

BIG DATA:

KONSEP, TEKNOLOGI, DAN APLIKASI DI ERA DIGITAL



Penulis :

Andy Dharmalau; Ike Kurniati; Tuhfatul Habibah Hasibuan; Fitri Handayani; Lela Nurlaela; Adi Sopian; Jamah Sari; Jhon Larsen

BIG DATA: KONSEP, TEKNOLOGI, DAN APLIKASI DI ERA DIGITAL

**Andy Dharmalau; Ike Kurniati;
Tuhfatul Habibah Hasibuan; Fitri Handayani;
Lela Nurlaela; Adi Sopian; Jamah Sari; Jhon Larsen**

Editor: Ahmad Fitriansyah



PT. Mustika Sri Rosadi

BIG DATA: KONSEP, TEKNOLOGI, DAN APLIKASI DI ERA DIGITAL

Penulis:

Andy Dharmalau; Ike Kurniati; Tuhfatul Habibah Hasibuan; Fitri Handayani; Lela Nurlaela; Adi Sopian; Jamah Sari; Jhon Larsen.

Editor: Ahmad Fitriansyah

Layout: Tim PT. Mustika Sri Rosadi

Desain Sampul: Tim PT. Mustika Sri Rosadi

ISBN: 978-634-7535-66-5 (PDF)

Cetakan Pertama: 1 Maret 2026

Hak Cipta 2026

Hak Cipta Dilindungi Oleh Undang-Undang

Dilarang memperbanyak karya tulis ini dalam bentuk dan dengan cara apapun tanpa izin tertulis dari penerbit.

Diterbitkan oleh Penerbit Mustika Sri Rosadi

Alamat Penerbit: Citra Indah City, Bukit Heliconia AG 23/32,
Kecamatan Jonggol, Kab. Bogor.

Email: mars.mustikasrirosadi@gmail.com

KATA PENGANTAR

Puji syukur ke hadirat Tuhan Yang Maha Esa atas limpahan rahmat dan karunia-Nya sehingga buku " BIG DATA: KONSEP, TEKNOLOGI, DAN APLIKASI DI ERA DIGITAL" dapat diselesaikan dengan baik.

Buku ini hadir untuk menjawab kebutuhan untuk mengelola dan menganalisis Big Data dengan menyajikan materi secara sistematis, mulai dari konsep dasar, arsitektur teknologi, metode analisis, hingga isu keamanan, privasi, dan etika. Dilengkapi dengan contoh penerapan nyata dan latihan soal, buku ini diharapkan dapat menjadi referensi bagi mahasiswa, akademisi, dan praktisi yang ingin mendalami Big Data.

Kami menyadari buku ini masih memiliki kekurangan. Kritik dan saran yang membangun sangat kami harapkan untuk perbaikan ke depan. Terima kasih kepada semua pihak yang telah mendukung penyusunan buku ini. Akhir kata, selamat membaca dan semoga buku ini dapat membantu memahami dan memanfaatkan Big Data secara bijak dan bertanggung jawab.

Bogor, 1 Maret 2026

Penulis

DAFTAR ISI

KATA PENGANTAR	I
DAFTAR ISI.....	II
BAB 1. PENGENALAN BIG DATA	
ANDY DHARMALAU	1
A. Pendahuluan.....	1
B. Evolusi Data Digital	2
C. Perbedaan Big Data dan Data Tradisional.....	5
D. Konsep 5V Big Data	6
E. Big Data Dari Aktivitas Digital Sehari-hari	8
F. Latihan Soal.....	8
BAB 2. ARSITEKTUR DAN INFRASTRUKTUR BIG DATA	
IKE KURNIATI	11
A. Pendahuluan.....	11
B. Komponen Utama Arsitektur dan Infrastruktur Big Data.....	12
C. Arsitektur Ekosistem Big Data Berbasis Hadoop.....	14
D. Infrastruktur Big Data.....	17
E. Model Arsitektur Big Data	19
F. Latihan Soal.....	22
BAB 3. TEKNOLOGI DAN ALAT BIG DATA	
TUHFATUL HABIBAH HASIBUAN.....	25
A. Pendahuluan.....	25

B. Cara Kerja Big Data	26
C. Teknologi Big Data.....	29
D. Alat Big Data.....	30
E. Latihan Soal.....	33
 BAB 4. DATA MINING DAN MACHINE LEARNING DALAM BIG DATA	
FITRI HANDAYANI	36
A. Pendahuluan.....	36
B. Konsep <i>Data Mining</i> dalam <i>Big Data</i>	36
C. Konsep <i>Machine Learning</i> dalam <i>Big Data</i> ..	37
D. Teknik <i>Clustering</i>	38
E. Teknik <i>Classification</i>	39
F. Studi Kasus dan Contoh Aplikasi <i>Machine Learning</i> dalam Big Data.....	40
G. Tantangan dan Peluang Pengembangan.....	41
H. Latihan Soal.....	42
 BAB 5. VISUALISASI DAN ANALITIK BIG DATA	
LELA NURLAELA.....	44
A. Pendahuluan.....	44
B. Visualisasi Big Data	51
C. Pengenalan Tools Visualisasi	55
D. Contoh Penerapan Teknik Visualisasi.....	57
E. Pengembangan Dashboard Analitik.....	63
F. Interpretasi Hasil Analisis Data	66

G. Visualisasi Data Sebagai Alat Pendukung Keputusan.....	72
H. Latihan Soal.....	74
BAB 6. KEAMANAN DAN PRIVASI BIG DATA	
ADI SOPIAN	77
A. Pendahuluan.....	77
B. Konsep Dasar Keamanan Big Data.....	79
C. Karakteristik Big Data dan Implikasinya terhadap Keamanan Data	82
D. Ancaman Keamanan dalam Lingkungan Big Data	85
E. Privasi Data dalam Big Data.....	88
F. Teknologi Keamanan Big Data.....	90
G. Regulasi dan Kebijakan Keamanan serta Privasi Data	93
H. Tantangan dan Isu Etika	95
I. Latihan Soal.....	97
BAB 7. APLIKASI BIG DATA DI BERBAGAI INDUSTRI	
JAMAH SARI.....	100
A. Pendahuluan.....	100
B. Pemanfaatan Aplikasi Big Data di Berbagai Industri.....	101
C. Penerapan Big Data Pada Sektor Kesehatan	103

D. Penerapan Big Data Pada Sektor Ekonomi dan Bisnis.....	104
E. Penerapan Big Data Pada Sektor Transportasi dan Logistik.....	105
F. Penerapan Big Data Pada Sektor Pemerintahan dan Sosial	105
G. Latihan Soal.....	106
BAB 8. TANTANGAN DAN MASA DEPAN BIG DATA	
JHON LARSEN.....	110
A. Pendahuluan.....	110
B. Tantangan Utama dalam Implementasi Big Data	114
C. Masa Depan Big Data: Menuju Era Intelijen Otonom	115
D. Latihan Soal.....	119
DAFTAR PUSTAKA.....	123
BIOGRAFI PENULIS	132
LAMPIRAN	140
SINOPSIS.....	142

BAB 1. PENGENALAN BIG DATA

A. Pendahuluan

Big data merupakan sebuah konsep yang telah lama ada, namun baru dalam beberapa tahun ini orang mulai memanfaatkannya. Big data menjadi isu yang berkembang pada saat ini seiring dengan perkembangan teknologi yang semakin pesat. Perkembangan teknologi mendorong manusia untuk terus memproduksi data baik itu data personal maupun data institusi ataupun negara (Adriyana & Darumoyo, 2020).

Berbagai organisasi baik swasta maupun pemerintahan mulai memanfaatkan Big data untuk mendapatkan wawasan lebih mendetail tentang operasi bisnis mereka, sehingga memungkinkan dalam membuat sebuah keputusan bisnis yang lebih baik.

Big data bekerja berdasarkan prinsip empat V yaitu: *Volume*, *Velocity*, *Variety*, dan *Veracity* (Dharma & Indra, 2024). Sesuai namanya, big data artinya sejumlah besar data dihasilkan setiap hari, mencapai skala *petabyte* (Maryanto, 2017). Bentuk data yang dihimpun memiliki berbagai bentuk seperti terstruktur, semi-terstruktur, dan tidak terstruktur. Data ini berasal dari segala sumber dan dihasilkan oleh banyak sistem dan perangkat, seperti sistem transaksional, berkas log, perangkat GPS, ponsel

pintar, pembaca RFID, kamera pengawas, jaringan sensor, *Internet of Things (IoT)*, media sosial dan lain lain (Hapsari, 2020). Big data pada akhirnya menjadi sebuah aset penting bagi perusahaan atau organisasi, yang fokusnya tertuju pada kepercayaan data dan sumbernya.

Pesatnya pertumbuhan eksponensial dalam volume, variasi, dan kecepatan data, pendekatan tradisional untuk manajemen dan pemrosesan data menjadi tidak memadai lagi. Hal ini telah mendorong pengembangan konsep Big Data (Dharma & Indra, 2024; Suratnoaji et al., 2019), dengan memanfaatkan teknologi dan metode canggih untuk menangani serta mengekstraksi wawasan berharga dari data dalam skala yang sangat besar. Pendekatan ini mencakup penggunaan algoritma analitik, kecerdasan buatan, dan infrastruktur komputasi modern untuk mengatasi tantangan dan peluang yang ditawarkan oleh ekosistem data yang berkembang pesat.

B. Evolusi Data Digital

Sejarah big data berawal dari masa-masa awal komputasi dan kemunculan data digital. Berikut ikhtisar singkat tentang tonggak-tonggak penting dalam evolusi big data (Mohan, 2024):

1. Era Munculnya Basis Data: Pada tahun 1960-an, konsep basis data dan sistem manajemen data mulai terbentuk. Pengembangan sistem manajemen basis

data relasional (RDBMS) seperti System R milik IBM dan RDBMS milik Oracle meletakkan fondasi bagi penyimpanan dan pengambilan data terstruktur.

2. Era Pergudangan Data: Pada tahun 1980-an, konsep pergudangan data muncul. Pergudangan data melibatkan konsolidasi data dari berbagai sumber ke dalam repository pusat untuk keperluan analisis dan pelaporan. Pendekatan ini memungkinkan organisasi untuk memanfaatkan aset data mereka untuk pengambilan keputusan.
3. Era Internet dan Web 1.0: Kebangkitan dan kemunculan World Wide Web pada tahun 1990-an menyebabkan peningkatan besar dalam pembuatan data. Web menjadi platform untuk menerbitkan, berbagi, dan mengumpulkan data, yang menyebabkan pertumbuhan sumber data tidak terstruktur dan semi terstruktur.
4. Era Transformasi Digital dan Sistem Perusahaan: Organisasi mulai menghasilkan data terstruktur dalam jumlah besar dengan meluasnya adopsi sistem perusahaan seperti Perencanaan Sumber Daya Perusahaan (ERP) dan Manajemen Hubungan Pelanggan (CRM). Hal ini menandai dimulainya pengambilan keputusan berbasis data dalam bisnis.
5. Era Hadoop dan Komputasi Terdistribusi: Pada tahun 2005, Doug Cutting dan Mike Cafarella mengembangkan Hadoop, sebuah kerangka kerja sumber terbuka yang dirancang untuk memproses

kumpulan data besar dalam lingkungan komputasi terdistribusi. Sistem berkas terdistribusi (HDFS) Hadoop dan model pemrograman MapReduce memudahkan penanganan big data dengan memecahnya menjadi bagian-bagian yang lebih kecil dan memprosesnya secara paralel.

6. Era Kebangkitan NoSQL: Basis data relasional tradisional kesulitan menangani volume, variasi, dan kecepatan big data. Akibatnya, pengembangan basis data NoSQL (Not Only SQL) menjadi terkenal di akhir tahun 2000-an. Basis data NoSQL menawarkan desain skema yang fleksibel dan skalabilitas horizontal, sehingga cocok untuk menangani data tidak terstruktur dan semi-terstruktur.
7. Era Analisis Lanjutan dan Pembelajaran Mesin: Adanya Big data memungkinkan berbagai organisasi untuk mengekstrak wawasan dan pola berharga dari data tersebut. Teknik analisis lanjutan, seperti penambangan data, pembelajaran mesin, dan pemodelan prediktif, semakin diminati untuk mengungkap wawasan yang dapat ditindaklanjuti dan mendukung pengambilan keputusan.
8. Era Komputasi Awan: Munculnya komputasi awan di akhir tahun 2000-an menyediakan infrastruktur yang scalable dan hemat biaya bagi organisasi untuk menyimpan, memproses, dan menganalisis big data.
9. Era Internet of Things (IoT): Banyaknya perangkat dan sensor yang terhubung, jumlah data yang dihasilkan

oleh perangkat IoT meroket. IoT berkontribusi pada ekspansi big data, karena data streaming real-time dari berbagai sumber menjadi tersedia untuk dianalisis.

10. Tren Saat Ini: Analitik big data terus berkembang, didorong oleh kemajuan teknologi seperti kecerdasan buatan, pemrosesan bahasa alami, dan pembelajaran mendalam. Fokusnya telah bergeser ke arah analitik real-time, tata kelola data, masalah privasi, dan penggunaan data yang etis.

Pengelolaan dan analisis kumpulan data yang besar dan beragam terus berkembang, dan berbagai organisasi terus mengeksplorasi pendekatan dan teknologi inovatif untuk memanfaatkan potensi big data untuk wawasan dan pengambilan keputusan.

C. Perbedaan Big Data dan Data Tradisional

Sistem Data Tradisional adalah sistem pengelolaan data yang digunakan untuk mengatasi semua permasalahan bisnis, menggunakan pengelolaan data secara tradisional dengan cara menyimpan record-record pada file-file yang terpisah yang disebut juga sistem pemrosesan file.

Pada data tradisional, sistem pengelolaannya setiap aplikasi membutuhkan file dan program tersendiri. Organisasi file pada aplikasi tersebut pada umumnya sederhana serta logika bergantung pada file dan

program tertentu, tentunya sangat tidak efisien dan rumit. Akses data pada sistem tradisional ini sulit diakses pihak lain di luar unit yang bersangkutan, karena sistem informasi tidak terkoneksi antara satu unit dan unit lainnya.

Perbedaan antara Big data dengan data tradisional dapat dilihat dari perbedaan ukuran data yang jauh lebih besar, kecepatan data yang terus mengalir, banyaknya bentuk data serta nilai yang terkandung dalam sebuah data. Pada data tradisional biasanya terstruktur dalam tabel dan angka, namun pada Big Data dalam berbagai bentuk, mulai dari teks, gambar, video, hingga data yang tidak terstruktur seperti komentar di media sosial. Analisis Big Data membutuhkan pendekatan yang lebih canggih, seperti *Machine learning* dan kecerdasan buatan (AI), untuk menemukan pola tersembunyi. Sedangkan dalam pengolahan data tradisional, analisis biasanya dilakukan dengan metode statistik sederhana.

D. Konsep 5V Big Data

Fenomena Big Data, dimulai pada tahun 2000-an ketika seorang analis industri Doug Laney menyampaikan konsep Big Data yang terdiri dari tiga bagian penting (Hartawan et al., 2022) (Wali et al., 2023), diantaranya:

1. Volume: Data dikumpulkan dari berbagai sumber, termasuk transaksi bisnis, media sosial dan informasi dari sensor atau mesin.
2. Kecepatan: Pemrosesan Aliran data harus cepat dan tepat bisa melalui hardware maupun software.
3. Variasi: Data yang dikumpulkan mempunyai format berbeda-beda baik dari jenis data, termasuk terstruktur, semi-terstruktur, dan tidak terstruktur.

Sesuai dengan perkembangan teknologi lalu di tambahkan dua V lagi yaitu: nilai dan kebenaran. Penambahan dua V ini untuk membantu ilmuwan data mengartikulasikan dan mengomunikasikan karakteristik penting big data secara lebih efektif. Karakteristik 5V dalam big data mengacu pada lima karakteristik utama yaitu: *Volume* (jumlah data yang sangat besar), *Velocity* atau Kecepatan (laju data dibuat dan bergerak), *Variety* atau Variasi (beragam jenis dan format data), *Veracity* atau Kebenaran (keakuratan dan kualitas data), dan *Value* atau Nilai (manfaat yang bisa diperoleh dari data tersebut). Karakteristik 5V ini merupakan kerangka kerja untuk memahami tantangan dan peluang yang ditawarkan Big Data. Organisasi yang mampu mengelola kelima aspek ini dengan efektif dapat memperoleh keunggulan kompetitif melalui pengambilan keputusan yang lebih baik, inovasi produk, dan peningkatan efisiensi operasional.

E. Big Data Dari Aktivitas Digital Sehari-hari

Era revolusi 4.0 dan berkembang pesatnya teknologi Big Data, membuat pertumbuhan eksponensial dalam sejumlah data akibat dari pesatnya perkembangan teknologi yang digunakan dalam aktivitas sehari-hari. Berikut ini beberapa Contoh Big Data dari Aktivitas Digital Sehari-hari seperti: aktivitas pada Platform media sosial, Situs belanja online, Penggunaan Platform Google Maps atau Waze, Platform seperti Netflix dan Spotify dll.

Manfaat Big Data dalam kehidupan sehari-hari, dapat digunakan untuk peningkatan efisiensi, pengambilan keputusan yang Lebih Baik, merekomendasikan hal-hal yang lebih personal, dan memberikan wawasan yang mendalam tentang perilaku konsumen dan tren pasar.

F. Latihan Soal

1. Berbagai organisasi memanfaatkan Big data untuk mendapatkan:
 - a. Wawasan lebih mendetail tentang operasi bisnis.
 - b. Keuntungan dari bisnis yang dijalankan.
 - c. Efisiensi dapat dilakukan pada operasi bisnis.
 - d. Efektivitas lebih detail dari operasional bisnis.
2. Berikut ini beberapa contoh sumber Big Data dari aktivitas digital sehari-hari kecuali:

- a. Aktivitas media sosial.
 - b. Aktivitas belanja online.
 - c. Aktivitas mencari rute terdekat.
 - b. Aktivitas membaca buku.
3. Doug Laney menyampaikan konsep Big Data yang terdiri dari tiga bagian penting yaitu:
- a. Volume, Ketepatan, Variasi
 - b. Variasi, Volume, Kecepatan
 - c. Volume, Variasi, Teknologi
 - d. Kecepatan, Kebenaran, Variasi
4. Berikut ini merupakan manfaat Big Data dalam kehidupan sehari-hari kecuali:
- a. Memudahkan Konsumen akan kebutuhannya.
 - b. Peningkatan Efisiensi.
 - c. Pengambilan keputusan yang Lebih Baik.
 - d. Merekomendasikan yang lebih personal.
5. Untuk menganalisis Big Data membutuhkan pendekatan teknologi yang lebih canggih, seperti:
- a. Kecerdasan buatan (AI) dan Data Science
 - b. Machine Learning dan Kecerdasan Intelligence
 - c. Kecerdasan buatan (AI) dan Machine Learning
 - d. Kecerdasan buatan (AI) dan Robotika
6. Pernyataan berikut ini cara yang dilakukan pada Sistem Data Tradisional:
- a. Menyimpan record pada file yang terkoneksi.
 - b. Setiap File Terhubung dengan file lainnya.

- c. Analisis dengan metode statistik sederhana.
 - d. Menggunakan Aplikasi My SQL.
7. Pernyataan berikut ini merupakan contoh bentuk data pada Big data kecuali:
- a. Gambar
 - b. Komentar di Media sosial.
 - c. Video.
 - d. Pamflet
8. Analisis yang mendalam dalam pada Big data dilakukan untuk:
- a. Menemukan pola tersembunyi
 - b. Menemukan Sistem terkini
 - c. Menemukan Metoda terbaru
 - d. Menemukan Gambaran dari jenis data
9. Konsep karakteristik Big data pada awalnya dibahas dalam konteks tiga V, lalu di tambahkan dua V lagi yaitu:
- a. Sinergi dan Kecepatan.
 - b. Kebenaran dan Nilai.
 - c. Kompleksitas dan Validitas
 - d. Nilai dan Kompleksitas
10. Sumber data seperti apa yang menyebabkan Big data merupakan data yang besar?
- a. Sumber data berupa teks
 - b. Sumber data dari berbagai bentuk
 - c. Sumber data berupa angka statistic
 - d. Sumber data hanya dari kuesioner

BAB 2. ARSITEKTUR DAN INFRASTRUKTUR BIG DATA

A. Pendahuluan

Big Data merupakan istilah yang digunakan untuk menggambarkan kumpulan data dengan karakteristik volume yang sangat besar, kecepatan pertumbuhan yang tinggi, serta keragaman format yang kompleks. Data tersebut dapat bersumber dari transaksi digital, sensor *Internet of Things* (IoT), media sosial, maupun sistem informasi organisasi. Kompleksitas yang melekat pada Big Data menuntut adanya pendekatan teknologi yang berbeda dari sistem pengolahan data konvensional.

Dalam konteks ini, arsitektur Big Data berperan sebagai kerangka konseptual yang mengatur alur data mulai dari proses akuisisi, penyimpanan, pemrosesan, hingga analisis. Sementara itu, infrastruktur Big Data menyediakan fondasi teknis berupa perangkat keras, perangkat lunak, serta jaringan yang memungkinkan sistem berjalan secara skalabel, andal, dan aman. Tanpa dukungan arsitektur dan infrastruktur yang memadai, potensi Big Data tidak dapat dimanfaatkan secara optimal dan justru berisiko menjadi beban operasional.

Di Indonesia, urgensi penerapan arsitektur dan infrastruktur Big Data semakin nyata seiring dengan pesatnya perkembangan ekonomi digital. Asosiasi

Penyelenggara Jasa Internet Indonesia (APJII) mencatat bahwa pada tahun 2025 jumlah pengguna internet mencapai 229,4 juta jiwa dari total populasi 284,4 juta, dengan tingkat penetrasi sebesar 80,66% (Shabrina, Sarah, 2025). Sementara itu, Bank Indonesia melaporkan nilai transaksi e-commerce mencapai Rp44,4 triliun pada Juli 2025, meningkat 6,41% secara tahunan (Gita Amanda, 2025). Industri telekomunikasi juga menjadi salah satu penghasil data terbesar, dengan ratusan juta nomor seluler aktif yang digunakan untuk analisis pola konsumsi layanan digital (Riyanto, 2025). Pemerintah pun mulai mengintegrasikan Big Data dalam program *Smart City*, seperti Jakarta *Smart City*, guna memantau transportasi, kesehatan, dan lingkungan secara real-time. Fakta-fakta ini menunjukkan bahwa arsitektur dan infrastruktur Big Data bukan sekadar kebutuhan teknis, melainkan fondasi strategis bagi pertumbuhan industri dan tata kelola modern di Indonesia.

B. Komponen Utama Arsitektur dan Infrastruktur Big Data

Secara umum, arsitektur dan infrastruktur Big Data terdiri atas sejumlah komponen utama yang saling berinteraksi secara integratif, yaitu:

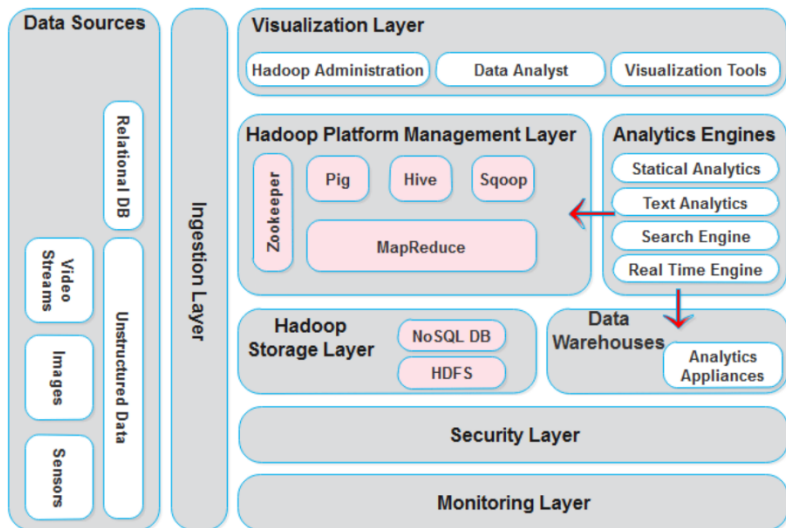
1. Sumber Data (*Data Sources*): mencakup sensor, aplikasi, media sosial, serta transaksi bisnis yang menjadi titik awal akuisisi data.

2. Lapisan Ingesti (*Data Ingestion Layer*): berfungsi sebagai mekanisme untuk memasukkan data ke dalam sistem, baik melalui pemrosesan *batch* maupun *streaming real-time*.
3. Lapisan Penyimpanan (*Data Storage Layer*): meliputi *data lake*, *data warehouse*, sistem file terdistribusi seperti Hadoop Distributed File System (HDFS), serta basis data NoSQL yang mendukung penyimpanan berskala besar.
4. Lapisan Pemrosesan (*Data Processing Layer*): menggunakan kerangka kerja seperti MapReduce, Apache Spark, atau Flink untuk mengolah data dalam jumlah masif secara efisien.
5. Lapisan Analisis (*Data Analysis Layer*): melibatkan penerapan teknik machine learning, business intelligence, serta visualisasi data guna menghasilkan wawasan strategis yang relevan bagi pengambilan keputusan.
6. Lapisan Akses dan Keamanan (*Data Access & Security Layer*): mencakup API, autentikasi, enkripsi, serta kontrol akses yang bertujuan menjaga integritas, kerahasiaan, dan keamanan data.

Keseluruhan komponen tersebut membentuk suatu ekosistem teknologi yang terintegrasi, memungkinkan organisasi untuk mengubah data mentah menjadi informasi yang bernilai, sekaligus mendukung proses pengambilan keputusan berbasis data secara lebih akurat dan efektif.

C. Arsitektur Ekosistem Big Data Berbasis Hadoop

Arsitektur Big Data modern dirancang untuk menangani volume data yang sangat besar, beragam jenis data, serta kebutuhan analisis yang kompleks dan real-time. Salah satu pendekatan yang umum digunakan adalah ekosistem Hadoop, yang menyediakan kerangka kerja terdistribusi untuk penyimpanan dan pemrosesan data skala besar. Diagram arsitektur ini mengilustrasikan sembilan lapisan utama yang membentuk sistem Big Data yang komprehensif.



Gambar 2.1 Arsitektur Big Data Berbasis Hadoop

1. Data Sources

Lapisan ini mencakup berbagai sumber data, baik terstruktur maupun tidak terstruktur. Data terstruktur berasal dari Relational Database Management

Systems (RDBMS), sedangkan data tidak terstruktur meliputi video streams, images, dan data sensor dari perangkat IoT. Keberagaman sumber ini mencerminkan karakteristik variety dalam paradigma Big Data.

2. Ingestion Layer

Lapisan ini berfungsi sebagai pintu masuk data ke dalam sistem Hadoop. Proses data ingestion dapat dilakukan secara batch maupun streaming, memungkinkan sistem untuk menangani data historis dan data real-time secara simultan.

3. Hadoop Platform Management Layer

Lapisan ini menyediakan berbagai komponen manajemen dan pemrosesan data, seperti:

- a. Zookeeper: koordinasi layanan terdistribusi
- b. Pig: bahasa skrip berbasis data flow untuk analisis data
- c. Hive: antarmuka SQL untuk query data di Hadoop
- d. Sqoop: alat untuk transfer data antara Hadoop dan RDBMS
- e. MapReduce: model pemrosesan paralel berbasis pemetaan dan reduksi

4. Hadoop Storage Layer

Lapisan penyimpanan ini terdiri dari:

- a. HDFS (Hadoop Distributed File System): sistem file terdistribusi yang dirancang untuk menyimpan data dalam jumlah besar secara efisien

- b. NoSQL Database: basis data non-relasional yang mendukung fleksibilitas skema dan skalabilitas tinggi.

5. Analytics Engines

Lapisan ini mendukung berbagai jenis analisis data, termasuk:

- a. Statistical Analytics: analisis berbasis statistik
- b. Text Analytics: analisis teks dan bahasa alami
- c. Search Engine: pencarian data terindeks
- d. Real-Time Engine: analisis data secara langsung saat data masuk

6. Data Warehouses

Lapisan ini berfungsi sebagai repositori hasil analisis, menggunakan analytics appliances untuk menyimpan dan mengolah data secara efisien guna mendukung pengambilan keputusan strategis.

7. Visualization Layer

Lapisan visualisasi memungkinkan pengguna seperti data analyst dan administrator Hadoop untuk menampilkan hasil analisis dalam bentuk grafik, dashboard, dan laporan interaktif menggunakan berbagai visualization tools.

8. Security Layer

Lapisan keamanan bertanggung jawab atas perlindungan data melalui mekanisme autentikasi,

otorisasi, dan enkripsi, guna memastikan integritas dan kerahasiaan data dalam seluruh proses.

9. Monitoring Layer

Lapisan ini memantau performa sistem, mendeteksi anomali, dan memastikan bahwa seluruh komponen arsitektur berjalan optimal dan stabil.

Struktur ini mencerminkan pendekatan modular dan skalabel dalam pengelolaan Big Data, di mana setiap lapisan memiliki fungsi spesifik yang saling melengkapi. Integrasi komponen-komponen ini memungkinkan organisasi untuk mengolah data dalam skala besar secara efisien, aman, dan bernilai strategis.

D. Infrastruktur Big Data

Infrastruktur Big Data merupakan fondasi teknis yang memungkinkan organisasi untuk mengelola, menyimpan, dan menganalisis data dalam skala besar secara efektif. Komponen utama dari infrastruktur ini adalah hardware, yang mencakup server berperforma tinggi, cluster komputasi terdistribusi, jaringan berkecepatan tinggi, serta sistem penyimpanan masif. Perangkat keras tersebut berfungsi sebagai tulang punggung yang mendukung proses komputasi paralel dan distribusi data, sehingga sistem mampu menangani beban kerja yang kompleks.

Selain perangkat fisik, perkembangan teknologi komputasi awan telah menghadirkan *cloud*

infrastructure seperti Amazon Web Services (AWS), Microsoft Azure, dan Google Cloud. Infrastruktur berbasis awan ini menawarkan fleksibilitas, skalabilitas, serta efisiensi biaya melalui model *pay-as-you-go*. Dengan demikian, organisasi dapat menyesuaikan kapasitas komputasi sesuai kebutuhan tanpa harus melakukan investasi besar pada perangkat keras. Dalam konteks ini, terdapat perbedaan mendasar antara *on-premise deployment* dan *cloud deployment*. Pendekatan on-premise memberikan kontrol penuh terhadap data dan sistem, sementara cloud menawarkan kemudahan integrasi, elastisitas, serta pengurangan biaya operasional.

Lebih lanjut, teknologi virtualisasi dan containerization menjadi komponen penting dalam arsitektur Big Data modern. Virtualisasi memungkinkan pemanfaatan sumber daya komputasi secara optimal melalui mesin virtual, sedangkan containerization dengan platform seperti Docker dan Kubernetes mendukung orkestrasi aplikasi secara ringan, portabel, dan skalabel. Hal ini mempercepat siklus pengembangan, pengujian, serta penerapan aplikasi berbasis Big Data, sekaligus meningkatkan efisiensi operasional.

Akhirnya, infrastruktur Big Data harus dirancang dengan prinsip scalability dan fault tolerance. Skalabilitas memastikan sistem mampu menangani peningkatan volume data secara bertahap, sedangkan fault tolerance

menjamin keberlangsungan layanan meskipun terjadi kegagalan pada sebagian komponen. Kedua aspek ini menjadi krusial dalam menjaga keandalan sistem Big Data yang beroperasi dalam lingkungan dinamis dan penuh ketidakpastian.

Struktur ini menegaskan bahwa infrastruktur Big Data bukan sekadar perangkat keras atau perangkat lunak, melainkan ekosistem terintegrasi yang mendukung pengolahan data dalam skala besar dengan tingkat keandalan tinggi.

E. Model Arsitektur Big Data

Pemodelan arsitektur Big Data merupakan kerangka konseptual yang menentukan bagaimana data dikumpulkan, diproses, dan dianalisis dalam skala besar. Pemilihan model arsitektur tidak hanya bersifat teknis, tetapi juga strategis, karena berimplikasi langsung terhadap kecepatan, akurasi, serta efisiensi sistem analitik. Terdapat beberapa pendekatan utama yang telah berkembang, yaitu *Lambda Architecture*, *Kappa Architecture*, serta model hibrida yang mengintegrasikan keduanya.

1. Lambda Architecture

Model ini dirancang untuk mengatasi tantangan pemrosesan data dalam jumlah besar dengan menggabungkan dua pendekatan: batch processing dan real-time processing. Lapisan batch bertugas mengolah data historis dalam volume besar dengan

tingkat akurasi tinggi, sementara lapisan real-time berfungsi untuk memproses data yang masuk secara langsung sehingga menghasilkan respons cepat. Dengan demikian, Lambda Architecture mampu menyeimbangkan kebutuhan akan akurasi jangka panjang dan kecepatan analisis instan. Namun, kompleksitas pengelolaan dua jalur pemrosesan sering menjadi tantangan dalam implementasi praktis.

2. Kappa Architecture

Sebagai alternatif, Kappa Architecture mengeliminasi lapisan batch dan berfokus sepenuhnya pada pemrosesan streaming. Seluruh data diperlakukan sebagai aliran kontinu yang diproses secara real-time, sehingga arsitektur menjadi lebih sederhana dan mudah dikelola. Pendekatan ini sangat relevan untuk aplikasi yang membutuhkan analisis instan, seperti sistem deteksi penipuan (fraud detection), pemantauan sensor IoT, atau rekomendasi konten digital. Kelebihan utama Kappa Architecture adalah kesederhanaan desain dan kemampuan adaptasi terhadap data yang terus berubah, meskipun terkadang mengorbankan akurasi analisis historis.

3. Hybrid Approaches

Selain Lambda dan Kappa, muncul pula pendekatan hibrida yang berusaha mengintegrasikan keunggulan keduanya. Model ini memungkinkan organisasi untuk menyesuaikan metode pemrosesan berdasarkan jenis data, tingkat urgensi, dan tujuan analisis. Misalnya,

data transaksi keuangan dapat diproses secara real-time untuk mendeteksi anomali, sementara data historis pelanggan dianalisis secara batch untuk mendukung strategi pemasaran jangka panjang. Pendekatan hibrida menawarkan fleksibilitas tinggi, meskipun memerlukan perencanaan arsitektur yang matang agar tidak menimbulkan redundansi sistem.

Dalam praktiknya, Lambda Architecture banyak digunakan pada sektor e-commerce untuk menggabungkan analisis historis perilaku konsumen dengan rekomendasi produk secara real-time. Kappa Architecture lebih sesuai untuk sektor keuangan, khususnya dalam fraud detection, di mana kecepatan identifikasi transaksi mencurigakan menjadi prioritas. Sementara itu, pendekatan hibrida sering diterapkan pada industri telekomunikasi, yang membutuhkan analisis trafik jaringan secara real-time sekaligus evaluasi tren jangka panjang untuk perencanaan kapasitas.

Dengan demikian, pemilihan model arsitektur Big Data harus mempertimbangkan kebutuhan spesifik organisasi, karakteristik data yang dikelola, serta tujuan analisis yang ingin dicapai. Pendekatan yang tepat akan menghasilkan sistem yang tidak hanya efisien dan responsif, tetapi juga mampu memberikan nilai strategis bagi pengambilan keputusan

F. Latihan Soal

1. Perbedaan mendasar antara on-premise deployment dan cloud deployment adalah....
 - a. On-premise lebih murah daripada cloud
 - b. On-premise memberikan kontrol penuh terhadap data, sementara cloud menawarkan elastisitas dan pengurangan biaya operasional
 - c. Cloud tidak dapat diskalakan, sementara on-premise sangat elastis
 - d. On-premise hanya untuk perusahaan kecil, cloud untuk perusahaan besar

2. Lapisan dalam arsitektur Hadoop yang berfungsi memantau performa sistem dan mendeteksi anomali adalah....
 - a. Security Layer
 - b. Visualization Layer
 - c. Data Warehouses
 - d. Monitoring Layer

3. Apache Hive dalam ekosistem Hadoop berfungsi untuk....
 - a. Transfer data antara Hadoop dan RDBMS
 - b. Koordinasi layanan terdistribusi
 - c. Antarmuka SQL untuk query data di Hadoop
 - d. Pemrosesan paralel berbasis pemetaan dan reduksi

4. Dua pendekatan pemrosesan data yang digabungkan dalam Lambda Architecture adalah....
 - a. Batch processing dan real-time processing
 - b. Structured processing dan unstructured processing
 - c. Local processing dan cloud processing
 - d. Sequential processing dan parallel processing
5. Contoh basis data non-relasional (NoSQL Database) yang mendukung fleksibilitas skema dan skalabilitas tinggi adalah komponen dari lapisan....
 - a. Hadoop Platform Management Layer
 - b. Hadoop Storage Layer
 - c. Analytics Engines
 - d. Visualization Layer
6. Teknologi yang mendukung orkestrasi aplikasi secara ringan, portabel, dan skalabel dalam infrastruktur Big Data modern adalah....
 - a. Virtual Machine
 - b. Mainframe Computer
 - c. Containerization (Docker dan Kubernetes)
 - d. Solid State Drive (SSD)
7. Lapisan dalam arsitektur Hadoop yang bertanggung jawab atas perlindungan data melalui mekanisme autentikasi, otorisasi, dan enkripsi adalah....
 - a. Visualization Layer
 - b. Monitoring Layer

- c. Security Layer
 - d. Analytics Layer
8. Berikut ini adalah komponen utama dalam arsitektur Big Data, kecuali....
- a. Sumber Data (Data Sources)
 - b. Lapisan Ingesti (Data Ingestion Layer)
 - c. Lapisan Input Data (Data Entry Layer)
 - d. Lapisan Visualisasi (Visualization Layer)
9. Lapisan Penyimpanan (Data Storage Layer)
Kerangka kerja pemrosesan paralel dalam ekosistem Hadoop yang menggunakan konsep pemetaan dan reduksi disebut....
- a. Apache Spark
 - b. MapReduce
 - c. Apache Flink
 - d. Zookeeper
10. Perangkat lunak yang berfungsi untuk transfer data antara Hadoop dan basis data relasional (RDBMS) adalah....
- a. Pig
 - b. Hive
 - c. Sqoop
 - d. Zookeeper

BAB 3. TEKNOLOGI DAN ALAT BIG DATA

A. Pendahuluan

Volume data yang dihasilkan di era digital ini, setiap detiknya telah mencapai skala yang luar biasa besar, sering disebut sebagai "Big Data". Fenomena yang terjadi ini dalam Big Data dicirikan oleh tiga karakteristik utama, yang dikenal sebagai "Tiga V" yaitu: Volume (jumlah data yang masif), Velocity (kecepatan data dihasilkan dan perlu diproses secara real-time), dan Variety (ragam format data, dari terstruktur hingga tidak terstruktur)(Santoso, 2020).

Kemampuan untuk mengekstrak wawasan berharga dari besarnya volume data dan informasi ini telah menjadi pembeda utama dalam keunggulan kompetitif di berbagai sektor industri, mulai dari keuangan, kesehatan, hingga perdagangan elektronik (Dharma & Indra, 2024).

Kemampuan untuk mengekstrak wawasan berharga dari besarnya volume data dan informasi ini telah menjadi pembeda utama dalam keunggulan kompetitif di berbagai sektor industri, mulai dari keuangan, kesehatan, hingga perdagangan elektronik.

Permasalahannya ada pada proses mengelola dan menganalisis data yang sangat besar ini, sudah tidak mungkin dan jauh melampaui kemampuan basis data

tradisional dan metode pemrosesan konvensional. Untuk itu diperlukan pendekatan infrastruktur dan perangkat lunak yang sepenuhnya baru untuk menangani tantangan komputasi dalam pengolahan data dengan volume yang sangat besar ini. Di sinilah peran krusial teknologi dan alat Big Data menjadi sentral.

Teknologi digunakan untuk mengeksplorasi dan komputasi yang digunakan untuk setiap tahap siklus hidup data, yaitu penyerapan, penyimpanan, pemrosesan, dan analisis data.

Memahami dan memanfaatkan teknologi ini sangat penting bagi organisasi yang ingin menggali potensi penuh data mereka untuk mendorong inovasi, mengoptimalkan operasi, dan membuat keputusan yang lebih cerdas dan berbasis bukti.

B. Cara Kerja Big Data

Big data adalah istilah yang menggambarkan kumpulan data dalam jumlah besar dan kompleks, sehingga sulit dikelola dengan metode tradisional (Kamal & Saxena, 2019). Di era digital, big data menjadi kunci untuk memahami pola dan tren yang tersembunyi dalam sebuah kumpulan data yang sangat besar.

Penerapan Big Data dapat diaplikasikan pada berbagai sektor, seperti kesehatan, pemasaran, dan media (Dharma & Indra, 2024). Misalnya, di sektor

kesehatan, big data membantu memprediksi wabah penyakit dan meningkatkan efisiensi perawatan (Wali et al., 2023). Dalam pemasaran, data ini memungkinkan optimasi harga dan layanan pelanggan yang lebih baik. Big data dapat memberikan sajian alternatif pilihan dalam pengambilan Keputusan berdasarkan informasi yang dihasilkan dari kumpulan data. Dengan analisis data yang mendalam, dapat memahami kebutuhan konsumen, meningkatkan efisiensi operasional, dan membuat keputusan yang lebih tepat. Teknologi ini memungkinkan organisasi untuk mengubah data menjadi wawasan yang berharga.

Big data memungkinkan untuk menganalisis pola dan tren yang tersembunyi dalam data, Hasil analisis ini menjadikan big data menjadi sebuah alat penting dalam pengambilan keputusan berbasis data. Big data bekerja melalui serangkaian proses yang melibatkan pengumpulan, penyimpanan, pemrosesan, hingga analisis data. Setiap langkah ini memanfaatkan teknologi canggih untuk mengelola data dalam jumlah besar dan kompleks. Berikut adalah langkah-langkah utama dalam proses ini:

1. Mengumpulkan data dari berbagai sumber, termasuk sensor IoT, aplikasi, atau situs web.
2. Menyimpan data dalam basis data komputer atau wadah khusus seperti Hadoop dan cloud.

3. Menggunakan perangkat lunak khusus untuk menganalisis data yang telah dikumpulkan. Proses ini memastikan bahwa semua data yang relevan tersedia untuk langkah-langkah berikutnya.
4. Setelah data terkumpul, langkah berikutnya adalah menyimpannya. Teknologi penyimpanan modern memungkinkan Anda menyimpan data dalam jumlah besar dengan efisiensi tinggi (Hapsari, 2020). Salah satu teknologi yang paling umum digunakan adalah Hadoop. Teknologi ini memungkinkan penyimpanan data dalam berbagai format, baik terstruktur maupun tidak terstruktur.
5. Setelah data disimpan, langkah berikutnya adalah Pemrosesan data bertujuan untuk mengubah data mentah menjadi informasi yang dapat digunakan. Salah satu alat yang sering digunakan adalah Apache Spark, dirancang untuk menangani tantangan data besar dengan kinerja yang cepat dan skalabilitas tinggi. Hasil dari memproses data dalam waktu singkat dan mendapatkan wawasan yang berharga untuk pengambilan keputusan.
6. Analisis Data merupakan tahapan akhir dari Big data, Proses yang dilakukannya adalah mengolah, membersihkan, mentransformasi, dan memodelkan data untuk menemukan informasi yang berguna, menarik kesimpulan yang valid, dan mendukung pengambilan keputusan yang akurat (Prasojo & Haryatmi, 2021).

7. Visualisasi dan Pengambilan Keputusan, Visualisasi data adalah langkah penting dalam proses big data. Dengan visualisasi, dapat disajikan data dalam bentuk grafik, diagram, atau peta yang mudah dipahami. Hal ini membantu untuk melihat pola, tren, dan anomali yang mungkin tidak terlihat jika hanya melihat data mentah.

C. Teknologi Big Data

Teknologi big data pastinya merujuk pada sistem dan alat untuk mengelola, menyimpan, menganalisis, dan memvisualisasikan kumpulan data yang sangat besar dan kompleks yang dihasilkan dari berbagai sumber.

Teknologi dapat mengatasi keterbatasan perangkat lunak tradisional dengan memanfaatkan karakteristik seperti volume (jumlah data besar), velocity (kecepatan data tinggi), dan variety (keragaman format data). Contoh teknologi yang sering digunakan adalah Apache Hadoop dan Apache Spark(Wali et al., 2023).

Komponen teknologi big data yaitu:

1. Penyimpanan data: Solusi seperti data lake yang fleksibel untuk menyimpan data dalam berbagai bentuk.
2. Penambangan data: Proses untuk menemukan pola tersembunyi dalam data.

3. Analisis data: Pengolahan data menggunakan alat seperti Apache Spark untuk analisis batch dan real-time.
4. Visualisasi data: Menyajikan hasil analisis dalam bentuk yang mudah dipahami.

Teknologi big data mencakup kerangka kerja terdistribusi (Hadoop, Spark), Basis data NoSQL (MongoDB, Cassandra), Analitik data (Data Mining, NLP, AI/ML), dan Platform awan (AWS, Azure, GCP) untuk mengumpulkan, menyimpan, memproses, dan menganalisis volume data besar dan beragam (Hartawan et al., 2022).

D. Alat Big Data

Alat-alat pada big data meliputi Hadoop, Apache Spark, MongoDB, Tableau, Google BigQuery, serta teknologi seperti Natural Language Processing (NLP) dan Machine Learning (ML) untuk mendapatkan wawasan berharga guna pengambilan keputusan bisnis, personalisasi layanan, dan prediksi pasar.

Ada banyak alat Big Data yang tersedia, baik gratis (open source) maupun berbayar, yang dirancang untuk berbagai tahap dalam siklus hidup data (penyerapan, penyimpanan, pemrosesan, dan analisis) (Tarumingkeng, 2025).

Beberapa alat Big Data populer meliputi:

1. Alat Sumber Terbuka (Open Source)

Ekosistem Apache menyediakan banyak alat sumber terbuka yang banyak digunakan dalam pemrosesan Big Data (Tarumingkeng, 2025):

- a. Apache Hadoop: Kerangka kerja fundamental untuk penyimpanan terdistribusi (HDFS) dan pemrosesan paralel (MapReduce) data dalam jumlah besar di kluster komoditas.
- b. Apache Spark: Mesin pemrosesan data terdistribusi yang dikenal dengan kecepatan dan kemampuan komputasi dalam memori, cocok untuk analitik real-time dan pembelajaran mesin.
- c. Apache Kafka: Platform event streaming berthroughput tinggi dan toleran terhadap kesalahan yang digunakan untuk membangun pipeline data real-time.
- d. Apache Flink: Kerangka kerja untuk komputasi stream processing yang canggih dengan latensi rendah, ideal untuk analitik real-time.
- e. Apache Cassandra/HBase/MongoDB: Basis data NoSQL yang dirancang untuk ketersediaan tinggi dan skalabilitas, digunakan untuk menyimpan data tidak terstruktur atau semi-terstruktur dalam volume besar.
- f. Apache Hive: Sistem gudang data yang menyediakan antarmuka seperti SQL untuk mengkueri dan mengelola data yang berada di penyimpanan terdistribusi seperti HDFS.

2. Platform Komersial dan Berbasis Cloud

- a. Google BigQuery: Gudang data cloud nirserver (serverless) yang sangat skalabel dan hemat biaya, mendukung kueri SQL pada set data yang besar.
- b. Amazon EMR/Redshift: Layanan terkelola di Amazon Web Services (AWS) untuk menjalankan kerangka kerja Big Data (EMR) dan solusi gudang data (Redshift).
- c. Microsoft Azure Synapse Analytics: Layanan analitik perusahaan yang menyatukan integrasi data, gudang data perusahaan, dan analitik data besar.
- d. Snowflake: Platform data cloud-native yang menawarkan gudang data fleksibel dan skalabel.
- e. Databricks: Platform terpadu untuk rekayasa data dan ilmu data, dibangun di atas Apache Spark.

3. Alat Analisis dan Visualisasi

Setelah data diproses dan disimpan, alat-alat ini membantu dalam analisis dan visualisasi(Rizal, 2019):

- a. Tableau: Alat visualisasi data yang kuat dengan dukungan komunitas yang besar, meskipun biaya lisensinya tinggi.
- b. Microsoft Power BI: Alat intelijen bisnis dengan fitur drag-and-drop yang mudah digunakan dan terintegrasi dengan ekosistem Microsoft.
- c. Python/R: Bahasa pemrograman serbaguna dengan banyak pustaka (library) untuk analisis