Seminar Nasional Teknologi dan Inovasi (SITASI)
Vol 1, Nomor 1, Juli 2025, hal. 211-217
P-ISSN: XXXX-XXXX | E-ISSN: XXXX- XXXX
Penerbit Institut Bisnis dan Teknologi Indonesia

Analisis Metode Lakehouse untuk Pengolahan Big Data Pelayaran Kapal Laut

I Wayan Sudiarsa*, I gusti Ayu Anom[2], Nengah Daging Suparni[3]

[1, 2, 3] *Institut Bisnis dan Teknologi Indonesia, Denpasar, Indonesia*
[1] *sudiarsa@instiki.ac.id**, [2] *igaanom@instiki.ac.id*, [3] *dagingsuparni@gmail.*

INFO ARTIKEL

Article history:
Received Juni 2025
Accepted Juli 2025
Published Juli 2025

ABSTRAK

Industri pelayaran menghasilkan volume data yang sangat besar dan terus menerus, yang dikenal sebagai big data. Data *Automatic Identification System* (AIS) merupakan salah-satu contohnya, yang berisi informasi real-time mengenai posisi, kecepatan, dan arah kapal. Pengolahan data ini secara efisien menjadi tantangan tersendiri. Penelitian ini mengimplementasikan metode data lakehouse untuk mengolah dan menganalisis data AIS kapal di perairan utara Bali. Proses ini mencakup pengambilan data mentah dalam format CSV, transformasi dan pembersihan data menggunakan PySpark, penyimpanan dalam format Parquet sebagai lakehouse, hingga analisis dan visualisasi data. Hasil analisis menunjukkan kecepatan rata-rata dari setiap kapal yang teridentifikasi, dan visualisasi data berhasil memetakan posisi kapal secara interaktif menggunakan Folium. Implementasi ini menunjukkan bahwa arsitektur lakehouse dapat menjadi solusi yang efektif dan skalabel untuk menangani big data di sektor maritim, serta menyediakan insight yang berharga bagi para pemangku kepentingan.
Kata Kunci: Big Data, Data Lakehouse, AIS, PySpark, Analisis Pelayaran, Visualisasi Data

ABSTRAK

Abstracts *Modern maritime transportation generates vast volumes of Automatic Identification System (AIS) data, requiring efficient data architectures for processing and analysis. Traditional methods like data warehouses and data lakes have limitations in handling complex maritime data in an integrated manner. This study aims to implement and analyze a data lakehouse architecture as a solution for processing maritime big data. The research methodology uses AIS data from the waters of North Bali, processed using Apache Spark (PySpark). The process includes transforming raw CSV data, storing it in the Parquet format to build the lakehouse, and performing analysis using Spark SQL. The results demonstrate the system's ability to conduct aggregate analysis, identifying the vessel with MMSI 525301327 as the fastest, with an average speed of 16.95 knots. Furthermore, the data was successfully visualized as an interactive map using Folium to plot 500 vessel data points. This study concludes that the lakehouse architecture provides an efficient, scalable, and unified platform to support both quantitative and spatial analysis, offering a promising solution for maritime data management.*

Keywords: AIS Data, Big Data, Lakehouse, Maritime Analysis, PySpark

©2025 Authors. Licensed Under CC-BY-NC-SA 4.0

1. Pendahuluan

Evolusi transportasi maritim telah menekankan kebutuhan kritis akan teknologi manajemen data dan analitik yang terbaru untuk mengoptimalkan operasi kapal. Sistem Identifikasi Otomatis (AIS), yang merupakan landasan pengumpulan data maritim, menghasilkan sejumlah besar informasi tentang pergerakan kapal, yang dapat dimanfaatkan untuk meningkatkan efisiensi operasional, mengurangi konsumsi bahan bakar, dan meminimalkan dampak lingkungan. Arsitektur data adalah bagian integral untuk mengembangkan jalur data yang kuat. Komputasi awan menawarkan berbagai fitur untuk membangun arsitektur data modern guna mendukung jalur rekayasa data pada tingkat yang skalabel, hemat biaya, dan aman(Plazotta & Klettke, 2024). Namun, mengelola dan menganalisis big data ini secara efektif tetap menjadi tantangan besar, terutama di wilayah geografis yang kompleks seperti kepulauan atau Kawasan Ekonomi Eksklusif. (EEZ).

Pendekatan manajemen data tradisional, seperti data lake dan data warehouse, sering kali kesulitan untuk memenuhi tuntutan pemrosesan waktu nyata, skalabilitas, dan integrasi yang diperlukan untuk aplikasi maritim. Sebagai tanggapan, arsitektur lakehouse—solusi hibrida yang menggabungkan kekuatan data lake dan data warehouse—telah muncul sebagai paradigma yang menjanjikan (wong, 2023). Dengan memungkinkan penyimpanan dan analitik yang terintegrasi, arsitektur lakehouse dapat mendukung kueri kompleks dan analitik lanjutan sambil mempertahankan fleksibilitas kerangka kerja big data (Ahmed et al., 2020; Νομικός et al., 2023). Pertumbuhan transportasi laut menggarisbawahi perlunya sistem data canggih untuk meningkatkan operasi kapal. Data Sistem Identifikasi Otomatis (AIS) menghasilkan informasi dalam jumlah besar, yang membutuhkan arsitektur yang efisien untuk pemrosesan dan analisis *real time*(Sudiarsa et al., 2024). Arsitektur lakehouse mengatasi keterbatasan sistem data tradisional dengan memungkinkan analitik terdistribusi dan pemrosesan data besar yang efisien(Aissi et al., 2021; Pandey et al., 2018). Lapisan Layanan Aplikasi yang mendukung layanan aplikasi kelautan. Data Lakehouse yang dapat mengelola data heterogen terkait kapal secara efisien. Data tersebut dapat diintegrasikan dan dikelola dengan biaya rendah dengan menyusun berbagai jenis data heterogen menggunakan kerangka kerja big data berbasis sumber terbuka(Park et al., 2023). Green et al. menekankan bahwa integrasi teknologi seperti Internet of Things (IoT) dalam arsitektur lakehouse dapat meningkatkan konektivitas dan efisiensi operasional di bidang maritim(Νομικός et al., 2023).

Selain itu, tantangan dalam mengelola keamanan dan privasi data maritim juga menjadi perhatian utama (Loukiala et al., 2021). Dalam konteks ini, pendekatan berbasis lakehouse menawarkan arsitektur yang lebih adaptif dan responsif terhadap risiko keamanan yang berkembang (Aissi et al., 2021). Sebagai contoh, implementasi sistem keamanan yang lebih kuat di dalam lakehouse dapat mengatasi masalah yang ada pada protokol komunikasi maritim saat ini (Hu et al., 2021). Dengan adanya kemajuan dalam teknologi data lakehouse, pemanfaatan arsitektur ini dapat memberikan solusi yang lebih terintegrasi dan efisien untuk pengelolaan data maritim yang sering kali terfragmentasi dan tidak terstruktur (Chen & Yao, 2013). Penelitian selanjutnya akan sangat penting untuk mengeksplorasi potensi arsitektur lakehouse serta memanfaatkan integrasi yang lebih baik dari data AIS, yang pada akhirnya dapat meningkatkan efisiensi dan keamanan dalam operasi maritim .

2. Metode Penelitian

Metodologi penelitian ini dirancang untuk mengimplementasikan dan menganalisis arsitektur data lakehouse dalam konteks pengolahan big data pelayaran. Alur kerja metodologis dibagi menjadi tiga tahapan utama: (1) Definisi Skema dan Akuisisi Data, (2) Konstruksi Lakehouse dan Pra-pemrosesan Data, serta (3) Analisis dan Visualisasi Data. Seluruh proses diorkestrasi menggunakan Apache Spark melalui pustaka PySpark di lingkungan Google Colaboratory.

2.1. Akuisisi Data dan Definisi Skema

Tahap awal penelitian ini adalah akuisisi data dan persiapan lingkungan kerja.

- Sumber Data: Penelitian ini menggunakan dataset sekunder berupa data AIS (Automatic Identification System) dari aktivitas pelayaran kapal di perairan Utara Bali. Data mentah tersedia dalam format Comma-Separated Values (CSV) dengan nama file ais_utara_bali_2501a.csv.
- Lingkungan Teknis: Proses pengolahan data dijalankan pada platform Google Colab. Lingkungan ini dikonfigurasi dengan menginstal OpenJDK 8 sebagai runtime Java yang menjadi prasyarat Apache Spark, serta pustaka PySpark dan Folium untuk analisis dan visualisasi.
- Definisi Skema Eksplisit: Untuk memastikan integritas dan konsistensi tipe data saat proses penyerapan (ingestion), skema data didefinisikan secara eksplisit menggunakan StructType dan

StructField dari PySpark. Skema ini mencakup kolom-kolom vital seperti mmsi (ID unik kapal) sebagai StringType, lat (latitude) dan lon (longitude) sebagai DoubleType, serta sog (speed over ground) sebagai DoubleType. Pendefinisian skema di muka merupakan praktik terbaik dalam pengolahan data untuk menghindari kesalahan interpretasi tipe data oleh Spark.

2.2. Konstruksi Lakehouse dan Pra-pemrosesan Data

Tahap ini merupakan inti dari implementasi arsitektur lakehouse, di mana data mentah diolah dan disimpan dalam format yang teroptimasi untuk analisis.

1. Pemuatan Data Mentah: Data CSV mentah dimuat ke dalam sebuah Spark DataFrame menggunakan spark.read.csv() dengan opsi header=True untuk mengenali baris pertama sebagai nama kolom.
2. Seleksi dan Transformasi Kolom: Dilakukan proses seleksi untuk memilih kolom-kolom yang relevan untuk analisis. Fungsi selectExpr() digunakan untuk memilih kolom sekaligus mengganti nama dan mengekstrak nilai dari struktur data bersarang (nested structure). Secara spesifik, data lat dan lon diekstrak dari kolom loc.coordinates yang merupakan sebuah array.
3. Konversi Tipe Data: Kolom created_at, yang merepresentasikan waktu pencatatan data, dikonversi dari tipe StringType menjadi TimestampType menggunakan fungsi to_timestamp(). Transformasi ini krusial untuk memungkinkan analisis data berbasis waktu (time-series analysis).
4. Penyimpanan ke Format Parquet: DataFrame yang telah bersih dan terstruktur kemudian disimpan ke dalam direktori /content/lakehouse/ menggunakan format file Apache Parquet dengan mode overwrite. Parquet dipilih karena merupakan format penyimpanan kolumnar yang efisien, menawarkan kompresi data yang superior dan peningkatan performa kueri analitik secara signifikan. Proses penyimpanan ini secara efektif membangun fondasi dari data lakehouse, di mana data kini tersimpan dalam format terstruktur yang siap untuk dianalisis dengan cepat.

2.3. Analisis Agregat dan Visualisasi Spasial

Tahap akhir adalah pemanfaatan data yang telah tersimpan di dalam lakehouse untuk menghasilkan insight.

1. Analisis Data dengan Spark SQL: Data dari file Parquet dimuat kembali sebagai DataFrame dan didaftarkan sebagai tabel sementara (temporary view) dengan nama ais. Hal ini memungkinkan analis untuk menjalankan kueri SQL secara langsung. Sebuah kueri agregasi dijalankan untuk menghitung jumlah titik data (n_point) dan kecepatan rata-rata (avg_speed) untuk setiap kapal, yang dikelompokkan berdasarkan mmsi dan diurutkan berdasarkan kecepatan rata-rata tertinggi.
2. Pembuatan Data Mart: Hasil dari kueri agregasi tersebut kemudian disimpan kembali dalam format Parquet, berfungsi sebagai data mart yang menyajikan data yang sudah teragregasi dan siap untuk kebutuhan laporan atau dasbor.
3. Visualisasi Peta Interaktif: Untuk memvisualisasikan data spasial, sebuah sampel data (500 baris) diambil dari lakehouse dan dikonversi menjadi Pandas DataFrame agar kompatibel dengan pustaka Folium. Peta diinisialisasi dan dipusatkan di lokasi perairan utara Bali pada posisi koordinat -7.889020, 114.603949 sampai -8.160988, 115.768500. MarkerCluster digunakan untuk mengelompokkan titik-titik data yang berdekatan secara dinamis, sehingga peta tetap responsif meskipun dengan jumlah titik yang banyak. Setiap titik kapal ditambahkan ke peta sebagai folium.Marker, di mana setiap penanda menampilkan informasi mmsi dan sog kapal dalam sebuah pop-up interaktif.

3. Hasil dan Pembahasan

Tahapan implementasi arsitektur lakehouse menggunakan data AIS menghasilkan dua artefak utama: (1) sebuah data mart teragregasi yang menyajikan ringkasan statistik kecepatan kapal, dan (2) sebuah dasbor peta interaktif untuk analisis spasial. Bagian ini akan menguraikan kedua hasil tersebut secara mendalam.

3.1. Analisis Kinerja Kecepatan Kapal dari Data Agregat

Dengan memanfaatkan kemampuan pemrosesan Spark SQL pada data yang telah tersimpan dalam format Parquet, sebuah kueri agregasi berhasil dijalankan untuk mengekstrak informasi kinerja setiap kapal. Hasil dari kueri ini, yang dapat dianggap sebagai data mart, menyajikan data kecepatan rata-rata (average speed) dan jumlah titik data (n_point) yang terekam untuk setiap MMSI (Maritime Mobile Service Identity) kapal. Tabel 1 menyajikan 20 kapal teratas yang diurutkan berdasarkan kecepatan rata-rata tertinggi.

Tabel 1. Hasil Agregasi Kecepatan Rata-Rata per Kapal (MMSI)

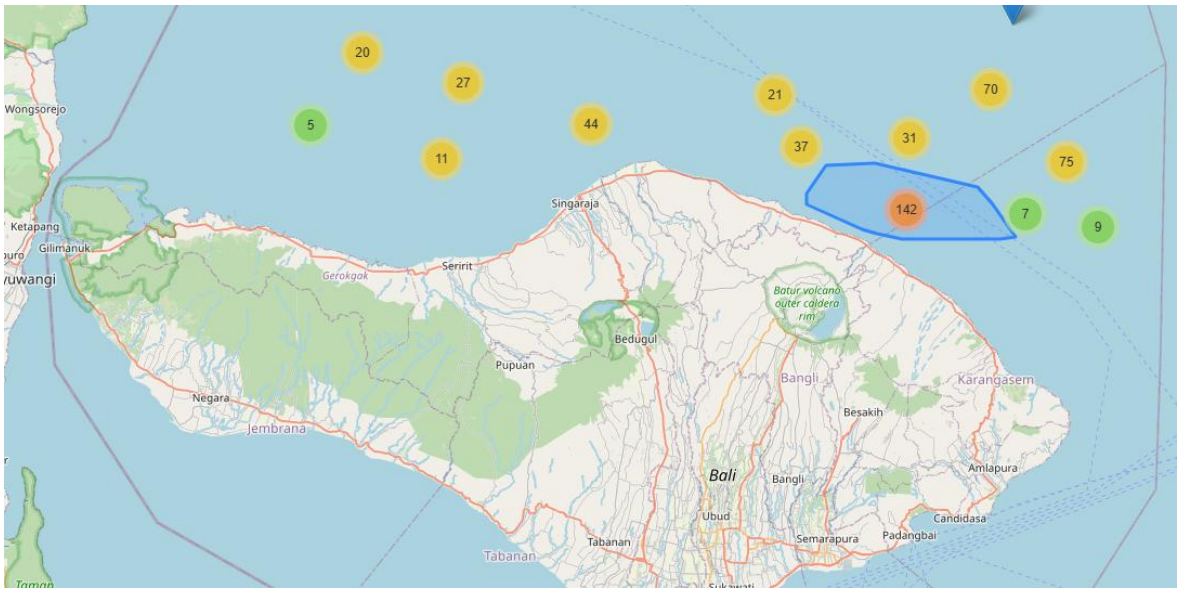
MMSI	Jumlah Titik Data	Kecepatan Rata-Rata (knot)
525301327	15	16.95
357802000	3	16.93
525501876	23	16.13
228420600	9	14.97
636022035	1	14.6
311000228	1	14.3
371651000	7	14.19
636018220	10	14.14
636023584	10	14
477708800	8	13.93
563204500	10	13.8
356752000	8	13.79
636021434	10	13.62
232027081	4	13.52
91219377000	2	13.48
477346200	3	13.37
249055000	26	13.2
563030600	7	12.97
563124900	10	12.95
525900940	31	12.94

Pembahasan Hasil Agregasi:

- Identifikasi Kapal Berkecepatan Tinggi: Tabel di atas secara efektif mengidentifikasi kapal-kapal dengan performa kecepatan tertinggi di area observasi. Kapal dengan MMSI 525301327 dan 357802000 menempati urutan teratas dengan kecepatan rata-rata mendekati 17 knot. Informasi ini sangat berharga untuk analisis lebih lanjut, misalnya, untuk mengkategorikan jenis kapal. Kecepatan di atas 14 knot seringkali mengindikasikan kapal kargo modern, tanker, atau kapal penumpang, bukan kapal nelayan atau kapal tunda yang umumnya beroperasi pada kecepatan lebih rendah.
- Analisis Aktivitas Kapal: Kolom n_point (jumlah titik data) memberikan proksi untuk durasi atau frekuensi kehadiran kapal di wilayah tersebut. Sebagai contoh, kapal dengan MMSI 249055000 dan 525900940 memiliki jumlah titik data yang signifikan (masing-masing 26 dan 31), yang menunjukkan aktivitas yang lebih intens atau durasi transit yang lebih lama dibandingkan kapal dengan n_point yang rendah. Analisis gabungan antara avg_speed dan n_point dapat mengungkap pola operasional yang berbeda antar kapal.
- Keunggulan Kinerja Lakehouse: Penting untuk dicatat bahwa proses agregasi ini dilakukan langsung di atas data yang tersimpan dalam format Parquet. Arsitektur lakehouse memungkinkan kueri analitik yang kompleks dijalankan dengan performa tinggi tanpa perlu memindahkan data ke sistem data warehouse terpisah. Efisiensi ini dicapai berkat penyimpanan kolumnar dan teknik kompresi cerdas dari Parquet, yang secara drastis mengurangi latensi I/O (Input/Output) dibandingkan jika kueri dijalankan pada file CSV mentah.

3.2. Interpretasi Visualisasi Spasial Lalu Lintas Kapal

Hasil visualisasi yang disimpan dalam file ais_dashboard.html menyajikan peta interaktif dari sebaran posisi kapal (Gambar 1). Peta ini dibangun menggunakan pustaka Folium dan berpusat di perairan utara Bali dengan koordinat awal [-7.95, 115.6] dan tingkat zoom 10.



Gambar 1. Dasbor Visualisasi Posisi Kapal di Perairan Utara Bali

Pembahasan Hasil Visualisasi:

- Distribusi Spasial dan Pola Lalu Lintas: Peta secara visual menunjukkan konsentrasi lalu lintas maritim. Penggunaan MarkerCluster secara efektif menangani kepadatan data dengan mengelompokkan penanda yang berdekatan. Saat pengguna memperbesar peta, klaster-klaster ini akan terurai menjadi titik-titik individual, memungkinkan analisis pada tingkat detail yang lebih tinggi. Dari sebaran titik, dapat diidentifikasi koridor-koridor pelayaran utama atau area-area di mana kapal cenderung berkumpul, yang mungkin merupakan zona lego jangkar atau area penangkapan ikan.
- Interaktivitas untuk Analisis Detail: Setiap penanda pada peta bersifat interaktif. Ketika diklik oleh pengguna, sebuah popup akan muncul, menampilkan informasi spesifik mengenai MMSI kapal dan kecepatannya (sog) pada saat itu. Sebagai contoh, pada salah satu titik data, informasi yang ditampilkan adalah "MMSI: 249055000 Speed: 14.2 knot". Fitur ini memungkinkan pengguna untuk melakukan investigasi cepat terhadap kapal-kapal individual langsung dari antarmuka peta, mengubah data mentah menjadi informasi kontekstual yang mudah dipahami.
- Manfaat Praktis: Dasbor visual ini memiliki implikasi praktis yang signifikan. Bagi otoritas pelabuhan atau penjaga pantai, peta ini dapat berfungsi sebagai alat pemantauan situasi maritim (maritime situational awareness). Pola kecepatan yang tidak biasa (misalnya, kapal yang tiba-tiba melambat atau berhenti di jalur pelayaran utama) atau deviasi kapal dari rute yang lazim dapat dideteksi secara visual, yang berpotensi menjadi indikator awal untuk insiden keselamatan atau aktivitas ilegal.

Secara keseluruhan, hasil analisis dan visualisasi ini mengkonfirmasi bahwa arsitektur lakehouse tidak hanya mampu menangani volume data AIS, tetapi juga memfasilitasi ekstraksi insight yang actionable, baik dalam bentuk agregat statistik maupun representasi spasial yang intuitif.

4. Kesimpulan

Penelitian ini berhasil menunjukkan bahwa arsitektur lakehouse dapat diimplementasikan secara efektif untuk pengolahan data pelayaran AIS. Dengan menggunakan PySpark, data mentah dari file CSV berhasil diolah, ditransformasi, dan disimpan dalam format Parquet yang efisien. Dari data yang tersimpan di lakehouse, analisis menggunakan Spark SQL mampu mengidentifikasi kapal dengan MMSI 525301327 sebagai yang tercepat dengan kecepatan rata-rata 16.95 knot. Lebih lanjut, data tersebut berhasil divisualisasikan menjadi sebuah peta interaktif menggunakan Folium yang memetakan 500 titik data kapal di perairan utara Bali, membuktikan bahwa pendekatan ini mampu mendukung analisis kuantitatif dan spasial dari satu sumber data terpadu.

Meskipun demikian, implementasi ini masih dalam skala purwarupa, menggunakan satu set data CSV yang terbatas. Oleh karena itu, penelitian lebih lanjut sangat disarankan untuk menguji dan memvalidasi arsitektur ini pada skala yang lebih besar. Pengujian dengan volume data hingga terabyte dan data yang bersifat streaming secara real-time akan memberikan gambaran yang lebih akurat mengenai performa dan skalabilitas sistem. Selain itu, penelitian mendatang dapat diperluas dengan menerapkan model machine learning untuk deteksi anomali perilaku kapal atau analisis prediktif untuk optimalisasi rute pelayaran, yang merupakan langkah logis selanjutnya setelah fondasi data yang solid terbangun.

5. Ucapan Terima Kasih

Penulis menyampaikan penghargaan dan terima kasih yang sebesar-besarnya kepada Rektor Institut Bisnis dan Teknologi Indonesia (INSTIKI) atas dukungan dan kepercayaan yang diberikan dalam pelaksanaan kegiatan penelitian ini. Ucapan terima kasih juga disampaikan kepada Direktorat Penelitian dan Pengabdian kepada Masyarakat (DRPM INSTIKI) yang telah mendanai kegiatan ini melalui skema INSTIKI Instiki Research and Development(IRDP). Dukungan yang diberikan telah menjadi faktor penting dalam keberhasilan implementasi sistem monitoring pada mitra usaha dan dalam tercapainya tujuan kegiatan penelitian ini.

Daftar Pustaka

Ahmed, N., Barczak, A. L. C., Sušnjak, T., & Rashid, M. A. (2020). A Comprehensive Performance Analysis of Apache Hadoop and Apache Spark for Large Scale Data Sets Using HiBench. *Journal of Big Data*, *7*(1). https://doi.org/10.1186/s40537-020-00388-5

Aissi, M. E. M. El, Benjelloun, S., Loukili, Y., Lakhrissi, Y., Boushaki, A. E., Chougrad, H., & Ali, S. E. B. (2021). *Data Lake Versus Data Warehouse Architecture: A Comparative Study*. 201–210. https://doi.org/10.1007/978-981-33-6893-4_19

Chen, D. D., & Yao, Z. G. (2013). Analysis on Ship Equipment Consumption Data Based on Data Mining. *Advanced Materials Research*, *846–847*, 1141–1144. https://doi.org/10.4028/www.scientific.net/amr.846-847.1141

Hu, Q., Chen, Y., & He, Y. (2021). Construction and Performance Analysis of Data Warehouse Based on Ship Propulsion System. *E3s Web of Conferences*, *248*, 01060. https://doi.org/10.1051/e3sconf/202124801060

Loukiala, A., Joutsenlahti, J.-P., Raatikainen, M., Mikkonen, T., & Lehtonen, T. (2021). *Migrating From a Centralized Data Warehouse to a Decentralized Data Platform Architecture*. 36–48. https://doi.org/10.1007/978-3-030-91452-3_3

Pandey, V., Kipf, A., Neumann, T., & Kemper, A. (2018). How Good Are Modern Spatial Analytics Systems? *Proceedings of the VLDB Endowment*, *11*(11), 1661–1673. https://doi.org/10.14778/3236187.3236213

Park, S., Yang, C. S., & Kim, J. W. (2023). Design of Vessel Data Lakehouse with Big Data and AI Analysis Technology for Vessel Monitoring System. *Electronics (Switzerland)*, *12*(8). https://doi.org/10.3390/electronics12081943

Plazotta, M., & Klettke, M. (2024). *Data Architectures in Cloud Environments*. 243–247. https://doi.org/10.1007/s13222-024-00490-5

Sudiarsa, W., Made, I., Widyantara, O., Ary, N. M., & Wirastuti, E. D. (2024). *OPTIMIZING VESSEL ROUTING AND SUSTAINABILITY WITH LAKEHOUSE ARCHITECTURES: A SYSTEMATIC REVIEW*. *3*. http://eprosiding.idbbali.ac.id/index.php/imade

wong, B. (2023). *Navigating the Data Architecture Landscape: A Comparative Analysis of Data Warehouse, Data Lake, Data Lakehouse, and Data Mesh*. https://doi.org/10.20944/preprints202309.2113.v1

Νομικός, N., Gkonis, P. K., Bithas, P. S., & Trakadas, P. (2023). A Survey on UAV-Aided Maritime Communications: Deployment Considerations, Applications, and Future Challenges. *Ieee Open Journal of the Communications Society*, *4*, 56–78. https://doi.org/10.1109/ojcoms.2022.3225590