E-ISSN : 2988-1986
https://ejournal.cahayailmubangsa.institute/index.php/kohesi

Kohesi: Jurnal Multidisiplin Saintek
Volume 10 No 8 Tahun 2026

Evolusi dan Tantangan Platform Big Data Analytics: Perspektif Skalabilitas dan Tata Kelola Data Pasca-2020

Ahmeth Maulana Ishaq[1], Isa Faqihuddin Hanif[2], Mohammad Rafli Nugroho[3]
[1, 2] Sistem dan Teknologi Informasi, Fakultas Teknologi Industri dan Informatika
Universitas Muhammadiyah Prof. Dr. Hamka
maulanais860@gmail.com[1], isa@uhamka.ac.id[2] , bguys471@gmail.com[3]

ABSTRACT

This study analyzes the evolution of Big Data Analytics platforms in the post-2020 artificial intelligence era through a systematic library research approach. The fundamental transformation from monolithic on-premise Hadoop architectures to cloud-native data lake solutions presents a new paradigm in real-time data processing and storage efficiency. The adoption of compute-storage separation technology introduces scalability flexibility but increases network latency complexity. Data governance undergoes a significant shift toward Data Mesh principles with decentralized domain ownership, replacing the traditional centralized model. Privacy challenges, including compliance with global data protection regulations in a distributed environment, require the adoption of new privacy technologies. Performance optimization integrates AI-driven optimization to achieve operational cost efficiency in the modern data ecosystem. These findings provide strategic guidance for organizations in designing resilient and regulatory-compliant data architectures.

***Keywords**: big data, cloud-native analytics, data mesh, data governance, scalability*

ABSTRAK

Penelitian ini menganalisis evolusi platform Big Data Analytics dalam era kecerdasan buatan pasca-2020 melalui pendekatan systematic library research. Transformasi fundamental dari arsitektur on-premise Hadoop yang monolitik menuju solusi cloud-native data lakes menghadirkan paradigma baru dalam pemrosesan data real-time dan efisiensi penyimpanan. Adopsi teknologi pemisahan komputasi dan penyimpanan (compute-storage separation) memperkenalkan fleksibilitas skalabilitas namun meningkatkan kompleksitas latensi jaringan. Tata kelola data mengalami pergeseran signifikan menuju prinsip Data Mesh dengan desentralisasi kepemilikan domain, menggantikan model sentralisasi tradisional. Tantangan privasi meliputi kepatuhan terhadap regulasi perlindungan data global dalam lingkungan terdistribusi, memerlukan adopsi teknologi privasi baru. Pengoptimalan kinerja mengintegrasikan AI-driven optimization untuk mencapai efisiensi biaya operasional dalam ekosistem data modern. Temuan ini memberikan panduan strategis bagi organisasi dalam merancang arsitektur data yang tangguh dan patuh regulasi.

Kata Kunci: big data, cloud-native analytics, data mesh, tata kelola data, skalabilitas

Article History
Received: Januari 2026
Reviewed: Januari 2026
Published: Januari 2026

Plagiarism Checker No 235
Prefix DOI :
10.8734/Kohesi.v1i2.365
Copyright : Author
Publish by : Kohesi

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

PENDAHULUAN

Analisis terhadap literatur terkini menunjukkan bahwa transformasi fundamental ekosistem Big Data pasca-2020 telah mengubah paradigma tradisional dalam tiga aspek kritis: kecepatan pemrosesan (velocity), variasi data (variety), dan tata kelola (governance). Di era sebelumnya, fokus utama organisasi adalah semata-mata pada volume penyimpanan. Namun, evolusi ini tidak dapat dipisahkan dari konteks historis yang dimulai dari era Data Warehouse tradisional pada tahun 1990-an yang berfokus pada data terstruktur, berkembang menjadi ekosistem Hadoop pada dekade 2010-an untuk menangani data tidak terstruktur, hingga menjadi Cloud Data Platforms modern saat ini.

Pergeseran ini didorong oleh kegagalan arsitektur Hadoop on-premise dalam memenuhi tuntutan bisnis modern yang membutuhkan kelincahan (agility). Klaster Hadoop tradisional sering kali menghadapi masalah kompleksitas operasional, biaya perawatan yang tinggi, dan kesulitan dalam penskalaan sumber daya komputasi secara independen dari penyimpanan. Transformasi digital yang didorong oleh adopsi Artificial Intelligence (AI) telah mendorong evolusi lebih lanjut dari sistem analitik batch-oriented menuju arsitektur streaming yang lebih sophisticated.

Studi terbaru mengidentifikasi bahwa kebutuhan akan real-time decision making menjadi akselerator utama dalam adopsi arsitektur Lakehouse. Dalam paradigma ini, organisasi dipaksa untuk menyatukan keandalan dan struktur dari data warehouse dengan fleksibilitas dan skalabilitas biaya dari data lake. Fenomena ini juga diperumit dengan adanya regulasi data global yang ketat, menuntut arsitektur yang tidak hanya cepat tetapi juga governable (dapat dikelola) dan aman. Penelitian ini bertujuan untuk membedah evolusi tersebut secara komprehensif, menyoroti tantangan teknis dan manajerial yang muncul setelah tahun 2020.

TINJAUAN PUSTAKA

2.1 Evolusi Modern Data Stack

Konsep Modern Data Stack (MDS) merujuk pada rangkaian alat yang beroperasi secara cloud-native, modular, dan mudah diintegrasikan. Berbeda dengan legacy stack yang sering kali bersifat all-in-one (seperti suite ERP tradisional), MDS memisahkan fungsi ingest, storage, transformation, dan visualization. Jain & Nambiar (2023) mencatat bahwa tren ini didorong oleh kebutuhan untuk mengurangi vendor lock-in dan meningkatkan kecepatan deployment analitik.

2.2 Paradigma Lakehouse

Arsitektur Lakehouse muncul sebagai jawaban atas dikotomi antara Data Lake (murah, tidak terstruktur, rawan menjadi "data swamp") dan Data Warehouse (mahal, terstruktur, performa tinggi). Armbrust et al. (2021) mendefinisikan Lakehouse sebagai sistem yang mengimplementasikan struktur data dan manajemen transaksi (ACID) mirip warehouse langsung di atas penyimpanan objek cloud yang murah. Hal ini memungkinkan eksekusi beban kerja BI (Business Intelligence) dan ML (Machine Learning) dalam satu platform tunggal.

2.3 Data Mesh dan Desentralisasi

Sebagai antitesis dari pendekatan monolitik, Zhamak Dehghani (2022) memperkenalkan Data Mesh. Paradigma ini menolak sentralisasi data melalui satu tim teknis (data engineering team) yang sering menjadi bottleneck. Sebaliknya, Data Mesh mengusulkan arsitektur sosioteknikal yang mendesentralisasi kepemilikan data ke domain bisnis spesifik, memperlakukan data sebagai produk yang harus dikelola kualitasnya oleh pemilik domain tersebut.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan systematic library research sebagai metodologi utama untuk menganalisis evolusi arsitektur data secara kualitatif dan komparatif. Proses penelitian dirancang untuk memastikan cakupan literatur yang relevan dan mutakhir.

3.1 Prosedur Pengumpulan Data

Proses dimulai dengan identifikasi sumber literatur primer melalui database akademik bereputasi tinggi, yaitu IEEE Xplore, ACM Digital Library, ScienceDirect, dan Google Scholar. Fokus publikasi dibatasi pada rentang tahun 2020 hingga 2025 untuk menangkap fenomena pasca-pandemi dan ledakan AI generatif.

3.2 Kriteria Inklusi dan Eksklusi

Strategi pencarian menggunakan kombinasi kata kunci spesifik dan operator boolean, antara lain: "modern data stack", "lakehouse architecture", "compute-storage separation", "data mesh principles", dan "data governance challenges". Literatur yang dipilih harus membahas aspek teknis arsitektur atau aspek manajerial tata kelola data. Artikel yang hanya membahas algoritma analisis data tanpa menyinggung infrastruktur atau platform dieksklusi dari analisis.

3.3 Analisis Data

Data yang terkumpul dianalisis menggunakan metode sintesis tematik. Informasi dikategorikan ke dalam empat domain utama: Arsitektur Skalabilitas, Tata Kelola Terdesentralisasi, Keamanan Privasi, dan Optimasi Berbasis AI.

HASIL DAN PEMBAHASAN

4.1 Transformasi Arsitektur: Dari Monolitik Hadoop ke Modern Data Stack

Metamorfosis arsitektur data mendemonstrasikan pergeseran fundamental dari klaster Hadoop on-premise yang kaku menuju ekosistem Modern Data Stack (MDS) yang modular. Arsitektur Hadoop tradisional (HDFS) memiliki kelemahan fatal: coupled compute and storage. Artinya, untuk menambah kapasitas penyimpanan, organisasi juga harus menambah CPU dan RAM, yang sering kali mengakibatkan pemborosan sumber daya (over-provisioning) atau kekurangan daya (under-provisioning).

Evolusi pasca-2020 mengakar pada kebutuhan untuk mengakomodasi elastisitas sumber daya. Jika sistem tradisional mengandalkan coupled architecture, arsitektur modern memisahkan kedua lapisan tersebut (disaggregated architecture). Hal ini memungkinkan penskalaan independen yang signifikan secara ekonomi; penyimpanan dapat ditingkatkan hingga ukuran Petabyte di Cloud Object Storage (seperti Amazon S3 atau Google Cloud Storage) tanpa harus menyewa ribuan server komputasi secara permanen.

Namun, transformasi ini bukan tanpa tantangan. Pemisahan komputasi dan penyimpanan memperkenalkan latensi jaringan yang tidak ada pada sistem lokal. Untuk mengatasi ini, teknologi format tabel terbuka (open table formats) seperti Apache Iceberg, Delta Lake, dan Apache Hudi telah menjadi enabler utama. Teknologi ini menyediakan lapisan metadata cerdas yang memungkinkan mesin query untuk melakukan data skipping (hanya membaca file yang relevan), sehingga performa tetap tinggi meskipun data tersimpan di penyimpanan objek yang relatif lambat. Giebler et al. (2021) menekankan bahwa lapisan metadata ini adalah kunci untuk membawa kemampuan transaksional (ACID) ke dalam lingkungan data lake.

Aspek	Arsitektur Tradisional (Hadoop/Legacy DW)	Arsitektur Modern (Cloud-Native/Lakehouse)
Penyimpanan	Terikat dengan node komputasi (HDFS/SAN)	Terpisah, berbasis objek (S3/ADLS)
Skalabilitas	Vertikal atau Horizontal (berat)	Horizontal, Elastis, & Otomatis
Format Data	Proprietary atau File Mentah	Open Table Formats (Iceberg/Delta)

Biaya	CapEx tinggi (Investasi Awal)	OpEx (Pay-as-you-go)
Latensi	Rendah (Data lokal di disk)	Bervariasi (Tergantung jaringan & caching)

Tabel 1: Perbandingan Arsitektur Data Tradisional vs Modern

4.2 Kompleksitas Tata Kelola: Implementasi Paradigma Data Mesh

Transisi menuju arsitektur terdesentralisasi menghadirkan tantangan organisasional yang kompleks. Diidentifikasi bahwa manajemen data terpusat (centralized data engineering) sering kali menjadi bottleneck inovasi. Dalam model lama, tim bisnis harus mengajukan tiket ke tim data pusat hanya untuk mendapatkan laporan sederhana, menciptakan antrian pekerjaan yang panjang dan lambat.

Sebagai respons, konsep Data Mesh muncul sebagai solusi sosioteknikal, memperlakukan "data sebagai produk" dan mendesentralisasi kepemilikan ke domain bisnis masing-masing.

Namun, implementasi ini menghadapi tantangan budaya yang berat. Loshin (2024) mencatat bahwa pendekatan ini memerlukan perubahan pola pikir yang drastis; tim domain yang sebelumnya tidak teknis kini harus memiliki kapabilitas data. Tanpa federated governance yang ketat, desentralisasi ini berisiko menciptakan data silos baru yang terisolasi dan tidak konsisten.

4.3 Evolusi Keamanan dan Privasi Data dalam Lingkungan Terdistribusi

Lanskap keamanan dalam sistem Big Data terdistribusi telah bergeser dari sekadar kontrol akses berbasis peran (Role-Based Access Control - RBAC) statis menuju Attribute-Based Access Control (ABAC) yang dinamis. Dalam lingkungan cloud modern, data sering kali diakses oleh ribuan pengguna dan mesin dari berbagai lokasi, membuat RBAC tradisional sulit dikelola.

Tantangan menjadi semakin pelik dalam lingkungan di mana data tersebar di berbagai cloud region atau bahkan multi-cloud. Mekanisme enkripsi konvensional (enkripsi saat istirahat dan saat transit) tidak lagi cukup. Sidllo et al. (2022) menyoroti adopsi mekanisme kriptografi canggih seperti Enkripsi Homomorfik (Homomorphic Encryption) dan Privasi Diferensial (Differential Privacy). Teknologi ini memungkinkan analisis statistik dilakukan pada data terenkripsi tanpa perlu mendekripsinya terlebih dahulu, menjaga kerahasiaan data mentah dari analis data sekalipun.

Selain aspek teknis, aspek kepatuhan menjadi krusial. Kepatuhan terhadap regulasi seperti GDPR (Eropa) dan UU PDP (Indonesia) menjadi persyaratan arsitektural fundamental. Hal ini memunculkan konsep Data Sovereignty dalam arsitektur platform, di mana platform harus mampu melakukan geo-fencing data—memastikan data warga negara tertentu tidak pernah keluar dari wilayah yurisdiksi fisik negara tersebut meskipun diproses di cloud.

4.4 Perspektif Masa Depan: Konvergensi AI, MLOps, dan FinOps

Trajektori perkembangan sistem data menunjukkan konvergensi kuat antara Data Engineering dan Machine Learning Operations (MLOps). Kreuzberger et al. (2023) menjelaskan bahwa batas antara pipa data analitik dan pipa pembelajaran mesin semakin kabur. Platform data masa depan tidak hanya menyimpan data, tetapi juga menjadi tempat penyimpanan feature store dan vector embeddings untuk mendukung aplikasi AI Generatif.

Selain itu, platform diproyeksikan akan bersifat otonom (self-healing dan auto-optimizing). Nair & Gupta (2024) menemukan bahwa integrasi AI dalam query optimizer dapat secara signifikan mengurangi biaya komputasi. AI digunakan untuk memprediksi kebutuhan komputasi (predictive scaling), merekomendasikan partisi data yang optimal, dan bahkan menulis ulang query SQL yang tidak efisien secara otomatis.

Terakhir, aspek FinOps (Financial Operations) menjadi komponen integral. Dengan model biaya pay-as-you-go, inefisiensi kode sekecil apapun dapat menyebabkan lonjakan biaya yang masif. Oleh karena itu, observability biaya dan atribusi biaya ke level unit bisnis menjadi fitur

wajib dalam platform data pasca-2020. Pengoptimalan kinerja kini tidak hanya soal kecepatan waktu eksekusi, tetapi juga efisiensi biaya per query.

KESIMPULAN

Investigasi komprehensif terhadap evolusi platform Big Data mengindikasikan transformasi paradigmatik dari infrastruktur berat berbasis Hadoop menuju arsitektur cloud-native yang fleksibel, modular, dan terpisah antara komputasi dan penyimpanan. Metamorfosis ini dikatalisasi oleh kebutuhan bisnis akan analitik real-time, efisiensi biaya, dan dukungan terhadap beban kerja AI.

Meskipun menawarkan skalabilitas superior, arsitektur baru ini menghadirkan tantangan kompleks yang belum sepenuhnya terpecahkan, terutama dalam hal tata kelola terdesentralisasi (Data Mesh) yang membutuhkan kedewasaan budaya organisasi. Selain itu, kepatuhan privasi lintas yurisdiksi dalam lingkungan terdistribusi memerlukan adopsi teknologi privasi baru seperti enkripsi homomorfik.

Ke depan, integrasi AI dalam manajemen infrastruktur data akan menjadi kunci keberhasilan. Platform yang mampu melakukan self-optimization dan menyediakan tata kelola otomatis akan mendominasi pasar, memungkinkan organisasi untuk fokus pada nilai bisnis data daripada kerumitan infrastruktur. Organisasi disarankan untuk mulai mengadopsi prinsip desentralisasi data secara bertahap dan berinvestasi pada platform yang mendukung standar terbuka untuk menjamin keberlanjutan jangka panjang.

REFERENSI

Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021). Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Data Lakes. Proceedings of the 11th Conference on Innovative Data Systems Research (CIDR).

Dehghani, Z. (2022). Data Mesh: Delivering Data-Driven Value at Scale. O'Reilly Media.

Giebler, C., Gröger, C., Hoos, E., Schwarz, H., & Mitschang, B. (2021). The Data Lake Architecture Framework: A Comprehensive Overview of Layers and Components. Enterprise Modelling and Information Systems Architectures, 16(1). https://doi.org/10.18417/emisa.16.5

Jain, A., & Nambiar, R. (2023). The Evolution of Modern Data Stack: Trends and Challenges in Cloud-Native Analytics. International Journal of Information Management Data Insights, 3(2), 100-112.

Kreuzberger, D., Kühl, N., & Kollmann, S. (2023). Machine Learning Operations (MLOps): Overview, Definition, and Architecture. IEEE Access, 11, 31866-31879. https://doi.org/10.1109/ACCESS.2023.3262138

Loshin, D. (2024). Data Governance for the Modern Data Ecosystem: Ensuring Trust in Distributed Environments. Morgan Kaufmann.

Nair, M., & Gupta, S. (2024). AI-Driven Query Optimization in Distributed Data Systems: A Comparative Study. Journal of Big Data, 11(1), 45-62.

Sidllo, C., et al. (2022). Privacy-Preserving Big Data Analytics: A Comprehensive Survey of Homomorphic Encryption and Differential Privacy. IEEE Communications Surveys & Tutorials, 24(3).