07266-6.

Junaidi, S., Beno, I. S., Farkhan, M., Supartha, I. K. D. G., Pasaribu, A. A.,

Kmurawak, R. M. B., Supiyanto, S., Sroyer, A. M., Reba, F., Fitriyanto, R., & Others. (2024). *Buku Ajar Machine Learning*. PT. Sonpedia Publishing Indonesia.

Lindholm, A., Wahlström, N., Lindsten, F., & Schön, T. B. (2019). *Supervised Machine Learning Lecture Notes For The Statistical Machine Learning Course*. Department of Information Technology, Uppsala University.

Liu, B. (2018). Text Sentiment Analysis Based on CBOW Model And Deep Learning In Big Data Environment. *Journal of Ambient Intelligence And Humanized Computing*, *11*(2), 451–458. https://doi.org/10.1007/s12652-018-1095-6.

Luengo, J., García-Gil, D., Ramírez-Gallego, S., García, S., & Herrera, F. (2019). *Big Data Preprocessing Enabling Smart Data*. Springer Nature Switzerland AG 2020.

Maulud, D., & Abdulazeez, A. M. (2020). A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends*, *1*(2), 140–147. https://doi.org/10.38094/jastt1457.

Nurhalizah, R. S., Ardianto, R., & Purwono, P. (2024). Analisis Supervised dan Unsupervised Learning Pada Machine Learning: Systematic Literature Review. *Jurnal Ilmu Komputer Dan Informatika*, *4*(1), 61–72. https://doi.org/10.54082/jiki.168.

Obi, J. C. (2023). A Comparative Study of Several Classification Metrics And Their Performances on Data. *World Journal of Advanced Engineering Technology and Sciences*, *8*(1), 308–314. https://doi.org/10.30574/wjaets.2023.8.1.0054.

Pérez-Rave, J. I., Correa-Morales, J. C., & González-Echavarría, F. (2019). A Machine Learning Approach To Big Data Regression Analysis of Real Estate Prices For Inferential And Predictive Purposes. *Journal of Property Research*, *36*(1), 59–96. https://doi.org/10.1080/09599916.2019.1587489.

Retnoningsih, E., & Pramudita, R. (2020). Mengenal Machine Learning Dengan Teknik Supervised dan Unsupervised Learning

Menggunakan Python. *Bina Insani ICT Journal*, *7*(2), 156–165. https://www.python.org/.

Saber, A. Y. (2018). *2017 IEEE Symposium Series on Computational Intelligence*. 2017 IEEE Symposium Series on Computational Intelligence (SSCI).

Shahnawaz, M., & Kumar, M. (2025). A Comprehensive Survey on Big Data Analytics: Characteristics, Tools And Techniques. In *ACM Computing Surveys* (Vol. 57, Number 8, pp. 1–33). Association for Computing Machinery. https://doi.org/10.1145/3718364.

Tatachar, A. V. (2021). Comparative Assessment of Regression Models Based on Model Evaluation Metrics. *International Research Journal of Engineering and Technology*. www.irjet.net.

Xing, W., & Bei, Y. (2020). Medical Health Big Data Classification Based on KNN Classification Algorithm. *IEEE Access*, *8*, 28808–28819. https://doi.org/10.1109/ACCESS.2019.2955754

PROFIL PENULIS

Ir. Fauzi Nur Iman, S.Kom., M.Kom.

Ir. Fauzi Nur Iman, S.Kom., M.Kom. dilahirkan di Jember, Jawa Timur pada tanggal 18 Agustus 1989. Pendidikan Sarjana (S1) diselesaikan pada tahun 2014 di Universitas Mercu Buana, Program Studi Informatika. Selanjutnya, pendidikan Magister (S2) diselesaikan pada tahun 2017 di Universitas Budi Luhur, Program Studi Magister Ilmu Komputer. Penulis juga telah menyelesaikan Program Profesi Insinyur di Universitas Mercu Buana pada tahun 2025 sebagai bagian dari penguatan kompetensi profesional.

Saat ini, penulis merupakan Dosen Tetap pada Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana. Dalam menjalankan tugas tridharma perguruan tinggi, penulis aktif dalam kegiatan pengajaran, penelitian, dan pengabdian kepada masyarakat, khususnya pada bidang pengembangan perangkat lunak, sistem informasi, dan teknologi berbasis web. Penulis dapat dihubungi melalui email: fauzi@mercubuana.ac.id.

BAB 10
ENSEMBLE METHODS DAN *MODEL SELECTION STRATEGIES*

Dr. Rusdah, S.Kom., M.Kom.
Universitas Budi Luhur

Pendahuluan

Dalam ekosistem *big data,* paradigma pemodelan prediktif telah bergeser dari pencarian "model tunggal terbaik" menuju penggabungan berbagai model untuk mencapai stabilitas dan akurasi yang lebih tinggi. Fenomena *Volume, Velocity,* dan *Variety* menuntut model yang tidak hanya cerdas secara statistik, tetapi juga tangguh terhadap *noise* dan variansi data yang ekstrem.

Masalah utama dalam *Data Mining* skala besar adalah risiko *overfitting* pada model yang terlalu kompleks atau *underfitting* pada model yang terlalu sederhana. *Ensemble Methods* muncul sebagai solusi teknis yang mengadopsi prinsip *"Wisdom of the Crowd"* yaitu kombinasi beberapa prediksi dapat mengurangi error dan meningkatkan generalisasi.

Dalam dunia *machine learning*, *Ensemble Method* adalah teknik yang menggabungkan beberapa model (sering disebut *base learners* atau *weak learners*) untuk menghasilkan satu model prediksi yang lebih kuat dan akurat. Seperti meminta pendapat dari sekelompok ahli daripada hanya satu orang, *ensemble method* bertujuan untuk mengurangi kesalahan seperti *bias* (salah arah) dan *variance* (terlalu sensitif pada data latih/*overfitting*). Sementara itu, model *selection strategies* berperan dalam memilih model terbaik atau kombinasi model yang optimal berdasarkan kriteria tertentu seperti akurasi, kompleksitas, dan generalisasi.

Konsep Dasar *Ensemble Method*

Ensemble learning adalah pendekatan yang menggabungkan beberapa model untuk menghasilkan prediksi yang lebih baik dibandingkan dengan model individu. Prinsip utama dari metode ini adalah *wisdom of the crowd*, yaitu kombinasi beberapa prediksi dapat mengurangi *error* dan meningkatkan generalisasi. Secara matematis, model *ensemble* dapat direpresentasikan sebagai:

$$\hat{y} = \sum_{i=1}^{M} w_i f_i(x)$$

Dimana:
$f_i(x)$ adalah model ke-i.
w_i adalah bobot model ke-I.
M adalah jumlah model dalam *ensemble.*
$\hat{y}$ adalah *output* prediksi akhir

Jenis Utama *Ensemble Method*

Secara garis besar, teknik ini dibagi menjadi tiga kategori utama berdasarkan cara model-model tersebut dikombinasikan:

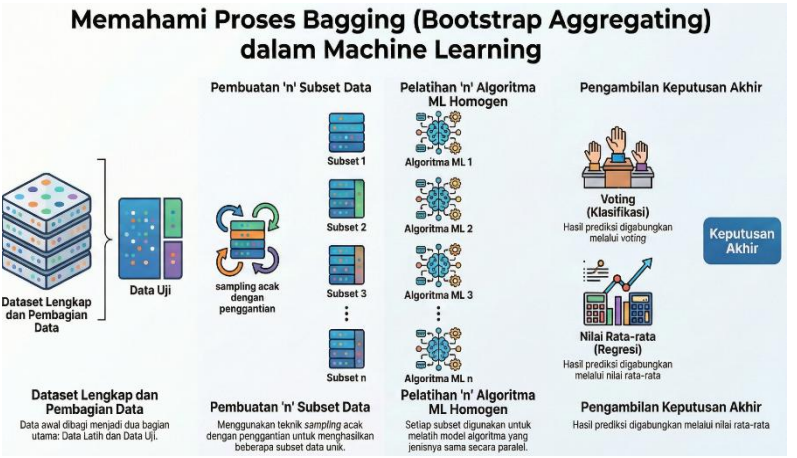
1. ***Bagging* (*Bootstrap Aggregating*)**

   Bagging atau *Bootstrap Aggregation* secara resmi diperkenalkan oleh Leo Breiman (Breiman, 1996). Teknik ini mengurangi kesalahan pembelajaran melalui implementasi serangkaian algoritma pembelajaran mesin yang homogen secara paralel pada *subset data* yang berbeda (diambil secara acak dengan penggantian/*replacement*). Hasil akhirnya ditentukan melalui pemungutan suara terbanyak (*voting*) untuk klasifikasi atau rata-rata untuk regresi. Tujuan utama teknik bagging adalah untuk mengurangi *variance* (mencegah *overfitting*).

   a. Konsep

      Bagging bekerja dengan cara membuat beberapa *subset data* dari *dataset* pelatihan secara acak dengan pengembalian (*bootstrap sampling*). Setiap subset digunakan untuk melatih satu model secara paralel. Hasil akhirnya diambil melalui *voting* (klasifikasi) atau rata-rata (regresi). Dua komponen utama teknik *bagging* adalah: pengambilan sampel acak dengan penggantian (*bootstraping*) dan serangkaian algoritma *Machine Learning* (ML) yang homogen (*ensemble learning*).

      Proses bagging cukup mudah dipahami (Han *et al.,* 2023). Pertama-tama diekstrak "n" *subset* dari set pelatihan, kemudian subset ini digunakan untuk melatih "n" *base learner* dengan tipe yang sama. Untuk membuat prediksi, masing-masing dari "n" *base learner* diberi sampel uji, *output* dari setiap *base learner* dirata-ratakan (dalam kasus regresi) atau divoting (dalam kasus klasifikasi). Gambar 10.1 menunjukkan gambaran umum arsitektur *bagging*.



Gambar 10.1: *Bagging*

Sumber: Modifikasi Oleh Penulis Dari Berbagai Sumber.

Jumlah *subset* serta jumlah item per *subset* akan ditentukan oleh sifat masalah ML yang anda pilih. Menurut Breiman (1996), untuk masalah klasifikasi diperlukan lebih banyak subset dibandingkan dengan masalah regresi. Salah satu keunggulan utama *bagging* adalah dapat dieksekusi secara paralel karena tidak ada ketergantungan antar estimator. Untuk *dataset* kecil, beberapa estimator sudah cukup, sedangkan *dataset* yang lebih besar mungkin memerlukan lebih banyak estimator.

b. Algoritma Populer

Random forest adalah metode *ensemble* yang menggunakan pendekatan *bagging* (*Bootstrap Aggregating*). Algoritma ini membangun beberapa *decision tree* pada *subset* acak dari data, lalu menggabungkan hasil prediksi dari semua pohon tersebut. Setiap pohon keputusan dilatih pada data yang berbeda, yang dihasilkan melalui teknik *bootstrap*. Hasil akhir dari *random forest* merupakan *voting majority* (untuk klasifikasi) atau rata-rata (untuk regresi) dari prediksi setiap pohon. *Bagged Decision Trees*: versi standar di mana setiap pohon menggunakan seluruh fitur yang tersedia pada sampel data yang berbeda.

Daftar Pustaka

Annisa, M., & Rusdah. (2022). Prediction of Non-Performing Loans for Credit Application Analysis of Rural Bank Using Random Forest. *2022 9th International Conference on Electrical Engineering, Computer Science And Informatics (EECSI)*, 111–114. https://doi.org/10.23919/EECSI56542.2022.9946628.

Ayaz, M., Shaukat, F., & Raja, G. (2021). Ensemble Learning Based Automatic Detection of Tuberculosis In Chest X-Ray Images Using Hybrid Feature Descriptors. *Physical and Engineering Sciences in Medicine*, *44*(1), 183–194. https://doi.org/10.1007/s13246-020-00966-0.

Bhakti, D. S., Prasetyo, A., & Arsi, P. (2024). Implementation of Hyperparameter Tuning In Random Forest Algorithm For Loan Approval Prediction. *Jurnal Teknik Informatika (Jutif)*, *5*(4), 63–69. https://doi.org/10.52436/1.jutif.2024.5.4.2032.

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655.

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning And An Application to Boosting. *Journal of Computer and System Sciences*, *55*, 119–139.

Han, J., Pei, J., & Tong, H. (2023). *Data Mining: Concepts And Techniques*.

Hazizah, N., Sharipuddin, & Feranika, A. (2025). Implementasi Algoritma Random Forest Dalam Klasifikasi Risiko Gagal Bayar Kartu Kredit Pada Nasabah Bank. *Jurnal Manajemen Teknologi dan Sistem Informasi (JMS)*, *5*(1), 1050–1059. https://doi.org/10.33998/jms.v5i1.

Mawarni, A. C., Rusdah, R., Hin, L. L., & Anubhakti, D. (2023). Deteksi Dini Gejala Awal Penyakit Diabetes Menggunakan Algoritma Random Forest. *IDEALIS: InDonEsiA JournaL Information System*, *6*(2), 165–171. https://doi.org/10.36080/idealis.v6i2.3018.

Muhajir, M., & Widiastuti, J. (2022). Random Forest Method to Customer Classification Based on Non-Performing Loan in Micro Business. *Jurnal Online Informatika*, *7*(2), 177–183. https://doi.org/10.15575/join.v7i2.842.

Osamor, V. C., & Okezie, A. F. (2021). Enhancing The Weighted Voting Ensemble Algorithm For Tuberculosis Predictive Diagnosis. *Scientific Reports*, *11*(1), 1–11. https://doi.org/10.1038/s41598-021-94347-6.

Pebrianti, D., Istinabiyah, D. D., Bayuaji, L., & Rusdah. (2022). Hybrid Method for Churn Prediction Model In The Case of Telecommunication Companies. *2022 9th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 161–166. https://doi.org/10.23919/EECSI56542.2022.9946535.

Rusdah, Bregastantyo, B. A., & Riwurohi, J. E. (2023). Prognosis Model of The Treatment Period of Tuberculosis Patients With Medication Compliance Parameters Using The Gradient Boosting Algorithm. *2023 6th International Seminar on Research of Information Technology And Intelligent Systems (ISRITI)*, 220–225. https://doi.org/10.1109/ISRITI60336.2023.10467254.

Rusdah, R., Painem, P., & Kusumaningsih, D. (2025). Enhancing Prediction of Treatment Duration In New Tuberculosis Cases: A Comprehensive Approach With Ensemble Methods And Medication Adherence. *Jurnal Teknik Informatika (JUTIF)*, *6*(2).

Rusdah, Winarko, E., & Wardoyo, R. (2015). Preliminary Diagnosis of Pulmonary Tuberculosis Using Ensemble Method. *The 2nd International Conference on Data and Software Engineering (ICoDSE) 2015*.

Saleh, A., Mukhtar, R., & Rusdah, R. (2025). Early Detection of Dengue Hemorrhagic Fever Using Patient Medical Data With Ensemble Learning Methods. *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)*, *8*(3), 635–645. https://doi.org/10.24014/ijaidm.v8i3.38088.

Schapire, R. E. (2003). The Boosting Approach to Machine Learning: An Overview. In D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, & B. Yu (Eds.), *Nonlinear Estimation and Classification. Lecture Notes In Statistics* (Vol. 171, pp. 149–171). Springer. https://doi.org/10.1007/978-0-387-21579-2_9.

PROFIL PENULIS

Dr. Rusdah, S.Kom, M.Kom.

Penulis menyelesaikan pendidikan S1 di jurusan Sistem Informasi, program studi Komputerisasi Akuntansi pada Universitas Budi Luhur Jakarta pada tahun 2001. Penulis melanjutkan pendidikan S2 Ilmu Komputer pada universitas yang sama dan lulus tahun 2006. Kemudian penulis menyelesaikan pendidikan pada program doktoral S3 Ilmu Komputer di Universitas Gadjah Mada dan lulus pada tahun 2018. Saat ini, penulis menjadi dosen tetap pada program studi S2 Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Budi Luhur.

Selain sebagai dosen dengan pengalaman mengajar lebih dari 25 tahun, penulis juga aktif sebagai *reviewer* jurnal, asesor BKD Nasional, *reviewer* penelitian, dan narasumber pada pertemuan ilmiah (seperti lomba, pelatihan dan seminar) baik di tingkat nasional maupun internasional. Penulis memiliki kepakaran di bidang *Data Science, Data Mining, Machine Learning,* dan *Decision Support Systems.* Sebagai dosen profesional yang sudah tersertifikasi, penulis pun aktif sebagai peneliti di bidang kepakarannya tersebut dengan mendapatkan pendanaan baik internal perguruan tinggi maupun pendanaan dari Kemenristek DIKTI.

Email Penulis: rusdah@budiluhur.ac.id.

BAB 11
TEXT MINING DAN *NATURAL LANGUAGE PROCESSING*

Lukman Hakim, S.T., M.Kom.
Universitas Mercu Buana

Pendahuluan

Perkembangan teknologi informasi telah menciptakan banyak sekali data dalam berbagai bentuk, termasuk data berupa teks. Teks dapat berasal dari berbagai sumber seperti artikel berita, media sosial, dokumen ilmu, laporan organisasi, email, maupun ulasan pelanggan yang terdapat di berbagai *platform digital*. Sebagian besar data tersebut tidak memiliki bentuk yang teratur, sehingga diperlukan cara khusus untuk mengambil informasi yang terkandung di dalamnya.

Istilah *Natural Language Processing* yang merupakan cabang kecerdasan buatan yang menggunakan teknik membaca teks dengan analisis ucapan atau tulisan (Danar *et al.,* 2024). Berkembangnya model bahasa besar atau *Large Language Model* (LLM) meningkatkan kemampuan multibahasa serta kemampuan menjawab pertanyaan, dengan mengintegrasikan data multimodal seperti teks, suara, dan visual, NLP semakin kompleksitas (Zhang & Iglesias, 2025).

NLP menjadi bagian kehidupan manusia dalam bermacam-macam aktivitas seperti *chatbot*, mesin pencarian, GPS dengan perintah suara, banyak lainnya (Stryler Cole, 2026). Dalam dunia pengolahan data saat ini, dua bidang yang sangat berguna untuk menganalisis teks adalah *Text Mining* dan *Natural Language Processing* (NLP). Kedua bidang tersebut termasuk dalam perkembangan kecerdasan buatan yang bertujuan membantu komputer memahami bahasa manusia dengan lebih baik. Teknologi pengolahan bahasa alat penemuan pola dan informasi yang berguna dari data teks yang berjumlah besar. Sementara itu, NLP adalah teknologi yang memungkinkan komputer memahami, memproses, dan membuat bahasa alami yang digunakan manusia.

Dengan menggabungkan kedua pendekatan ini, berbagai aplikasi cerdas dapat dikembangkan, seperti analisis sentimen, sistem rekomendasi, *chatbot*, sistem penerjemah bahasa, hingga sistem pencarian informasi yang lebih cerdas. Di masa kini yang disebut era *big data,* kemampuan mengolah dan menganalisis informasi berupa teks semakin dibutuhkan, karena sebagian besar data yang ada di internet berupa teks. Maka itu, text mining dan NLP menjadi bidang penelitian yang semakin berkembang dalam ilmu komputer dan *data science.*

Konsep *Text Mining*

Text mining adalah sebagai proses atau cara memanfaatkan informasi yang berguna dari banyak dokumen teks dengan menggunakan teknik-teknik seperti analisis data, statistik, dan pembelajaran mesin (Dian Prawira, Nurul Mutiah, 20224). Menurut Feldman dan Sanger, text mining adalah cara otomatis atau hampir otomatis yang digunakan untuk mencari pola yang bermakna dari sekumpulan data teks yang sangat besar.

Berbeda dengan data berbentuk angka yang memiliki pola yang teratur, teks memiliki sifat yang lebih rumit karena mencakup berbagai aspek linguistik seperti tata bahasa, makna, dan situasi yang mengelilinginya. Pengolahan data teks biasanya melibatkan beberapa tahap utama untuk mengubah teks menjadi bentuk yang dapat diproses oleh komputer. Tujuan utama text mining antara lain:

1. Mengidentifikasi pola dalam data teks.
2. Mengklasifikasikan dokumen berdasarkan kategori tertentu.
3. Menemukan topik utama dalam kumpulan dokumen.
4. Menganalisis opini atau sentimen dalam teks.
5. Mendukung proses pengambilan keputusan berbasis data.

Pengertian *text mining: text mining*, atau penambangan teks, adalah proses pengambilan informasi yang terkandung dalam teks dengan menggunakan teknik komputasi dan analisis data. Teknik ini memungkinkan pengguna untuk menemukan pola, tendensi, dan informasi yang tersembunyi dalam dokumen teks yang besar, berbagai bentuk. *Text mining* digunakan untuk menjawab berbagai kebutuhan seperti analisis pasar, pemrosesan data kesehatan, pendidikan, manajemen pemerintah, maupun penelitian ilmiah.

Konsep Dasar *Natural Language Processing*

Pemrosesan Bahasa Alami (NLP) adalah bidang dalam kecerdasan buatan yang mempelajari cara komputer dapat memahami dan memproses bahasa manusia secara otomatis. Menurut Jurafsky dan Martin, NLP merupakan bidang ilmu yang menggabungkan beberapa disiplin seperti ilmu komputer, linguistik, dan *machine learning*, sehingga komputer dapat berinteraksi dengan bahasa manusia (Jurafsky, Daniel, 2026).

NLP memiliki tujuan terkait teknologi ini diciptakan yaitu sebagai fasilitas komunikasi antara manusia dengan mesin dengan menggunakan bahasa alami atau natural (Widiantoro *et al*., 2024). Bahasa manusia memiliki ciri yang kompleks dan seringkali memiliki sifat ambigu. Sebuah kata bisa punya arti yang berbeda-beda tergantung bagaimana kalimat itu dipakai. Berikut adalah ulasan ulang dari kalimat tersebut:

Selain itu, struktur kalimat juga dapat berbeda-beda meskipun memiliki arti yang sama. Tantangan ini menjadikan pengembangan sistem NLP sebagai salah satu topik penelitian yang menarik dalam bidang kecerdasan buatan. Beberapa tugas utama dalam NLP antara lain:

1. Tokenisasi adalah proses membagi teks menjadi bagian-bagian seperti kata atau kalimat.
2. Penandaan kategori kata, yaitu mengenali jenis tata bahasa suatu kata.
3. Pengenalan Entitas Nama (*Named Entity Recognition*) adalah proses mengenali entitas seperti nama orang, lokasi, dan organisasi.
4. Menerjemahkan teks dari satu bahasa ke bahasa lainnya dengan menggunakan mesin.
5. *Text Summarization*, yaitu merangkum dokumen secara otomatis. Perkembangan teknologi *deep learning* telah meningkatkan kemampuan sistem pemrosesan bahasa alami (NLP) dalam memahami konteks dan makna bahasa secara lebih baik.

Pendekatan Terhadap NLP

NLP menggabungkan kekuatan linguistik komputasional dengan algoritma pembelajaran mesin (*Machine Learning*) dengan kemampuan menggunakan ilmu data (*data Science*) dalam menganalisis bahasa yang diucapkan, kemampuan sintaksis dan analisis *semantic* yang menafsirkan kata-kata atau ucapan menjadi makna dalam struktur kalimat. NLP mendukung pembelajaran mandiri (*Self supervised learning*) yang membutuhkan banyak dataset dalam melatih model AI (Stryler Cole, 2026). Ada 3 pendekatan dalam NLP yaitu:

Daftar Pustaka

Akritidis, L., & Bozanis, P. (2025). Machine Learning Advances and Applications on Natural Language Processing (NLP). *In Electronics (Switzerland)* (Vol. 14, Issue 16). https://doi.org/10.3390/electronics14163282.

Danar, R., Kom, D. M., Kom, M. M., Bahtiar, A., Kom, M., & Ali, I. (2024). Dasar Dasar Natural Language Processing (NLP). *Dasar Dasar Natural Language Processing (NLP)*, xii–78. https://repository.minhajpustaka.id/media/publications/593976-dasar-dasar-natural-language-processing-6e7ee8eb.pdf.

Dian Prawira, Nurul Mutiah, I. R. (20224). *Pengantar Text Mining Dan Implementasinya*. KAPI.

Hakim, L., Dalimunthe, M. V., Danuputri, C., & Widyaningrum, D. (2024). Sentimen Analisis Mengenai Polusi Udara Menggunakan Algoritma Support Vector Machine dan Random Forest. *Jurnal Ilmiah FIFO*, *15*(2), 91. https://doi.org/10.22441/fifo.2023.v15i2.001.

Jurafsky, Daniel, J. H. M. (2026). *Speech And Language Processing* (Stanford University (ed.); 3rd ed.).

Naf'an, E., Islami, F., & Gushelmi, G. (2022). *Dasar-Dasar Deep Learning dan Contoh Aplikasinya* (Rendi Fernandes (ed.)). CV. Mitra Cendekia Media. http://repository.upiyptk.ac.id/9683/1/Buku%2B1%2BDasar-Dasar%2BDeep%2BLearning.pdf.

Stryler Cole, J. H. (2026, March 31). What is NLP? *IBM*. https://www.ibm.com/think/topics/natural-language-processing.

Widiantoro, A., Dwiyoga Mustafid Sanjaya, & Ridwan. (2024). *Pengantar NLP Dan Topik Model*. https://repository.unika.ac.id/36457/1/buku pengantar NLP LDA-Based Topic Modeling labeling.pdf.

Wijaya, H., & Hakim, L. (2023). Analisis Sentimen Kebijakan Jaminan Hari Tua (JHT) Pada Twitter Menggunakan Naïve Bayes. *Jurnal Algoritma, Logika Dan Komputasi*, *6*(1), 509–518.

https://doi.org/10.30813/j-alu.v6i1.3552.

Zhang, J., & Iglesias, C. (2025). Special Issue on Recent Applications of Machine Learning in Natural Language Processing (NLP). *In Applied Sciences (Switzerland)* (Vol. 15, Issue 11). https://doi.org/10.3390/app15116110.

PROFIL PENULIS

Lukman Hakim, S.T., M.Kom.
Lahir di Tangerang, Kab Tangerang, 27 Oktober 1977. Jenjang Pendidikan S1 Teknik Komputer ditempuh di Universitas Yarsi, Kota Jakarta lulus tahun 2001. Pendidikan S2 Ilmu Komputer Indonesia, lulus tahun 2004 di Universitas Bunda Mulia. Saat ini menjabat sebagai Kepala Laboratorium Fakultas Ilmu Komputer di Universitas Mercu Buana. Penulis merupakan aktif sebagai *reviewer*, editor jurnal, penulis dan peneliti.
Ketertarikan penulis pada bidang rekayasa perangkat lunak dan *Data Science* mengajar Pemrograman Berorientasi Objek, Rekayasa Perangkat Lunak, *Machine Learning, Data Mining*, Proses *Mining*. Pengalaman mengajar di berbagai perguruan tinggi seperti Universitas Mercu Buana Jakarta, Universitas Bunda Mulia, Universitas Esa Unggul, Universitas Bina Nusantara, Kalbis. Email: lhakim2710@gmail.com atau lukman_hakim@mercubuana.ac.id. WA: 081314410170.

BAB 12
STUDI KASUS: IMPLEMENTASI *BIG DATA ANALYTICS* DALAM BERBAGAI SEKTOR

Ida Farida, S.T., M.Kom.
Universitas Mercu Buana

Pendahuluan

Big data memiliki peranan yang sangat penting dalam berbagai bidang, karena kemampuannya untuk menyediakan wawasan dan analisis yang lebih mendalam dari volume data yang sangat besar. Dengan teknologi yang ada saat ini, *big data* memungkinkan organisasi di berbagai sektor untuk meningkatkan efisiensi, inovasi, dan pengambilan keputusan berbasis data. *Big data* terus berkontribusi pada pengembangan teknologi dan otomatisasi, memungkinkan organisasi untuk meningkatkan efisiensi, responsivitas, dan akurasi dalam pengambilan keputusan strategis (Arifulsyah *et al.,* 2023).

Sesuai namanya, *big data* didefinisikan sebagai kumpulan data yang berukuran sangat besar. Ukuran big data bisa sebesar *terabyte* bahkan *petabyte*. Menurut Gartner (Janvrin & Weidenmier Watson, 2017) *big data* didefinisikan sebagai kumpulan data yang dilihat melalui pendekatan 3V, yaitu: skala (*Volume*), distribusi (*velocity*) dan keragaman (keragaman). Dalam perkembangannya, konsep *big data* terdapat beberapa V tambahan yang menjadi bagian integral dari karakteristik Big data (González García & Álvarez-Fernández, 2022), yaitu *Variability, Veracity, Value,* dan *Visualization.*

Dengan begitu banyak karakteristik yang berkontribusi pada konsep *big data,* hal ini menjelaskan bahwa *big data* bukan sekedar volume data yang besar dan beragam, akan tetapi juga melibatkan berbagai aspek terutama pada nilai analisis data yang diperlukan untuk pengambilan keputusan, disinilah peran *big data analytics* sangat diperlukan. *Big data analytics* digunakan untuk menemukan pola tersembunyi, tren pasar, dan preferensi konsumen untuk kepentingan pengambilan keputusan suatu perusahaan.

Menurut (Morabito, 2015), *big data analytics* memungkinkan organisasi untuk mengidentifikasi pola dan tren tersembunyi dalam data yang besar dan kompleks, sehingga mendukung pengambilan keputusan yang lebih baik. Didukung dengan kemampuan menganalisis data secara *real-time, big data* membantu perusahaan merespons perubahan pasar dengan cepat dan efektif (Budiarto *et al.,* 2024). Studi literatur menunjukkan bahwa penerapan teknologi *big data* dalam analisis data bisnis memberikan wawasan mendalam terkait perannya dalam berbagai sektor, seperti pemerintahan, bisnis, kesehatan, dan pendidikan.

Implementasi *big data analytics* terbukti dapat meningkatkan efisiensi dan inovasi di berbagai sektor dengan mengubah data mentah menjadi sebuah keputusan strategis berbasis data, seperti optimalisasi rantai pasok, analisis prediktif, dan personalisasi layanan. Teknologi ini mempercepat pengambilan keputusan, mengurangi risiko, dan memprediksi tren masa depan.

Konsep Dasar *Big Data Analytic*

1. Pengertian *Big Data Analytic*

Big data analytics merupakan proses penggalian informasi dengan beberapa tahapan proses dari mulai mengumpulkan, mengolah dan menganalisis data dari berbagai jenis kumpulan data yang berukuran besar untuk mendapatkan Informasi (*Insight*) yang berguna dalam pengambilan keputusan yang akurat dan strategis. *Big data analytics* banyak sekali digunakan untuk menemukan pola tersembunyi, tren pasar, dan preferensi konsumen atau masyarakat, serta prediksi masa depan sehingga dapat membantu dalam proses pengambilan keputusan suatu perusahaan berdasarkan data historis sehingga keputusan bisnis yang diambil lebih cerdas dan cepat. *Big data analytics* merupakan solusi inovatif dalam analisis data bisnis, pengambilan keputusan, dan penerapan strategi keuangan dengan cara memproses kumpulan data yang sangat besar dan kompleks (Sunata, O. A. 2025). Selain itu, studi dalam *Journal of Big Data* (Tosi, D. *et al.*, 2024) menjelaskan bahwa *big data analytics* merupakan bagian penting dari ekosistem kecerdasan buatan dan *machine learning*, yang digunakan untuk memahami pola, tren, serta tantangan dari data dalam skala besar.

2. Cara Kerja *Big Data Analytic*

Dalam suatu proses *big data analytics*, terdapat beberapa langkah dan teknologi yang digunakan.

a. Pengumpulan Data (*Data Collection*): data diperoleh dari berbagai sumber, seperti media sosial, transaksi bisnis, sensor IoT, maupun *database* internal perusahaan. Pada proses ini melibatkan integrasi data dalam jumlah besar, data yang

dikumpulkan dapat berupa teks, gambar, video, audio ataupun metadata. Pengumpulan data juga melibatkan pengumpulan data dari sumber terstruktur dan tidak terstruktur (Angin, J. T *et al.,* 2025).

b. Penyimpanan dan Pemrosesan: data kemudian disimpan dalam sistem penyimpanan terdistribusi dan dapat menangani volume besar, seperti data *warehouse* atau *data lake* dan bisa juga disimpan dalam media penyimpanan berbasis *cloud* seperti *Hadoop* atau *cloud storage* agar lebih mudah diakses dan dilakukan analisis (Berisha, B., Mëziu, E., & Shabani, I, 2022). Pada proses ini juga dilakukan pembersihan data agar dapat dianalisis dengan efisien, seperti memperbaiki atau membuang data yang tidak relevan, salah, duplikat, atau tidak lengkap untuk memastikan kualitas data (*veracity*) sebelum dianalisis.
c. Analisis Data (*Data Analyze*): tahap ini merupakan inti dari *big data analytics* dan dapat berlangsung secara berulang (iteratif), khususnya dalam analisis data yang bersifat eksploratif, dimana proses analisis dilakukan kembali hingga ditemukan pola atau hubungan yang relevan. Berbagai metode seperti algoritma analisis statistik, *machine learning*, model prediktif dan kecerdasan buatan digunakan untuk menemukan pola, tren, serta anomali dalam data.
d. Visualisasi (*Data Visualization*) dan *Insight*: hasil analisis disajikan dalam bentuk *dashboard* interaktif atau laporan dalam bentuk grafik, gambar, *chart* atau *trend* yang memudahkan perusahaan dalam mengambil keputusan secara lebih akurat, seperti mengidentifikasi peluang pertumbuhan bisnis, memahami perilaku pelanggan, atau mengoptimalkan operasi perusahaan. Tahapan ini berperan dalam menampilkan makna baru serta menginterpretasikan data dalam skala besar, sehingga memudahkan proses eksplorasi data dan menyederhanakan analisis *big data* yang kompleks (Yoo, K. H., Leung, C. K., & Nasridinov, A, 2022). Proses ini sering kali bersifat siklis (*lifecycle*), di mana hasil analisis dapat memicu pengumpulan data baru untuk meningkatkan akurasi.

Daftar Pustaka

Albores, E. L. I. (2022). Consumption Prediction on Netflix: Audience Tracking Analysis Based on The Recommendation Algorithm In Times of Pandemic. *In Predictive Technology In Social Media* (pp. 52–74). CRC Press.

Alifia, R. A., Safitri, N. R., Irhami, D. M., Hidayat, N. R., & Kusumasari, I. R. (2024). Challenges And Solutions For Decision Making In The Era of Big Data. *Jurnal Bisnis dan Komunikasi Digital*, 2(2), 13.

Angin, J. T. K. P., Purnomo, W. A., Juansa, A., Robet, R., & Pribadi, O. (2025). *Pengantar Big Data: Konsep, Teknologi, Dan Aplikasinya*. Star Digital Publishing.

Berisha, B., Mëziu, E., & Shabani, I. (2022). Big Data Analytics In Cloud Computing: An Overview. *Journal of Cloud Computing*, 11(1), 24.

Chen, Y. (2024). Research on Supply Chain Optimization At Amazon. *In Proceedings of The 3rd International Conference on Financial Technology And Business Analysis*. https://doi.org/10.54254/2754-1169/105/20241983.

Dzakiyyah, B. H., Putri, K. D., Salsabila, N. Y., Rafania, T. A., & Prawira, I. F. A. (2023). Pemanfaatan Big Data Untuk Meningkatkan Kepuasan Pelanggan Shopee. *Innovative: Journal of Social Science Research*, 3(5), 10441–10455.

Ellahi, E., Talha, M., Vidhate, D. A., Mann, G., Chauhan, S., & Singh, V. (2024). Fraud Detection And Prevention In Finance: Leveraging Artificial Intelligence And Big Data. *Dandao Xuebao Journal of Ballistics*, 36(1), 54–62. https://doi.org/10.52783/dxjb.v36.141.

Faizah Ats Tsaniyah, N., Ningsih, S. P., Cyntia, D. Y., Kusumasari, I. R., & Hidayat, N, R. (2024). The Role of Big Data Analytics In Supporting Decision-Making Theories In Companies. *Jurnal Bisnis dan Komunikasi Digital*, 2(2), 10. https://doi.org/10.47134/jbkd.v2i2.3458.

Gong, S. (2024). Implementation of Big Data Techniques In Banking: Evidence From Real-Time Fraud Detection And Customer Segmentation. *In Proceedings of the 3rd International Conference on*

Financial Technology And Business Analysis. https://doi.org/10.54254/2754-1169/134/2024.18598.

González García, C., & Álvarez-Fernández, E. (2022). What Is (Not) Big Data Based On Its 7Vs Challenges: A Survey. *Big Data And Cognitive Computing*, 6(4). https://doi.org/10.3390/bdcc6040158.

Huo, Z. (2024). Data Statistical Analysis on Amazon E-Commerce Platform For Recommender System. *Applied And Computational Engineering*, 51(1), 97–103. https://doi.org/10.54254/2755-2721/51/20241183.

Janvrin, D. J., & Weidenmier Watson, M. (2017). Big Data: A New Twist To Accounting. *Journal of Accounting Education*, 38, 3–8. https://doi.org/10.1016/j.jaccedu.2016.12.009.

Jerendi, C., Ayu, F., Nasution, R., & Sitorus, S. P. (2026). *Kecerdasan Teknologi Big Data Dalam Transformasi Digital*.

Kumar, P., Gowda, D. Y., & Prakash, A. M. (2024). Machine Learning In Cybersecurity: A Comprehensive Survey of Data Breach Detection, Cyber-Attack Prevention, And Fraud Detection. *In Pioneering smart healthcare 5.0 With IoT, Federated Learning, And Cloud Security* (pp. 175–197).

Lestari, D. A., & Nasution, M. I. P. (2025). Pengelolaan Big Data: Inovasi Solusi Dan Tantangan Dalam Era Informasi Modern. *Journal Sains Student Research*, 3(3), 502–511.

Luo, S., & Pan, L. (2026). Fundamentals of Big Data Analysis. *In Big Data Analysis: Theory And Technology* (pp. 1–36). Springer Nature Singapore.

Maharani, A. (2025). Penerapan Big Data Dalam Perencanaan Strategis Dan Pengambilan Keputusan Bisnis. *JMEB Jurnal Manajemen Ekonomi & Bisnis*, 3(01).

Mittal, S. (2024). Big Data Application For Service Operation Application. *In 2024 1st International Conference on Sustainable Computing And Integrated Communication In Changing Landscape of AI (ICSCAI)* (pp. 1–10). IEEE.

Peddinti, S. R., Tanikonda, A., & Katragadda, S. R. (2024). *Deep Learning*

For Anomaly Detection In E-Commerce And Financial Transactions: Enhancing Fraud Prevention And Cybersecurity. SSRN.

Provost, F., & Fawcett, T. (2013). Data Science And Its Relationship To Big Data And Data-Driven Decision Making. *Big Data*, 1(1), 51–59.

Rahul, K., Banyal, R. K., Goswami, P., & Kumar, V. (2020). Machine Learning Algorithms For Big Data Analytics. *In Computational Methods And Data Engineering* (pp. 359–367). Springer Singapore.

Sari, M. M., & Avrianto, R. P. (2025). Big Data Dalam Bisnis: Studi Literatur Dan Penerapannya Di Indonesia. *Jurnal Inovasi Informatika*, 7(2), 25–36.

Setyawan, A. R., Fauzi, A., Yuniati, T., Mutaqin, A. Z., Anggraini, M. P., Mutazam, Z. H., ... & Muis, A. F. K. (2024). Peran Big Data Dalam Intelligence Business Pada Perkembangan E-Commerce. *Sentri: Jurnal Riset Ilmiah*, 3(6), 2728–2740.

Shoman, W., Yeh, S., Sprei, F., Köhler, J., Plötz, P., Todorov, Y., Rantala, S., & Speth, D. (2023). A Review of Big Data In Road Freight Transport Modeling: Gaps And Potentials. *Data Science For Transport*, 5(2). https://doi.org/10.1007/s42421-023-00065-y

Sunata, O. A. (2025). Penerapan Big Data Analytics Dalam Pengambilan Keputusan Bisnis. *Journal SAINS Student Research*, 3(2), 474–480. https://doi.org/10.61722/jssr.v3i2.4339.

Tosi, D., Kokaj, R., & Roccetti, M. (2024). 15 Years of Big Data: A Systematic Literature Review. *Journal of Big Data*, 11, 73.

Yoo, K. H., Leung, C. K., & Nasridinov, A. (2022). Big Data Analysis And Visualization: Challenges And Solutions. *Applied Sciences*, 12(16), 8248.

PROFIL PENULIS

Ida Farida, S.T., M.Kom.
menyelesaikan pendidikan Sarjana (S1) pada Program Studi Teknik Informatika Universitas Mercu Buana tahun 2002 dan melanjutkan pendidikan Magister (S2) Ilmu Komputer pada Universitas Budi Luhur tahun 2016. Saat ini penulis aktif sebagai dosen pada bidang Teknik Informatika, serta terlibat dalam berbagai kegiatan akademik yang meliputi pengajaran, penelitian, dan pengabdian kepada masyarakat. Selain itu, penulis juga memiliki pengalaman sebagai pengembang sistem (*web developer*) dengan fokus pada pengembangan aplikasi berbasis web dan *Learning Management System* (LMS) *Moodle*.

Penulis memiliki minat keilmuan pada bidang teknologi informasi, khususnya dalam pengembangan sistem informasi, *e-learning,* serta pemanfaatan teknologi dalam dunia pendidikan. Melalui karya buku ini, penulis berharap dapat memberikan kontribusi dalam pengembangan ilmu pengetahuan, khususnya di bidang teknologi informasi, serta menjadi referensi yang bermanfaat bagi mahasiswa, akademisi, dan praktisi.

Email Penulis: dae.farida@mercubuana.ac.id.

BAB 13
ETIKA DATA, *PRIVACY*, DAN TANTANGAN MASA DEPAN *BIG DATA ANALYTICS*

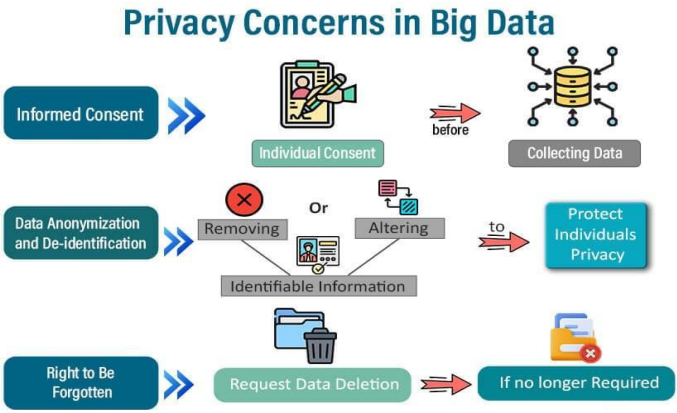
Dody, S.Kom., M.Kom.
Institut Teknologi Perusahaan Listrik Negara

Pendahuluan: Mengapa Etika dan Privasi Menjadi "Metode" di Era *Big Data*

Perkembangan *big data* telah membawa perubahan mendasar dalam cara data dikumpulkan, diproses, dan dianalisis. Karakteristik utama *big data* yang dikenal dengan volume, *velocity*, dan *variety* menjadikan data tidak lagi sekadar objek pasif untuk dianalisis, melainkan sumber daya strategis yang mempengaruhi pengambilan keputusan dalam skala besar.

Sebagaimana dibahas pada transformasi dari *small data* ke big data telah mengubah paradigma metodologi kuantitatif, sementara pemrosesan *data streaming* dan *real-time analytics* menegaskan bahwa kecepatan aliran data (*real-time* dan *streaming*) mempercepat siklus pengambilan keputusan berbasis analitik. Dalam konteks ini, analisis data tidak hanya bersifat retrospektif, tetapi semakin bersifat prediktif dan preskriptif, dengan dampak langsung terhadap individu, kelompok, maupun organisasi.

Keputusan yang dihasilkan oleh model statistik, machine learning, dan *deep learning* dapat menentukan akses terhadap layanan keuangan, pendidikan, kesehatan, serta membentuk preferensi dan perilaku pengguna melalui sistem rekomendasi. Oleh karena itu, etika dan privasi tidak lagi dapat diposisikan sebagai isu tambahan di luar metodologi, melainkan harus dipahami sebagai bagian integral dari metode analisis itu sendiri.



Gambar 13.1: Isu Privasi Dalam *Big Data* (Sari, 2026)

Sumber: Diolah Penulis.

Pendekatan kuantitatif yang dibahas pada fondasi matematika dan statistik untuk *big data analytics* dan probabilitas dan inferensi statistik dalam konteks *big data*, seperti probabilitas, inferensi statistik, serta teknik estimasi, pada dasarnya berangkat dari asumsi rasionalitas dan objektivitas. Namun, dalam praktik *big data analytics*, asumsi tersebut seringkali berhadapan dengan realitas sosial yang kompleks. Teknik *sampling*, uji hipotesis dan pengendalian kesalahan, regresi berdimensi tinggi hingga analisis deret waktu berskala besar, apabila diterapkan tanpa pertimbangan etika, berpotensi menghasilkan kesimpulan yang secara statistik valid tetapi berdampak merugikan bagi pihak tertentu. Urgensi etika dan privasi dalam *big data analytics* dapat dilihat dari berbagai risiko nyata. Kesalahan segmentasi kredit, misalnya, dapat menyebabkan kelompok tertentu secara sistematis dirugikan akibat bias data historis. Sistem rekomendasi yang manipulatif dapat membentuk filter *bubble* dan mengarahkan perilaku pengguna demi kepentingan tertentu. Sementara itu, kebocoran data dalam sistem terdistribusi dan *cloud* dapat mengakibatkan kerugian finansial, pelanggaran hak privasi, serta hilangnya kepercayaan publik terhadap teknologi analitik. Oleh karena itu, etika dan privasi perlu diposisikan sebagai "metode pengaman" (*methodological safeguards*) yang berjalan sejajar dengan metode kuantitatif dan teknik komputasional. Pertimbangan etika harus hadir sejak tahap perancangan analisis, pemilihan data, pembangunan model, hingga interpretasi dan implementasi hasil.

Dengan pendekatan ini, *big data analytics* tidak hanya menghasilkan model yang akurat dan efisien, tetapi juga bertanggung jawab, adil, dan berkelanjutan. Sub-bab ini memberikan landasan konseptual bagi pembahasan selanjutnya, dengan tujuan agar pembaca memahami bahwa keberhasilan *big data analytics* tidak semata diukur dari performa model, tetapi juga dari sejauh mana analisis tersebut menghormati nilai-nilai etika dan privasi dalam penerapannya di dunia nyata.

Prinsip Etika Data Dalam *Big Data Analytics*

Prinsip etika data merupakan fondasi normatif yang mengarahkan bagaimana data dikumpulkan, dikelola, dianalisis, dan dimanfaatkan

secara bertanggung jawab. Dalam konteks *big data analytics,* etika data menjadi semakin penting karena analisis tidak lagi sekadar menghasilkan informasi deskriptif, tetapi juga mendasari keputusan otomatis dan semi otomatis yang berdampak luas pada individu, organisasi, dan masyarakat. Berbeda dengan pendekatan analitik tradisional yang terbatas pada dataset kecil dan keputusan berbasis manusia, *big data analytics* menggabungkan skala data masif, kecepatan pemrosesan tinggi, serta kompleksitas model kuantitatif dan *machine learning*. Kondisi ini menuntut prinsip etika tidak hanya dipahami sebagai nilai moral abstrak, tetapi sebagai bagian integral dari metodologi analitik.

Gambar 13.2: Manfaat *Big Data Analytics*
Sumber: PG, 2018.

1. **Etika Data Sebagai Bagian Dari Metodologi Analitik**
   Dalam *big data analytics*, etika data tidak dapat dipisahkan dari metode kuantitatif yang digunakan. Setiap keputusan metodologis mulai dari pemilihan data, teknik sampling, model statistik, hingga interpretasi hasil yang mengandung implikasi etis (Jubaeli *et al.,* n.d.). Sebagai contoh, keputusan untuk menggunakan data historis tanpa koreksi bias dapat menghasilkan model yang memperkuat ketimpangan. Demikian pula, pemilihan metrik evaluasi yang hanya berfokus pada akurasi dapat mengabaikan dampak sosial dari

Daftar Pustaka

Amalia, A., & Putri, S. A. (2026). *Informatika Kesehatan: Integrasi Teknologi Dalam Transformasi Layanan Medis Modern*. Afdan Rojabi Publisher.

Jerendi, C., Ayu, F., Nasution, R., & Sitorus, S. P. (2026). Kecerdasan Teknologi Big Data Dalam Transformasi Digital. *Buku*, 85.

Jubaeli, A., Inayah, S., Hamid, S., & Wirasapta, A. H. (n.d.). *Metodologi Penelitian Berbasis Teknologi Dan Data Digital*.

Manurung, S., Saputra, O., Kusumastuti, S. Y., Rahman, R. A., Suryati, P., Manjaruni, M. I., & Budiasto, J. (2026). *Big Data: Konsep, Teknologi, dan Penerapan Era Modern*. Star Digital Publishing.

PG, D. S. W. (2018). Potential Benefits And Business Value of Big Data Analytics. *Majalah Ilmiah Bijak*, *15*(2), 106–114.

Sari, A. K. (2026). Perlindungan Hukum Terhadap Pengelolaan Big Data Dalam Perspektif Undang-Undang Perlindungan Data Pribadi di Indonesia. *Judge: Jurnal Hukum*, *6*(07), 2234–2239.

Zhong, Z. (2025). Big Data Engineering And Intelligent Analysis Framework For Compliance Investigation. *Academic Journal of Computing & Information Science (2025)*, *8*(11), 107–115.

PROFIL PENULIS

Dody, S.Kom., M.Kom.

Penulis merupakan seorang akademisi dan praktisi di bidang Teknologi Informasi dengan fokus pada pengembangan Sistem Informasi, *Knowledge Management*, dan *E-Government*. Saat ini berafiliasi dengan Institut Teknologi PLN dan aktif dalam berbagai kegiatan akademik serta pengembangan keilmuan berbasis teknologi *digital*. Pendidikan formal ditempuh melalui jenjang Sarjana (S1) Sistem Informasi di Universitas Budi Luhur, yang kemudian dilanjutkan pada jenjang Magister (S2) Teknologi Sistem Informasi di universitas yang sama.

Latar belakang pendidikan tersebut membentuk kompetensi dalam mengintegrasikan aspek teknis dan manajerial dalam pengelolaan sistem informasi. Memiliki minat yang kuat pada bidang *Knowledge Management*, *Information System, Business Process Modeling,* serta *E-Business*, penulis secara konsisten berkontribusi dalam pengembangan konsep dan implementasi teknologi informasi yang memberikan nilai tambah bagi organisasi dan masyarakat. Melalui buku ini, diharapkan pembaca dapat memperoleh wawasan, pemahaman, serta kontribusi nyata dalam pengembangan ilmu pengetahuan di bidang sistem informasi dan analitik data, khususnya dalam perspektif yang aplikatif dan relevan dengan kebutuhan masa kini.

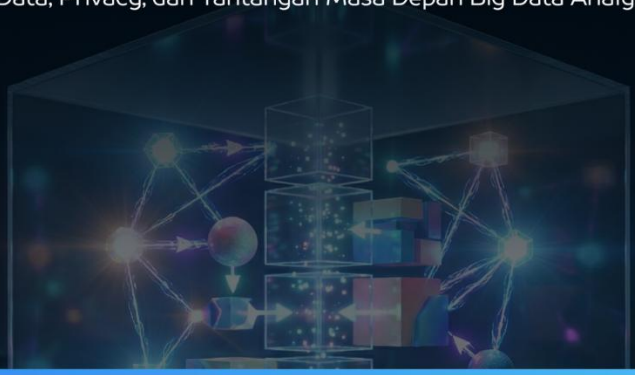
Email Penulis: dody@itpln.ac.id

METODE KUANTITATIF ERA BIG DATA

Teori dan Implementasi

Buku ini hadir sebagai respons atas pesatnya perkembangan teknologi dan ledakan data (big data) yang mengubah paradigma analisis kuantitatif di berbagai bidang, baik ekonomi, sosial, kesehatan, maupun bisnis. Perkembangan era big data tidak hanya membawa tantangan dalam hal volume, kecepatan, dan keragaman data, tetapi juga membuka peluang besar bagi peneliti dan praktisi untuk memperoleh wawasan yang lebih mendalam dan akurat. Metode kuantitatif konvensional perlu diperkaya dan diadaptasi dengan pendekatan komputasi modern, pembelajaran mesin (machine learning), serta teknik analisis data berskala besar. Buku ini disusun untuk menjembatani kesenjangan antara teori statistik klasik dan praktik analisis data di era digital. Materi dalam buku ini mencakup:

1. Pengantar Era Big Data dan Transformasi Metodologi Kuantitatif
2. Paradigma Baru dalam Analisis Data: Dari Small Data ke Big Data
3. Fondasi Matematika dan Statistik untuk Big Data Analytics
4. Probabilitas dan Inferensi Statistik dalam Konteks Big Data
5. Arsitektur Sistem Big Data: Hadoop, Spark, dan Cloud Computing
6. Database NoSQL dan Sistem Penyimpanan Terdistribusi
7. Visualisasi Data untuk Dataset Berskala Besar
8. Time Series Analysis pada Data Berskala Masif
9. Supervised Learning: Classification dan Regression pada Big Data
10. Ensemble Methods dan Model Selection Strategies
11. Network Analysis dan Graph Mining
12. Studi Kasus: Implementasi Big Data Analytics dalam Berbagai Sektor
13. Etika Data, Privacy, dan Tantangan Masa Depan Big Data Analytics



sada kurnia pustaka

ISBN 978-634-7522-52-8

9 786347 522528