Available online at https://ejournal.upnvj.ac.id/jsia
Jurnal Sistem Informasi dan Aplikasi
Volume 2 | Issue 2 | bulan DESEMBER (2024)
e-issn : 3025 – 9347
JSIA
Jurnal Sistem Informasi & Aplikasi

PENGEMBANGAN ARSITEKTUR DATA LAKE UNTUK MENGELOLA DATA TIDAK TERSTRUKTUR DALAM EKOSISTEM BIG DATA

Andrea Berliani Yoshita[1], Tjahjanto[2], Widya KhafaNofa[3]
[1]Program Studi Sistem Informasi, Fakultas Ilmu Komputer, UPN Veteran Jakarta
[2]Program Studi Magister Ilmu Komputer, Program Pascasarjana, Universitas Budi Luhur
2210512021@mahasiswa.upnvj.ac.id[1], tjahjanto@upnvj.ac.id[2], w_khafa@staff.gunadarma.ac.id[3]
Jl. Rs. Fatmawati, Pondok Labu, Jakarta Selatan, DKI Jakarta, 12450, Indonesia

Keywords:

Data Lake, Big Data, Unstructured Data, Data Architecture, Data Management, Literature Study

Abstract

Data Lake architecture has become an innovative solution to manage the rapidly growing unstructured data in the Big Data ecosystem. This research aims to develop an effective Data Lake architecture framework, with a primary focus on the integration and management of unstructured data. The research method used is a literature study, covering the latest journals and scientific articles since 2020. The purpose of this literature study is to identify key components and best practices in Data Lake implementation, so as to provide clear guidance for organizations to optimize the use of unstructured data.The results show that an effective Data Lake architecture should include several important components such as ingestion, storage, processing, and governance to achieve optimal efficiency and scalability. Ingestion ensures that data from various sources can be collected and integrated easily. Storage provides secure and scalable storage for data in various formats. Processing enables real-time or batch processing of data according to analysis needs. Governance ensures that data is well managed, meeting quality, security, and compliance standards. The combination of these components helps organizations maximize the value of their data and overcome the challenges of managing unstructured data.

Kata Kunci:

Data Lake, Big Data, Data Tidak Terstruktur, Arsitektur Data, Pengelolaan Data, Studi Literatur

Abstrak

Arsitektur Data Lake telah menjadi solusi inovatif untuk mengelola data tidak terstruktur yang berkembang pesat dalam ekosistem Big Data. Penelitian ini bertujuan untuk mengembangkan kerangka kerja arsitektur Data Lake yang efektif, dengan fokus utama pada integrasi dan pengelolaan data tidak terstruktur. Metode penelitian yang digunakan adalah studi literatur, mencakup jurnal dan artikel ilmiah terbaru sejak tahun 2020. Tujuan dari studi literatur ini adalah untuk mengidentifikasi komponen utama dan praktik terbaik dalam implementasi Data Lake, sehingga dapat memberikan panduan yang jelas bagi organisasi untuk mengoptimalkan penggunaan data tidak terstruktur. Hasil penelitian menunjukkan bahwa arsitektur Data Lake yang efektif harus mencakup beberapa komponen penting seperti ingestion, storage, processing, dan governance untuk mencapai efisiensi dan skalabilitas yang optimal. Ingestion memastikan bahwa data dari berbagai sumber dapat dikumpulkan dan diintegrasikan dengan mudah. Storage menyediakan penyimpanan yang aman dan scalable untuk data dalam berbagai format. Processing memungkinkan pemrosesan data secara real-time atau batch sesuai kebutuhan analisis. Governance menjamin bahwa data dikelola dengan baik, memenuhi standar kualitas, keamanan, dan kepatuhan. Kombinasi dari komponen-komponen ini membantu organisasi dalam memaksimalkan nilai dari data yang

mereka miliki, serta mengatasi tantangan dalam pengelolaan data tidak terstruktur.

1. Pendahuluan

Perkembangan teknologi informasi telah mendorong pertumbuhan data dengan sangat cepat, terutama data tidak terstruktur seperti teks, gambar, dan video. Dalam konteks ekosistem Big Data, pengelolaan data tidak terstruktur menjadi tantangan besar karena kompleksitas dan volume yang besar. Data tidak terstruktur sering kali sulit dianalisis dan diintegrasikan ke dalam sistem tradisional yang umumnya dirancang untuk menangani data terstruktur dengan skema kaku. Oleh karena itu, diperlukan solusi inovatif untuk menyimpan, mengelola, dan memproses data dalam bentuk aslinya tanpa melalui proses transformasi yang memakan waktu dan biaya. Salah satu solusi yang muncul untuk mengatasi tantangan ini adalah Data Lake.

Data Lake memungkinkan penyimpanan data dalam bentuk mentah, sehingga organisasi dapat menyimpan berbagai jenis data tanpa memerlukan skema yang ketat saat penyimpanan. Hal ini memberikan fleksibilitas tinggi dalam pengelolaan data, memudahkan integrasi dan analisis data tidak terstruktur dari berbagai sumber. Penelitian ini bertujuan mengembangkan arsitektur Data Lake yang dapat mengatasi tantangan pengelolaan data tidak terstruktur, serta menyediakan kerangka kerja yang dapat diterapkan dalam berbagai skenario bisnis. Selain itu, pentingnya kolaborasi antara departemen TI dan bisnis dalam merancang dan mengimplementasikan Data Lake juga semakin diakui, karena hal ini dapat memastikan bahwa kebutuhan analitik dan operasional organisasi dapat terpenuhi dengan lebih baik [1]. Dengan arsitektur yang tepat, Data Lake dapat menyederhanakan proses penyimpanan dan manajemen data, meningkatkan kemampuan organisasi dalam analisis data yang lebih mendalam dan real-time, sehingga menghasilkan wawasan yang lebih berharga untuk pengambilan keputusan strategis [2].

2. Metode Penelitian

Penelitian ini menggunakan metode studi literatur review untuk mengumpulkan dan menganalisis informasi dari jurnal-jurnal dan artikel ilmiah terbaru yang diterbitkan sejak tahun 2020. Pendekatan ini dipilih karena memungkinkan peneliti untuk mendapatkan pemahaman yang mendalam tentang perkembangan terkini dalam arsitektur Data Lake. Selain itu, metode ini memungkinkan pengumpulan berbagai perspektif dan temuan dari sejumlah besar studi yang telah dilakukan sebelumnya, yang membantu dalam membangun landasan teori yang kuat. Dengan demikian, penelitian ini dapat memberikan gambaran komprehensif mengenai tren terbaru dan inovasi dalam pengelolaan Data Lake, serta mengidentifikasi tantangan dan peluang yang muncul dalam pengembangan arsitektur tersebut [3].

Selanjutnya, artikel dan jurnal yang relevan diidentifikasi, dievaluasi, dan disintesis untuk membangun landasan teoritis dan praktis bagi pengembangan arsitektur yang diusulkan. Proses evaluasi dilakukan dengan mempertimbangkan kualitas dan relevansi sumber-sumber tersebut terhadap topik penelitian. Dengan menyatukan berbagai temuan dari literatur yang ada, penelitian ini berusaha untuk merumuskan praktik terbaik dalam pengelolaan data tidak terstruktur, yang merupakan komponen penting dari Data Lake. Hasil sintesis ini diharapkan dapat memberikan panduan praktis bagi para profesional dan akademisi dalam merancang dan mengimplementasikan arsitektur Data Lake yang efektif dan efisien, sesuai dengan kebutuhan dan perkembangan teknologi saat ini.

3. Hasil dan Pembahasan

Komponen Kunci Arsitektur Data Lake

Proses ingestion harus mampu menangani berbagai jenis data tidak terstruktur dari berbagai sumber secara real-time maupun batch. Alat seperti Apache NiFi dan Kafka sering digunakan untuk tujuan ini. Apache NiFi menyediakan kemampuan untuk mengotomatisasi aliran data antara sistem yang berbeda dengan mudah, sedangkan Kafka memungkinkan pengumpulan, penyimpanan, dan pemrosesan aliran data dalam skala besar secara real-time [4]. Penggunaan kedua alat ini memastikan bahwa data dapat diterima dan diproses dengan cepat dan efisien, memungkinkan perusahaan untuk segera mendapatkan wawasan dari data yang diterima.

Penyimpanan data tidak terstruktur memerlukan solusi yang scalable dan cost effective. Amazon S3 dan Hadoop Distributed File System (HDFS) adalah beberapa opsi yang populer dalam hal ini. Amazon S3 menawarkan penyimpanan objek yang sangat scalable dengan biaya yang relatif rendah, serta integrasi yang kuat dengan berbagai alat analitik dan machine learning [5]. Di sisi lain, HDFS menyediakan sistem file terdistribusi yang dirancang untuk menyimpan data dalam jumlah besar dengan tingkat keandalan yang tinggi, memungkinkan pemrosesan paralel yang efisien melalui ekosistem Hadoop.

Pemrosesan data tidak terstruktur melibatkan transformasi dan analisis data. Teknologi seperti Apache Spark dan Hadoop MapReduce digunakan untuk memproses data dalam skala besar, memberikan kemampuan untuk menganalisis data dalam waktu yang lebih singkat dan dengan kinerja yang lebih baik [6]. Dalam aspek tata kelola data, manajemen metadata, keamanan, dan kepatuhan merupakan elemen penting. Solusi seperti Apache Atlas dan AWS Glue dapat membantu dalam pengelolaan metadata dan data lineage, memastikan bahwa data dikelola dengan baik dan sesuai dengan regulasi yang berlaku [7]. Kombinasi alat-alat ini memberikan kerangka kerja yang komprehensif untuk menangani, menyimpan, memproses, dan mengelola data tidak terstruktur dengan efektif.

Tantangan Dan Solusi

Implementasi Data Lake menghadirkan beberapa tantangan utama, termasuk integrasi data, keamanan, dan tata kelola. Tantangan integrasi data mencakup kesulitan dalam menggabungkan data dari berbagai sumber yang beragam dan seringkali tidak terstruktur. Keamanan juga menjadi isu kritis, mengingat potensi risiko kebocoran data dan akses tidak sah. Untuk mengatasi tantangan-tantangan ini, solusi seperti enkripsi data dapat memastikan bahwa data tetap terlindungi selama proses penyimpanan dan transmisi. Selain itu, kontrol akses berbasis peran (RBAC) dapat diterapkan untuk memastikan bahwa hanya pengguna yang berwenang yang dapat mengakses data tertentu, memperkuat keamanan dan kepatuhan terhadap peraturan. Pemantauan real-time juga menjadi elemen kunci dalam tata kelola, memungkinkan deteksi dini anomali dan respons cepat terhadap potensi ancaman, sehingga meningkatkan efektivitas pengelolaan data secara keseluruhan [8].

4. Kesimpulan dan Saran

Penelitian ini menyoroti pentingnya pengembangan arsitektur Data Lake yang komprehensif untuk mengelola data tidak terstruktur dalam ekosistem Big Data. Arsitektur yang efektif harus mencakup proses ingestion, storage, processing, dan governance yang kuat. Proses ingestion memastikan bahwa data yang beragam dan berukuran besar dapat diintegrasikan dengan mulus ke dalam Data Lake. Storage yang efisien diperlukan untuk menyimpan data dengan aman dan dapat diakses dengan cepat, sementara processing yang canggih memungkinkan analisis data

secara real-time atau batch. Governance yang kuat sangat penting untuk menjamin kualitas, keamanan, dan kepatuhan data, sehingga organisasi dapat menghindari risiko terkait data yang tidak terstruktur.

Dengan mengadopsi praktik terbaik dan teknologi terbaru, organisasi dapat meningkatkan efisiensi pengelolaan data dan mendapatkan wawasan berharga dari data tidak terstruktur. Praktik terbaik mencakup penggunaan alat dan metodologi yang tepat untuk setiap tahap dalam arsitektur Data Lake, seperti penggunaan format penyimpanan yang optimal dan penerapan strategi pengelolaan metadata yang efektif. Teknologi terbaru, seperti kecerdasan buatan dan machine learning, dapat diterapkan untuk meningkatkan kapabilitas analitik dan menghasilkan wawasan yang lebih mendalam dari data tidak terstruktur. Dengan demikian, organisasi dapat membuat keputusan yang lebih informasional dan strategis, yang pada akhirnya meningkatkan daya saing dan inovasi dalam pasar yang semakin kompetitif.

Referensi

[1] Wibowo, H. (2022). Pendekatan Kolaboratif dalam Implementasi Data Lake. *Jurnal Sistem Informasi Indonesia*, 9(4), 98-110.

[2] Smith, J. (2020). "Data Lake Architecture for Managing Unstructured Data in Big Data Ecosystem." *Journal of Information Technology and Management*, 32(2), 123-135.

[3] Chen, J., Zhang, X., & Lee, W. (2021). Scalable Data Processing with Apache Spark. *Journal of Big Data Analytics*, 8(2), 123-137.

[4] Garcia-Molina, H., Ullman, J. D., & Widom, J. (2021). Data Governance in Modern Data Lakes. *Journal of Data Management*, 12(1), 45-59.

[5] Khan, M., Kumar, A., & Singh, R. (2021). Real-time Data Ingestion Techniques in Big Data Systems. *International Journal of Data Engineering*, 9(4), 210-223.

[6] Smith, T., Brown, L., & Zhao, Y. (2022). Metadata Management for Data Lakes. *IEEE Transactions on Knowledge and Data Engineering*, 34(3), 567-579.

[7] Hartono, A. (2021). Analisis Komprehensif Terhadap Praktik Terbaik dalam Pengelolaan Data Tidak Terstruktur Menggunakan Data Lake. *Jurnal Sistem Informasi Indonesia*, 8(3), 112-125.

[8] Zhang, Q., Wang, H., & Chen, Y. (2020). Cost-effective Storage Solutions for Big Data. *ACM Computing Surveys*, 53(5), 89-103.