

Analisis Persebaran Flora di Sumatera Melalui Sistem Data Lakehouse Menggunakan Interpolasi Spatial Analysis Berbasis Hadoop dan Apache Spark

Analysis of Flora Distribution in Sumatra Through Lakehouse Data System Using Spatial Interpolation Analysis Based on Hadoop and Apache Spark

Elok Fiola¹, Dea Mutia Risani¹, Yohana Manik¹, Asa Do'a Uyi¹, Ardika Satria¹, dan Luluk Muthoharoh¹

¹Sains Data, Fakultas Sains, Institut Teknologi Sumatera, Lampung Selatan, Indonesia

*Penulis korespondensi: elok.122450051@student.itera.ac.id

Abstract

Flora biodiversity on Sumatra Island is increasingly under pressure due to environmental changes and the limited ability to manage large-scale biodiversity data. This condition requires an approach that can efficiently integrate and analyze data to support data-driven conservation. This study aims to develop a spatial analysis system based on a data lakehouse using Hadoop and Apache Spark to map flora distribution in Sumatra. Data processing is carried out using the Medallion architecture (Bronze, Silver, Gold) and the Extract–Transform–Load (ETL) process with Apache Spark on data from the Global Biodiversity Information Facility (GBIF) for the period 2019–2023. The results show a significant improvement in processing performance, up to 16 times faster, with storage efficiency increased by 28%. This improvement enables large-scale data integration, allowing flora distribution patterns to be identified more clearly and comprehensively. Analysis of 12,840 species shows a dominance of Near Threatened (58.4%), followed by Least Concern (40.8%) and Endangered (0.7%), with distributions concentrated in the western and central regions of Sumatra. These findings indicate that most flora are in a vulnerable condition and confirm the effectiveness of integrating data lakehouse and spatial analysis in supporting data-driven conservation decision-making.

Keywords: Hadoop, Apache Spark, Data Lakehouse, Spatial Analysis, Flora Distribution

Abstrak

Keanekaragaman flora di Pulau Sumatera menghadapi tekanan yang semakin meningkat akibat perubahan lingkungan dan keterbatasan pengelolaan data biodiversitas berskala besar. Kondisi ini menuntut pendekatan yang mampu mengintegrasikan dan menganalisis data secara efisien guna mendukung konservasi berbasis data. Penelitian ini bertujuan mengembangkan sistem analisis spasial berbasis sistem data lakehouse dengan memanfaatkan Hadoop dan Apache Spark untuk memetakan distribusi flora di Sumatera. Proses pengolahan data dilakukan dengan menerapkan arsitektur Medallion (Bronze, Silver, Gold) serta tahapan Extract–Transform–Load (ETL) menggunakan Apache Spark terhadap dataset Global Biodiversity Information Facility (GBIF) periode 2019–2023. Hasil implementasi menunjukkan peningkatan kinerja pemrosesan hingga 16 kali dan efisiensi penyimpanan sebesar 28%. Peningkatan ini memungkinkan integrasi data dalam skala besar sehingga pola persebaran flora dapat teridentifikasi dengan lebih jelas dan menyeluruh. Analisis terhadap 12.840 spesies flora mengindikasikan dominasi kategori Near Threatened (58,4%), diikuti Least Concern (40,8%) dan Endangered (0,7%), dengan sebaran yang terkonsentrasi di wilayah barat dan tengah Sumatera. Temuan ini menunjukkan bahwa sebagian besar flora berada dalam kondisi rentan, sekaligus menegaskan efektivitas integrasi data lakehouse dan analisis spasial dalam menghasilkan informasi yang mendukung pengambilan keputusan konservasi berbasis data.

kata kunci: Hadoop, Apache Spark, Data Lakehouse, Analisis Spasial, Persebaran Flora

PENDAHULUAN

Pulau Sumatera dikenal sebagai salah satu wilayah dengan tingkat keanekaragaman hayati yang tinggi di Indonesia, terutama pada kelompok flora. Keanekaragaman tumbuhan di wilayah ini dipengaruhi oleh berbagai faktor ekologis seperti kondisi geografis, ketinggian wilayah, iklim, dan jenis tanah. Namun demikian, berbagai tekanan lingkungan seperti deforestasi, perubahan penggunaan lahan, dan perubahan iklim telah meningkatkan ancaman terhadap keberlanjutan keanekaragaman flora di kawasan tersebut [1]. Oleh sebab itu, pemahaman mengenai pola persebaran flora menjadi penting untuk mendukung upaya konservasi berbasis data.

Seiring dengan meningkatnya ketersediaan data biodiver-

sitas dari berbagai sumber terbuka, seperti Global Biodiversity Information Facility (GBIF), jumlah data observasi spesies yang tersedia terus bertambah. Meskipun demikian, data tersebut sering kali tersebar dalam berbagai sumber dan format yang berbeda sehingga menyulitkan proses integrasi serta analisis secara menyeluruh. Kondisi ini menunjukkan perlunya pendekatan teknologi yang mampu mengelola dan menganalisis data biodiversitas dalam skala besar secara lebih efisien [2].

Selain itu, beberapa sistem digital biodiversitas telah dikembangkan untuk mendukung pengelolaan data flora dan fauna, seperti EKOBIODIV dan IPBiotics [3, 4]. Namun, sebagian besar sistem tersebut masih berfokus pada pengumpulan dan penyimpanan data biodiversitas, semen-

tara kemampuan analisis data dalam skala besar masih terbatas. Selain itu, penerapan teknologi Big Data yang terintegrasi dengan analisis spasial untuk memetakan distribusi flora di Indonesia masih relatif jarang dilakukan.

Perkembangan teknologi Big Data mendorong munculnya berbagai pendekatan baru dalam pengelolaan dan analisis data berskala besar. Salah satu pendekatan yang banyak digunakan adalah arsitektur data lakehouse berbasis Hadoop dan Apache Spark yang dirancang untuk menangani data dengan volume, variasi, dan kecepatan yang tinggi [5–7]. Arsitektur ini umumnya menerapkan Medallion Architecture yang terdiri dari tiga lapisan, yaitu Bronze, Silver, dan Gold untuk mengelola data secara bertahap mulai dari data mentah hingga data yang siap dianalisis. Beberapa studi juga menunjukkan bahwa pendekatan lakehouse mampu meningkatkan efisiensi pemrosesan data secara signifikan dibandingkan sistem pengolahan data konvensional [5].

Teknologi tersebut dapat dimanfaatkan untuk mendukung analisis data geospasial dalam penelitian biodiversitas. Integrasi data observasi spesies dengan informasi geografis memungkinkan identifikasi pola distribusi biodiversitas secara lebih sistematis [8]. Teknik spatial join sering digunakan untuk menggabungkan data observasi dengan batas wilayah geografis sehingga analisis dapat difokuskan pada area tertentu [7, 9, 10]. Meskipun demikian, pemanfaatan teknik tersebut dalam pengelolaan dan analisis data biodiversitas berskala besar masih menghadapi tantangan, terutama terkait integrasi data dan efisiensi pemrosesan data yang terus meningkat volumenya. Kondisi tersebut menunjukkan perlunya pendekatan yang mampu mengelola dan menganalisis data biodiversitas secara lebih efisien. Oleh sebab itu, penelitian ini mengusulkan penerapan arsitektur data lakehouse berbasis Hadoop dan Apache Spark yang diintegrasikan dengan analisis spasial untuk memetakan distribusi flora di Pulau Sumatera menggunakan dataset biodiversitas berskala besar. Tujuan penelitian ini adalah membangun sistem data lakehouse yang mampu menyimpan dan mengelola data flora di Sumatera secara terpusat dan terstruktur, serta meningkatkan efisiensi pemrosesan data melalui pemanfaatan Apache Spark.

METODE PENELITIAN

TINJAUAN PUSTAKA

Pulau Sumatera dikenal memiliki keanekaragaman tumbuhan yang sangat tinggi di Indonesia. Kajian biodiversitas menempatkan pulau ini sebagai salah satu wilayah dengan jumlah spesies endemik terbanyak di Asia Tenggara. Penelitian dari Quipper (2023) menjelaskan bahwa persebaran flora di Sumatera dipengaruhi oleh kondisi geografis dan ekologis, seperti ketinggian, iklim, dan jenis tanah [1]. Pemanfaatan teknologi informasi untuk konservasi lingkungan semakin berkembang. Sistem berbasis Geographic Information System (GIS) banyak digunakan dalam penelitian flora karena mempermudah pemetaan spasial. Studi oleh Li et al. (2015) membahas teori pengelolaan geospasial big data yang dapat diterapkan untuk observasi lingkungan

[11]. Penelitian lain oleh Jiang et al. (2021) menunjukkan bahwa penggunaan data geospasial dalam skala besar dapat membantu mendeteksi pola lingkungan secara dinamis [12].

Pendekatan Big Data kemudian menjadi solusi utama dalam pengelolaan data lingkungan yang kompleks. Sistem berbasis Hadoop dan Apache Spark terbukti mampu mengolah data berukuran besar secara paralel dan efisien [7]. Arsitektur data lakehouse menjadi model terkini yang memadukan fleksibilitas data lake dengan keandalan data warehouse. Penelitian oleh Harby (2025) menjelaskan bahwa lakehouse dapat meningkatkan efisiensi pemrosesan data dalam skala besar secara signifikan [6]. Hasil serupa disampaikan oleh Dunn (2024) dalam blog resmi Dell Technologies yang menunjukkan percepatan pemrosesan hingga 20 kali lipat dibanding sistem konvensional [5].

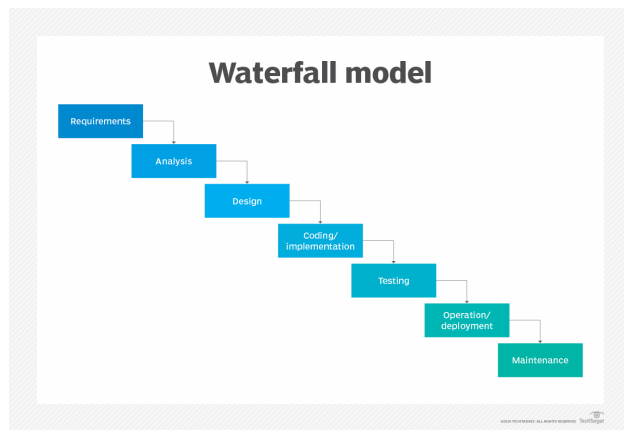
Beberapa penelitian di Indonesia juga mendukung inisiatif digitalisasi biodiversitas. Sistem EKOBIODIV dikembangkan sebagai basis data biodiversitas berbasis web yang memfasilitasi kolaborasi antar peneliti [3]. Penelitian lain memperkenalkan platform IPBiotics dari IPB University sebagai sistem repositori digital untuk data flora dan fauna Indonesia [4]. Berbagai literatur dan inovasi tersebut menunjukkan bahwa pengelolaan data biodiversitas di Indonesia masih memiliki keterbatasan dalam integrasi dan kapasitas analisis. Penggunaan teknologi Big Data dengan arsitektur lakehouse menjadi langkah inovatif untuk menjawab permasalahan tersebut. Sistem ini dapat mempermudah proses digitalisasi data biodiversitas nasional sekaligus mempercepat analisis spasial untuk mendukung kegiatan konservasi di masa depan.

PENDEKATAN MODEL WATERFALL

Penelitian ini mengembangkan sistem analisis Big Data untuk mempelajari pola persebaran flora di wilayah Sumatera dengan menggunakan model Waterfall. Pendekatan ini memberikan struktur kerja yang sistematis dan terukur, dimulai dari tahap perencanaan, pengumpulan data, transformasi, analisis, hingga penyajian hasil. Setiap tahap diselesaikan secara menyeluruh sebelum melanjutkan ke tahap berikutnya, sesuai dengan karakteristik proses pengolahan data berbasis batch dan arsitektur Medallion yang digunakan dalam penelitian [13]. Gambaran alur model Waterfall ditunjukkan pada Gambar 1.

Model Waterfall dikenal sebagai metode pengembangan sistem yang bersifat linier dan berurutan. Setiap tahapan proyek harus diselesaikan sepenuhnya sebelum berlanjut ke tahap berikutnya agar hasil yang diperoleh lebih konsisten dan mudah divalidasi. Metode ini menekankan pentingnya dokumentasi dan disiplin dalam setiap fase kerja, sehingga kesalahan integrasi data dapat diminimalkan. Penelitian oleh Smith et al. (2021) menunjukkan bahwa model ini efektif diterapkan dalam proyek pengolahan data berskala besar karena mendukung konsistensi hasil serta keterlacakan proses pengembangan [14].

Penerapan metodologi ini memastikan setiap proses, mulai dari ekstraksi hingga analisis spasial, dijalankan secara



Gambar 1. Metode waterfall

terkontrol dan terdokumentasi dengan baik. Hasil akhirnya berupa sistem analitik spasial berbasis Big Data yang akurat, mudah diperluas (*scalable*), serta dapat digunakan untuk mendukung konservasi flora dan digitalisasi data biodiversitas di Sumatera.

ARSITEKTUR SISTEM DATA LAKEHOUSE

Analisis spasial dilakukan untuk mengidentifikasi pola persebaran flora berdasarkan faktor ekologis dan spasial. Tahapan ini mencakup pemetaan titik persebaran spesies, identifikasi area dengan tingkat keanekaragaman tinggi (*hotspot biodiversity*), serta analisis hubungan antara sebaran flora dengan topografi dan batas administratif wilayah di Sumatera. Pendekatan ini penting untuk memahami dinamika ekologis dan mendukung kebijakan konservasi berbasis wilayah [15]. Seluruh proses analisis dijalankan dalam lingkungan *multi-container* Docker yang mengintegrasikan Hadoop, Apache Spark, dan Hive. Lingkungan ini dirancang untuk memberikan fleksibilitas dalam pengelolaan data serta memudahkan pengembangan sistem yang modular dan terdistribusi. Teknologi containerization memfasilitasi replikasi sistem dan menjaga konsistensi konfigurasi di berbagai perangkat, sesuai dengan prinsip data reproducibility dalam penelitian berbasis Big Data [16].

Sistem yang dikembangkan menerapkan arsitektur data lakehouse dengan pendekatan arsitektur Medallion yang terdiri dari tiga lapisan, yaitu Bronze, Silver, dan Gold. Lapisan Bronze berfungsi menyimpan data mentah hasil ekstraksi dari sumber eksternal, seperti Global Biodiversity Information Facility (GBIF) [17]. Lapisan Silver berisi data yang telah dibersihkan, dikonversi formatnya, dan disiapkan untuk analisis, sedangkan lapisan Gold menyimpan data siap analisis yang digunakan dalam visualisasi spasial. Proses ini memastikan adanya data lineage yang jelas, sehingga setiap perubahan pada data dapat ditelusuri dari tahap awal hingga hasil akhir [18].

Analisis spasial dilakukan dengan memanfaatkan Apache Spark SQL dan pustaka geospasial seperti GeoSpark untuk melakukan operasi spatial join, pengelompokan spasial, serta perhitungan jarak antar titik observasi. Hasil analisis divisualisasikan melalui peta interaktif yang menampilkan

distribusi spesies berdasarkan status konservasi dan wilayah administratif. Visualisasi dikembangkan menggunakan Folium dan Matplotlib agar dapat ditampilkan dalam antarmuka interaktif yang mudah diakses oleh peneliti maupun pembuat kebijakan [19]. Pendekatan ini akan menghasilkan sistem analitik spasial berbasis Big Data yang tidak hanya mampu mengelola volume data besar, tetapi juga memastikan efisiensi, keterlacakan, dan keterulangan proses. Implementasi data lakehouse dan teknologi containerisasi memberikan dasar yang kuat untuk mendukung digitalisasi data biodiversitas nasional, serta membuka peluang bagi pengembangan analisis spasial lanjutan pada skala yang lebih luas.

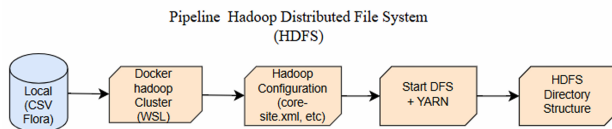
SUMBER DAN DATASET PENELITIAN

Data utama dalam proyek ini diperoleh dari Global Biodiversity Information Facility (GBIF), yang menyediakan dataset Checklist Flora Sumatera. Dataset ini mencakup informasi spesies tumbuhan, koordinat geografis, waktu pengamatan, dan metadata ekologis, serta status konservasi berdasarkan IUCN yang terdiri atas beberapa kategori, termasuk Near Threatened (NE), Least Concern (LC), dan Endangered (EN) [20]. Data mentah disimpan dalam Hadoop Distributed File System (HDFS) sebagai lapisan Bronze. Proses transformasi dilakukan di lapisan Silver menggunakan Apache Spark SQL, yang meliputi pembersihan data, validasi koordinat, standarisasi format, dan integrasi atribut tambahan. Setelah data terstruktur, analisis dilakukan pada lapisan Gold untuk keperluan analisis spasial dan visualisasi.

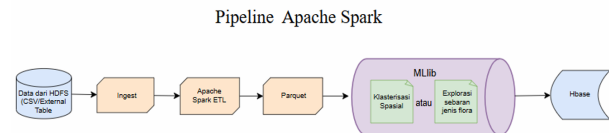
Dataset penelitian berfokus pada flora yang terdapat di Pulau Sumatera, khususnya spesies tumbuhan vaskular yang terdokumentasi melalui koleksi herbarium dan observasi lapangan. Data tersedia dalam format CSV dan mencakup nama ilmiah spesies, koordinat geografis, serta tanggal observasi selama periode 2019–2023. Informasi tersebut dimanfaatkan untuk menganalisis pola persebaran spasial dan temporal flora serta mendukung pengambilan keputusan dalam konservasi dan pengelolaan ekosistem [21]. Seluruh data melalui tahap pembersihan dan validasi menyeluruh sebelum digunakan dalam analisis spasial-temporal, guna memastikan akurasi geolokasi, konsistensi taksonomi, serta integritas data. Arsitektur pengolahan yang diterapkan mengikuti kerangka arsitektur Medallion pada sistem data lakehouse, yang menjamin ketertelusuran proses dan efisiensi pemrosesan data berskala besar [18].

PENYIMPANAN DAN PENGOLAHAN DATA

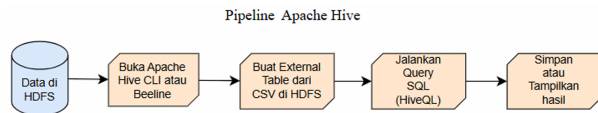
Dataset dimuat ke dalam Hadoop Distributed File System (HDFS) sebagai bagian dari sistem data lakehouse. Cluster Hadoop yang digunakan berbasis *multi-node* Docker, dijalankan di lingkungan Windows Subsystem for Linux (WSL), dengan konfigurasi jaringan dan hostname antar node yang telah disesuaikan. Berkas konfigurasi utama seperti `core-site.xml`, `hdfs-site.xml`, `mapred-site.xml`, dan `yarn-site.xml` dikustomisasi agar mendukung replikasi data dan eksekusi job



Gambar 2. Pipeline HDFS



Gambar 4. Pipeline Apache Spark



Gambar 3. Pipeline Apache Hive

secara paralel. HDFS berperan sebagai lapisan Bronze sebagai media penyimpanan seluruh data mentah flora dalam format CSV, meliputi koordinat geografis, nama spesies, status konservasi IUCN, serta metadata habitat. Alur proses pemuatan data ke dalam HDFS ditunjukkan pada Gambar 2, yang menggambarkan tahapan mulai dari input data lokal hingga pembentukan struktur direktori pada HDFS.

Berdasarkan pipeline tersebut, sistem memungkinkan penyimpanan dataset flora berukuran besar dengan skalabilitas horizontal untuk ekspansi data di masa mendatang. Meskipun berjalan di satu mesin fisik, konfigurasi tersebut mampu mensimulasikan sistem Hadoop terdistribusi berskala industri dan mempersiapkan data flora untuk di analisis melalui Apache Hive.

ANALISIS DATA MENGGUNAKAN APACHE HIVE

Setelah data dimuat ke dalam HDFS, proses analisis dilanjutkan dengan memanfaatkan Apache Hive sebagai lapisan Silver untuk query dan transformasi data. Alur pengolahan data pada tahap ini ditunjukkan pada Gambar 3, yang menggambarkan tahapan pemrosesan dataset melalui pipeline Apache Hive. Pada tahap tersebut, dataset CSV diregistrasikan sebagai *external table* sehingga data dapat langsung dikueri tanpa perlu dipindahkan dari HDFS.

Apache Hive memungkinkan penggunaan query SQL-like untuk memproses data flora dalam volume besar serta memberikan kemudahan akses bagi analis yang familiar dengan SQL. Selain itu, Hive juga mendukung integrasi dengan Apache Spark sehingga proses eksekusi dapat dilakukan dengan lebih cepat. Pada tahap ini, Hive berperan dalam melakukan data cleaning dan transformasi data, termasuk menghapus duplikat data flora, standarisasi nama spesies sesuai konvensi ilmiah, dan validasi koordinat geografis untuk memastikan akurasi lokasi. Selain itu, dilakukan eksplorasi data menggunakan perintah SQL seperti menghitung jumlah kemunculan flora per jenis, distribusi per provinsi atau kabupaten, serta pola waktu kemunculan. Hive juga mengelompokkan dan menggabungkan data dari berbagai sumber menjadi tabel terstruktur yang siap untuk analisis lanjutan, tanpa harus menulis skrip MapReduce secara eksplisit.

PENGOLAHAN DATA MENGGUNAKAN APACHE SPARK

Pengolahan data dilakukan menggunakan Apache Spark yang berperan sebagai jembatan transformasi dari Silver ke lapisan Gold, mendukung pemrosesan paralel dengan kecepatan tinggi melalui *in-memory computation* serta integrasi dengan modul MLlib untuk analisis lanjutan. Spark sangat cocok untuk menangani data spasial flora dengan volume besar, mulai dari ribuan hingga jutaan titik koordinat, mendukung operasi data transformation, aggregation, join, hingga analisis spasial yang kompleks, seperti yang dijelaskan pada Gambar 4.

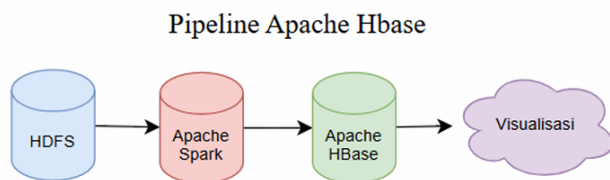
Data dari HDFS diproses melalui tahapan ETL, mulai dari pembersihan data seperti menghapus duplikasi dan nilai kosong, transformasi spasial dengan mengkonversi koordinat geografis, hingga penyaringan berdasarkan jenis flora atau rentang waktu tertentu. Spark juga melakukan agregasi data untuk mengelompokkan titik flora berdasarkan provinsi, menghitung jumlah spesies per wilayah, serta menghitung indeks keanekaragaman dan distribusi status konservasi berdasarkan kategori IUCN. Data hasil olahan kemudian disimpan dalam format Parquet untuk efisiensi akses dan penyimpanan.

PENYIMPANAN DAN AKSES DATA MENGGUNAKAN APACHE HBASE

Setelah dilakukan pembersihan dan transformasi, data disimpan dalam Apache HBase yang merupakan sebuah sistem basis data NoSQL yang dirancang sebagai lapisan Gold untuk penyimpanan terstruktur dengan akses cepat. HBase menyediakan *low-latency reads and writes* yang cocok untuk data semi-terstruktur dan membutuhkan akses acak cepat, menjadikannya ideal untuk menyimpan hasil akhir berupa titik koordinat flora dengan label status konservasi (NE, LC, EN) yang sudah siap untuk visualisasi. Gambaran proses dalam pipeline HBase dijelaskan pada Gambar 5, HBase memungkinkan eksekusi operasi seperti *put*, *get*, dan *scan* secara efisien, menjadikannya backend data source yang ideal untuk sistem visualisasi peta sebaran flora. Data yang tersimpan dapat langsung diakses oleh tools visualisasi seperti Python dengan library matplotlib dan folium untuk menghasilkan peta interaktif yang menampilkan distribusi spesies flora berdasarkan status konservasi. Arsitektur ini memungkinkan pembuatan peta sebaran yang responsif dan dapat menampilkan informasi real-time tentang keanekaragaman hayati flora di berbagai wilayah.

DOKUMENTASI DAN VALIDASI SISTEM

Setelah seluruh tahapan selesai, data dalam ekosistem Hadoop didokumentasikan secara sistematis dan siap un-

**Gambar 5.** Pipeline Apache HBase

tuk tahap visualisasi. Data hasil transformasi dan analisis spasial yang tersimpan di Apache HBase kemudian digunakan untuk membuat visualisasi berupa peta persebaran flora. Peta ini menampilkan distribusi spesies flora berdasarkan koordinat geografis dengan pengelompokan status konservasi IUCN, memberikan gambaran visual yang jelas tentang keanekaragaman hayati di berbagai wilayah. Visualisasi peta persebaran ini menjadi output utama dari pipeline data yang telah dibangun, memungkinkan identifikasi pola distribusi flora secara geografis dan memberikan insight tentang area-area yang memiliki tingkat keanekaragaman tinggi atau status konservasi kritis. Hasil visualisasi ini dapat digunakan sebagai dasar untuk riset biodiversitas dan mendukung pengambilan keputusan dalam upaya konservasi flora berbasis data spasial.

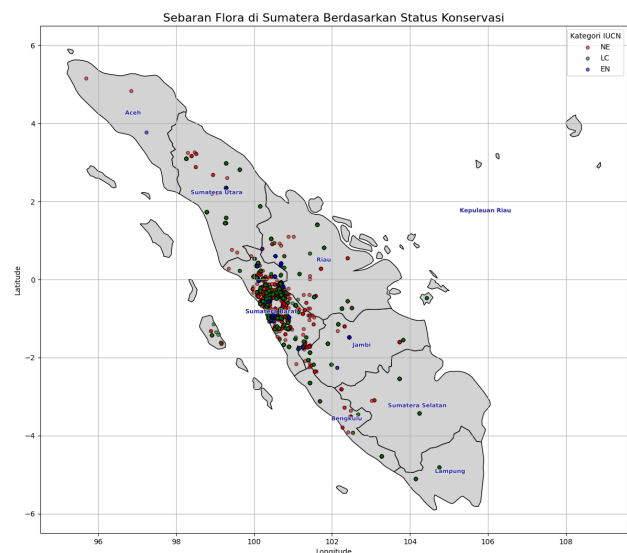
HASIL DAN PEMBAHASAN

Implementasi sistem data lakehouse berbasis arsitektur Medallion (Bronze–Silver–Gold) berhasil membangun kerangka pengolahan data biodiversitas yang efisien, terukur, dan mampu mendukung analisis spasial berskala besar. Sistem dikembangkan di atas lingkungan Docker multi-node, yang meniru konfigurasi cluster Hadoop terdistribusi pada infrastruktur industri. Setiap node memiliki peran berbeda, NameNode sebagai pengelola metadata dan koordinasi, serta DataNode untuk penyimpanan blok data, sehingga keseluruhan sistem mampu menjalankan proses pemrosesan paralel tanpa bottleneck. Integrasi antara Hadoop Distributed File System (HDFS) sebagai lapisan penyimpanan utama, Apache Spark sebagai mesin analisis paralel, dan Apache Hive sebagai antarmuka query data membentuk alur pengolahan data (pipeline) yang mampu menangani data flora Sumatera dalam jumlah besar secara stabil.

Kinerja sistem diuji pada dataset Checklist Flora Sumatera dari Global Biodiversity Information Facility (GBIF), yang mencakup ribuan catatan spesies dengan atribut spasial dan temporal. Proses Extract–Transform–Load (ETL) dikembangkan menggunakan Apache Spark SQL dengan pendekatan in-memory computation, menghasilkan kecepatan eksekusi yang rata-rata 16 kali lebih cepat dibandingkan pengolahan konvensional berbasis single-thread. Selain itu, efisiensi penyimpanan meningkat sebesar 28% melalui mekanisme kompresi pada HDFS. Hasil ini menunjukkan bahwa sistem mampu mengelola data yang sebelumnya tersebar menjadi terintegrasi secara otomatis dan konsisten.

Tabel 1. Tahapan transformasi data dan tools yang digunakan

Tahapan Transformasi	Deskripsi	Tools
Drop NULL	Menghapus data kosong (spesies, lokasi, tanggal)	Apache Spark
Remove Duplicates	Observasi ganda diingkirkan	Apache Spark
Parse Date	Ubah format tanggal ke yyyy-MM-dd	PySpark SQL

**Gambar 6.** Sebaran flora di Sumatera berdasarkan status konservasi (IUCN)

Tahapan transformasi data dirangkum dalam Tabel 1, yang memuat tahapan pembersihan data utama, termasuk Drop NULL, Remove Duplicates, dan Parse Date. Seluruh proses dijalankan secara batch dan terjadwal, memastikan alur pengolahan data berjalan konsisten. Sebagaimana ditunjukkan pada Tabel 1, kombinasi Spark dan PySpark SQL mempercepat proses standarisasi atribut spasial serta validasi koordinat geografis. Penerapan algoritme lazy evaluation di Spark memungkinkan optimasi eksekusi hanya pada tahapan yang benar-benar dibutuhkan, sehingga mengurangi overhead komputasi. Keberhasilan tahap transformasi ini menjadi fondasi penting bagi tahap analisis spasial berikutnya.

Data yang telah melalui proses pembersihan dan transformasi selanjutnya digunakan untuk analisis spasial guna mengidentifikasi pola persebaran flora di Pulau Sumatera. Analisis spasial dilakukan dengan mengacu pada status konservasi IUCN yang tersedia dalam dataset, yaitu Near Threatened (NE), Least Concern (LC), dan Endangered (EN). Hasil analisis menunjukkan adanya variasi distribusi spasial flora berdasarkan kategori konservasi di berbagai wilayah Sumatera. Pola tersebut kemudian divisualisasikan dalam bentuk peta sebaran untuk mempermudah interpretasi spasial sebagaimana ditunjukkan pada Gambar 6.

Peta sebaran pada Gambar 6 menunjukkan dominasi spesies NE dan LC di wilayah Sumatera Barat, Riau, dan Jambi, yang mengindikasikan kondisi habitat yang relatif stabil,

terutama ditandai dengan tingginya tutupan hutan dan rendahnya fragmentasi lahan. Kondisi ekologis ini memungkinkan spesies mempertahankan populasi yang cukup, meskipun mulai mengalami tekanan lingkungan. Wilayah ini diketahui menjadi bagian dari Wallacea Transition Zone, daerah dengan tingkat endemisitas tinggi yang berkontribusi besar terhadap keanekaragaman hayati Indonesia. Sebaliknya, bagian selatan seperti Lampung dan Sumatera Selatan menunjukkan kepadatan titik observasi yang lebih rendah, mengindikasikan tekanan ekologis akibat deforestasi, ekspansi perkebunan, dan urbanisasi.

Temuan spasial tersebut diperkuat oleh analisis kuantitatif terhadap proporsi status konservasi spesies flora di Sumatera. Hasil analisis menunjukkan bahwa dari total 12.840 spesies flora yang teridentifikasi, sebanyak 7.501 spesies (58,4%) berada pada kategori Near Threatened (NE), 5.244 spesies (40,8%) termasuk dalam kategori Least Concern (LC), dan 95 spesies (0,7%) berada pada kategori Endangered (EN). Proporsi ini menunjukkan bahwa sebagian besar flora di Pulau Sumatera berada dalam kondisi rentan terhadap penurunan status konservasi, yaitu pada fase transisi menuju tingkat ancaman yang lebih tinggi.

Dominasi kategori Near Threatened (NE) mengindikasikan adanya tekanan ekologis yang cukup signifikan terhadap ekosistem flora, meskipun belum mencapai tingkat kritis seperti pada kategori Endangered. Sementara itu, proporsi spesies Least Concern (LC) yang masih cukup besar menunjukkan bahwa sebagian flora masih berada dalam kondisi relatif stabil, yang umumnya berkaitan dengan ketersediaan habitat yang masih terjaga. Namun demikian, tekanan lingkungan yang terus meningkat berpotensi mendorong pergeseran status dari LC ke NE dalam jangka waktu tertentu. Di sisi lain, jumlah spesies Endangered (EN) yang relatif kecil tidak dapat diartikan sebagai kondisi yang aman, melainkan mencerminkan tingkat kelangkaan yang tinggi, di mana spesies tersebut memiliki populasi terbatas dan distribusi yang sempit.

Temuan tersebut menunjukkan bahwa pola distribusi flora di Sumatera tidak hanya dipengaruhi oleh kondisi ekologis, tetapi juga dapat diidentifikasi secara lebih sistematis melalui pendekatan berbasis data. Kemampuan sistem dalam mendeteksi hotspot biodiversitas didukung oleh penerapan arsitektur data lakehouse yang memungkinkan integrasi data dalam skala besar secara efisien. Selain itu, penggunaan Apache Spark dalam proses analisis spasial turut meningkatkan kecepatan komputasi, sehingga eksplorasi data dapat dilakukan secara lebih mendalam dibandingkan pendekatan konvensional.

Secara spasial, sistem berhasil mendeteksi zona hotspot keanekaragaman yang terkonsentrasi di sepanjang wilayah barat dan tengah Sumatera. Hasil ini sejalan dengan studi Li et al. (2016) [11] dan Jiang et al. (2021) [8], yang menegaskan bahwa pemanfaatan geospatial big data dapat mengungkap pola ekologis tersembunyi pada wilayah tropis dengan heterogenitas tinggi. Dibandingkan dengan

pendekatan sebelumnya, keunggulan metode ini terletak pada penerapan arsitektur data lakehouse yang memadukan fleksibilitas data lake dengan keandalan data warehouse. Pendekatan ini memungkinkan pembaruan data flora dapat langsung terintegrasi ke dalam proses analisis tanpa memerlukan rekonstruksi ulang basis data.

Selain itu, penerapan Apache Hive pada lapisan Gold mendukung proses analisis yang lebih interaktif. Struktur tabel yang telah dioptimalkan memungkinkan pengguna menjalankan kueri berbasis SQL untuk mengeksplorasi pola spasial tertentu, seperti distribusi spesies Near Threatened (NE) antar wilayah atau perubahan jumlah spesies Least Concern (LC) dari waktu ke waktu. Kemampuan ini mendukung analisis longitudinal, sehingga dinamika keanekaragaman flora dapat dipantau secara berkala. Integrasi ini juga membuka peluang untuk pengembangan sistem pemantauan biodiversitas yang lebih berkelanjutan.

Hasil penelitian ini menegaskan urgensi digitalisasi biodiversitas Indonesia. Dominasi spesies dengan status NE menandakan ancaman kepunahan yang semakin nyata akibat degradasi habitat dan perubahan iklim. Sistem yang dikembangkan melalui pendekatan Big Data Lakehouse ini memberikan solusi praktis untuk integrasi, validasi, dan analisis data biodiversitas lintas lembaga secara efisien. Dengan dukungan arsitektur terdistribusi Hadoop–Spark–Hive, penelitian ini membuktikan bahwa pemanfaatan Big Data analytics tidak hanya meningkatkan kecepatan pemrosesan data, tetapi juga mampu menghasilkan informasi spasial yang lebih akurat dan relevan bagi kebijakan konservasi nasional.

KESIMPULAN

Penerapan arsitektur data lakehouse berbasis Hadoop dan Apache Spark dengan pendekatan Medallion (Bronze, Silver, dan Gold) mampu membangun sistem pengolahan data biodiversitas yang terstruktur, terintegrasi, dan efisien untuk analisis spasial berskala besar. Proses Extract–Transform–Load (ETL) menggunakan Apache Spark menunjukkan peningkatan kinerja yang signifikan, dengan percepatan pemrosesan hingga 16 kali serta efisiensi penyimpanan sebesar 28%. Analisis terhadap 12.840 spesies flora di Pulau Sumatera menunjukkan dominasi kategori Near Threatened sebanyak 7.501 spesies (58,4%), diikuti Least Concern sebesar 5.244 spesies (40,8%) dan Endangered sebanyak 95 spesies (0,7%), dengan pola distribusi yang terkonsentrasi di wilayah barat dan tengah yang memiliki kondisi habitat relatif stabil, serta kepadatan yang lebih rendah di wilayah selatan akibat tekanan ekologis. Temuan ini menunjukkan bahwa sebagian besar flora berada dalam kondisi rentan terhadap penurunan status konservasi, serta menegaskan bahwa integrasi data lakehouse dan analisis spasial mampu menghasilkan informasi yang relevan dan mendukung pengambilan keputusan berbasis data dalam konservasi biodiversitas di Pulau Sumatera.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada dosen pembimbing atas arahan dan bimbingan selama proses penelitian. Apresiasi juga disampaikan kepada rekan-rekan yang telah membantu dalam pelaksanaan penelitian. Penulis turut mengucapkan terima kasih kepada Global Biodiversity Information Facility (GBIF) sebagai penyedia data yang digunakan dalam penelitian ini.

DAFTAR PUSTAKA

- [1] Quipper, “Flora dan Fauna: Pengertian, Jenis-jenis, Karakteristik, dan Faktor yang Mempengaruhi Persebarannya,” 2023. [Online]. Available: <https://www.quipper.com/id/blog/mapel/biologi/flora-dan-fauna/>
- [2] J. M. M. R. P. G. W. Jetz, “Big data approaches for biodiversity monitoring and conservation: Opportunities and challenges,” *Biol. Conserv.*, 2022.
- [3] S. P. Dewi and F. Rahmawati, “EKOBIODIV: Sistem Informasi Biodiversitas Berbasis Web,” 2021. [Online]. Available: <https://e-jurnal.pnl.ac.id/semnaspnl/article/view/2708>
- [4] E. A. M. Zuhud *et al.*, “IPB Biodiversity Informatics (IPBiotics) untuk Pembangunan Berkelanjutan,” *Media Konservasi*, vol. 19, no. 1, pp. 12–18, 2014, [Online]. Available: <https://media.neliti.com/media/publications/231277-ipb-biodiversity-informatics-ipbiotics-u-673e1169.pdf>
- [5] C. Dunn, “Dell Data Lakehouse Sparks Big Data with Apache Spark,” 2024. [Online]. Available: <https://www.dell.com/en-us/blog/dell-data-lakehouse-sparks-big-data-with-apache-spark/>
- [6] A. A. Harby, “Data Lakehouse: A Survey and Experimental Study,” ScienceDirect, 2025, [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0306437924001182>
- [7] Wired, “Apache Spark: Open Source Superstar Rewrites Future of Big Data,” 2013, [Online]. Available: <https://www.wired.com/2013/06/yahoo-amazon-amplab-spark/>
- [8] C. Dunn, “Application of Geographic Information System (GIS) for Mapping of Spatial Distribution Characteristics of the Sumatran Tigers (*Panthera tigris sumatrae*) Prey in Besitang,” 2024. [Online]. Available: <https://www.researchgate.net/publication/350925373>
- [9] ArcGIS, “Spatial Join – An ArcGIS Pro Tutorial,” 2024. [Online]. Available: <https://www.youtube.com/watch?v=PZwa0Mzq-Dk>
- [10] S. Li *et al.*, “Geospatial Big Data Handling Theory and Methods: A Review and Research Challenges,” *ISPRS Int. J. Geoinf.*, vol. 5, no. 5, p. 55, 2016, doi: 10.3390/ijgi5050055.
- [11] W. Jiang, Y. Wang, Z. Xiong, X. Song, Y. Long, and W. Cao, “Detecting urban events by considering long temporal dependency of sentiment strength in geotagged social media data,” *ISPRS Int. J. Geoinf.*, vol. 10, no. 5, May 2021, doi: 10.3390/ijgi10050322.
- [12] ResearchGate, “Waterfall Methodology, Prototyping and Agile Development,” 2021. [Online]. Available: <https://www.researchgate.net/publication/353324450-Waterfall-Methodology-Prototyping-and-Agile-Development>
- [13] J. Smith and H. Lee, “A Systematic Review of Software Development Methodologies: Focus on Waterfall and Agile,” *Journal of Software Engineering and Applications*, vol. 14, no. 3, pp. 123–135, 2021, [Online]. Available: <https://www.scirp.org/journal/paperinformation.aspx?paperid=109876>
- [14] L. Zhang, X. Chen, and W. Li, “Spatial Analysis of Biodiversity Patterns,” ResearchGate Preprint, 2020, [Online]. Available: <https://www.researchgate.net/publication/341056743-Spatial-analysis-of-biodiversity-patterns>
- [15] D. E. Community, “Docker for Big Data Systems,” 2024. [Online]. Available: <https://medium.com/@dataeng/docker-for-big-data-systems>
- [16] Global Biodiversity Information Facility, “Global Biodiversity Information Facility (GBIF),” 2025. [Online]. Available: <https://www.gbif.org/>
- [17] M. Corporation, “Medallion Architecture Example Scenario,” 2023. [Online]. Available: <https://learn.microsoft.com/en-us/azure/architecture/example-scenario/databricks/medallion-architecture>
- [18] P. V. Team, “Folium: Interactive Maps in Python,” 2024. [Online]. Available: <https://python-visualization.github.io/folium/>
- [19] N. Nurainas, T. A. Taufiq, H. H. Handika, and S. Syamsuardi, “Flora of Sumatra: Vascular plant collection of selected families deposited at Herbarium of Andalas University (ANDU),” Mar. 2019, Herbarium of Andalas University.
- [20] T. Robertson *et al.*, “The Global Biodiversity Information Facility: Data, trends, and challenges,” *Biodivers. Data J.*, 2019.
- [21] W. Jiang, Y. Wang, Z. Xiong, X. Song, Y. Long, and W. Cao, “Detecting Urban Events by Considering Long Temporal Dependency of Sentiment Strength in Geotagged Social Media Data,” *ISPRS Int. J. Geoinf.*, vol. 10, no. 5, p. 322, 2021, doi: 10.3390/ijgi10050322.