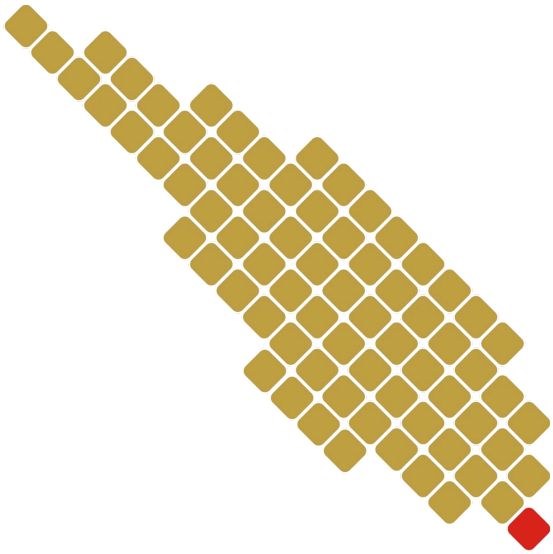
PROPOSAL TUGAS BESAR ANALISIS BIG DATA
Analisis Produksi Perikanan di Sumatera



Disusun oleh :

Dinda Joycehana	(122140048)
Elia Meylani Simanjuntak	(122450026)
Presilia	(122450081)
Randa Andriana Putra	(122450083)

PROGRAM STUDI SAINS DATA
FAKULTAS SAINS
INSTITUT TEKNOLOGI SUMATERA
2025

A. Latar Belakang

Dengan ribuan kilometer garis pantai dan akses ke wilayah perairan yang luas seperti Selat Malaka dan Samudera Hindia, Sumatera menjadi salah satu kontributor utama produksi perikanan nasional. Perikanan memainkan peran penting dalam mendukung pembangunan ekonomi dan ketahanan pangan bangsa. Provinsi seperti Sumatera Utara, Aceh, dan Sumatera Barat secara teratur mencatatkan tingkat produksi perikanan yang tinggi setiap tahunnya, menurut data dari Kementerian Kelautan dan Perikanan (Sri Endang Rahayu, 2022). Namun, meskipun ada potensi, pengelolaan data produksi perikanan masih menghadapi banyak masalah. Tantangan utama terletak pada jumlah data yang harus dianalisis dan kompleksitasnya. Data produksi mencakup jumlah hasil tangkapan, jenis ikan, teknik penangkapan, wilayah tangkapan, waktu panen, dan cara hasil didistribusikan. Karakteristik big data—volume besar, kecepatan tinggi, dan beragam jenis data—ditunjukkan oleh keragaman dan pertumbuhan data yang sangat cepat ini (Smith, T., & Zhang, Y., n.d.,).

Sebagai contoh, puluhan ribu entri dan hasil tangkapan yang dihasilkan setiap hari di Pelabuhan Perikanan Samudera Belawan di Sumatera Utara harus dicatat dan dipelajari untuk mendukung pengambilan kebijakan. Potensi ini dapat menjadi beban administratif dan menghambat respons kebijakan jika tidak ada sistem pengelolaan data yang memadai (Sugiarto, E., & Hartono, R., 2023,).

Teknologi Big Data mungkin merupakan solusi untuk masalah ini, khususnya dengan menggunakan ekosistem Hadoop, yang telah terbukti mampu menangani pengolahan data skala besar di berbagai industri. Hadoop memiliki sistem penyimpanan terdistribusi (HDFS) yang dapat menyimpan data dalam blok-blok di banyak node dan sistem pemrosesan paralel yang mempercepat proses komputasi melalui MapReduce. Keunggulan-keunggulan ini memungkinkan Hadoop digunakan untuk menyimpan dan menganalisis data produksi perikanan dalam jumlah besar secara efisien (Kamaluddin, A., & Rachman, H, n.d.,).

Selain itu, penggabungan komponen Hadoop seperti Apache Hive dan Apache HBase meningkatkan nilai pengolahan data. Meskipun data disimpan dalam HDFS, Hive memungkinkan peneliti melakukan kueri data dengan sintaks yang mirip dengan SQL, sehingga mempermudah proses analisis. Sementara itu, HBase memungkinkan peneliti menyimpan hasil analisis dalam bentuk basis data kolom yang dapat diakses secara acak dengan kecepatan rendah.. Dengan kombinasi ini, Hadoop adalah platform yang ideal untuk analisis data produksi perikanan yang membutuhkan skalabilitas, fleksibilitas, dan kecepatan (Smith, T., & Zhang, Y., n.d.,).

B. Tujuan

Tujuan proyek ini adalah untuk menggunakan ekosistem Hadoop untuk membangun pipeline analisis data produksi perikanan Sumatera yang mencakup pemrosesan, penyimpanan, kueri analitik, dan penyimpanan hasil. Secara khusus, tujuannya adalah:

a. Mengimplementasikan lingkungan Hadoop berbasis Docker secara lokal untuk simulasi cluster big data.
b. Melakukan ingestion dan pembersihan data produksi perikanan Sumatera ke dalam HDFS.
c. Membangun skrip MapReduce untuk menghitung agregasi produksi perikanan berdasarkan wilayah dan jenis komoditas.
d. Melakukan analisis deskriptif menggunakan HiveQL dan menyimpan hasil ringkasan ke dalam HBase untuk akses cepat.
e. Membangun model prediktif menggunakan algoritma regresi atau time-series untuk meramalkan produksi masa depan.
f. Mengintegrasikan seluruh proses dalam satu pipeline Hadoop yang teruji dan terdokumentasi.

C. Metodologi

Kebutuhan analisis data produksi perikanan di Sumatera yang sangat besar mendorong penggunaan ekosistem Hadoop, karena sistem ini dirancang untuk menangani data berukuran besar dengan perangkat keras komoditas. Di dalam arsitektur Hadoop, HDFS (Hadoop Distributed File System) berperan sebagai sistem penyimpanan terdistribusi. HDFS memecah data menjadi blok-blok kecil yang disimpan di setiap node, sehingga menghilangkan latensi jaringan dan menjamin throughput tinggi. Sebagai pelengkap, model pemrograman MapReduce digunakan untuk memproses data secara paralel. MapReduce membagi set data besar menjadi bagian-bagian kecil yang diproses bersamaan di berbagai node, kemudian menggabungkan hasil parsial menjadi output akhir. Dengan demikian, kombinasi HDFS dan MapReduce memungkinkan penanganan volume data perikanan yang besar secara efisien dan *scalable*.

Untuk memudahkan analisis data, Apache Hive diintegrasikan sebagai layer data warehouse. Hive menyediakan antarmuka SQL-like (HiveQL) yang memungkinkan peneliti menulis kueri analitis pada data di HDFS tanpa perlu menulis kode MapReduce secara manual. Misalnya, berbagai agregasi statistik produksi per provinsi atau jenis ikan dapat dihitung menggunakan HiveQL. Sementara itu, Apache HBase dipilih sebagai basis data NoSQL untuk menyimpan data hasil pengolahan secara kolom. HBase memanfaatkan HDFS sebagai lapisan penyimpanan dan dirancang untuk operasi baca/tulis acak dengan latensi rendah dan throughput tinggi. Hal ini memungkinkan akses cepat ke hasil ringkasan data (misalnya total produksi per provinsi) secara real-time. Perpaduan Hive dan HBase memungkinkan integrasi antara kueri analitik batch dan akses

data cepat pada skala besar, sehingga kelebihan masing-masing dapat dimanfaatkan dalam pipeline Hadoop.

Lingkungan implementasi Hadoop dirancang sebagai single-node cluster dengan kontainer Docker di Windows Subsystem for Linux (WSL2) pada sistem operasi Windows. Pada tahap ini, NameNode, DataNode, dan layanan Hadoop lainnya dijalankan di dalam kontainer Docker secara lokal untuk simulasi awal. Pendekatan ini memudahkan pengujian dan debug sebelum pengembangan ke lingkungan yang lebih besar. Skema modular ini pun mendukung perluasan ke multi-node jika diperlukan; misalnya dengan menggunakan Docker Compose untuk menambah kontainer worker baru. Dengan demikian, apabila data perikanan semakin bertambah atau kompleksitas analisis meningkat, arsitektur dapat diskalakan secara horizontal sesuai kebutuhan.

Alur data dimulai dari proses ingestion di mana dataset produksi perikanan Sumatera (format CSV) diunggah ke HDFS menggunakan perintah hdfs dfs -put dari sistem lokal. Setelah data tersedia di HDFS, tahap MapReduce dilakukan untuk transformasi awal, misalnya menghitung agregasi produksi total per provinsi atau jenis ikan. Pada tahap ini, fungsi map mengekstraksi entri data berdasarkan kunci (misalnya kode provinsi), sedangkan fungsi reduce menggabungkan hasil agregasi partial. Selanjutnya, Apache Hive digunakan untuk menjalankan kueri analitik pada data di HDFS. Tabel Hive didefinisikan dengan skema sesuai struktur CSV, sehingga kueri HiveQL (misalnya perhitungan rata-rata tahunan atau tren produksi) dapat dengan mudah dijalankan. Hasil penting seperti agregasi produksi yang telah dihitung kemudian disimpan ke dalam tabel HBase. Penyimpanan di HBase memfasilitasi akses acak dan real-time terhadap data ringkasan tanpa perlu memproses ulang seluruh dataset. Dengan demikian, setiap langkah (HDFS -> MapReduce -> Hive -> HBase) saling terintegrasi dalam satu pipeline Hadoop untuk analisis data perikanan.

Analisis deskriptif dilakukan dengan menghitung metrik statistik dasar dan pola pada data produksi. Misalnya, tim melakukan kueri Hive untuk memperoleh jumlah total produksi per provinsi, rata-rata produksi per tahun, serta grafik tren waktu produksi tiap jenis komoditas. Hasil-hasil ini membantu menginterpretasikan kondisi saat ini dan sebaran data secara keseluruhan. Untuk analisis prediktif, data historis digunakan untuk membangun model peramalan produksi masa depan. Teknik yang digunakan mencakup algoritma regresi atau time-series yang diimplementasikan di ekosistem Hadoop (misalnya menggunakan library seperti Apache Mahout atau modul pembelajaran mesin Hadoop lainnya). Model ini dilatih dan dievaluasi pada cluster, dengan mempertimbangkan karakteristik skala data. Arsitektur Hadoop mendukung skalabilitas horizontal, sehingga beban komputasi model prediktif dapat ditingkatkan sesuai kebutuhan. Hasil prediksi direpresentasikan sebagai proyeksi kuantitatif produksi perikanan di periode mendatang, yang dapat digunakan untuk pengambilan keputusan sektor perikanan di Sumatera.

D. Dataset

1. Jenis dan Sumber Data

   Dataset yang digunakan dalam analisis ini merupakan data sekunder kuantitatif yang bersumber dari website publikasi resmi Kementerian Kelautan dan Perikanan (KPP) dan Badan Pusat Statistik (BPS) Indonesia, terutama dari publikasi tahunan mengenai statistik perikanan nasional di berbagai provinsi. Data diambil dari dokumen yang memuat informasi Produksi Perikanan yang mencakup subsektor perikanan tangkap dan perikanan budidaya. Data ini bersifat publik dan telah melalui proses validasi dan standarisasi oleh instansi resmi pemerintah, sehingga dapat dijadikan dasar yang reliabel untuk keperluan analisis statistik dan spasial.

   Jenis data mencakup:

   - Perikanan Tangkap

     Data perikanan tangkap mencakup perikanan laut dan perairan umum daratan (sungai, danau, waduk, rawa dan genangan air lainnya). Dataset ini diklasifikasikan berdasarkan jenis penangkapan dan komoditas jenis ikan.
   - Perikanan Budidaya

     Data perikanan budidaya diklasifikasi berdasarkan jenis budidaya, kegiatan budidaya (Pembesaran, Pembenihan, ikan hias), dan lokasi budidaya (tambak, kolam, sawah, dan keramba).

2. Format dan Struktur Data

   Seluruh data produksi perikanan dari KKP dan BPS disediakan dalam bentuk tabel terstruktur dengan format CSV atau Excel. Dataset ini disusun secara longitudinal berdasarkan urutan waktu (time series) dan terstruktur secara hirarki berdasarkan provinsi, Sektor Perikanan, subsektor Jenis perikanan, tahun dan lainnya. Struktur tabel mencakup beberapa variabel utama seperti :

Variabel	Tipe Data	Deskripsi
Tahun	Integer	Tahun pencatatan data (2019-2023)
Provinsi	Kategori	Nama provinsi di wilayah Sumatera
Sub Sektor Perikanan	Kategori	Jenis perikanan (tangkap laut, tangkap umum, budidaya)
Jenis Ikan	Kategori	Komoditas utama ikan berdasarkan wilayah (bandeng, kerapu, udang, dan lainnya)
Teknik Penangkapan	Kategori	Teknik penangkapan pada sektor perikanan tangkap

Volume Produksi	Numerik (float)	Jumlah hasil perikanan dalam satuan ton
Nilai Produksi	Numerik (float)	Total nilai ekonomi dari hasil produksi perikanan dalam rupiah

Format data ini memungkinkan penerapan berbagai teknik analisis statistik deskriptif, komparatif antar wilayah, serta visualisasi berbasis waktu dan peta. Akan tetapi, beberapa variabel seperti Sektor dan subsektor perikanan, jenis ikan, teknik penangkapan tidak sepenuhnya tersedia dalam satu dataset terintegrasi. Oleh karena itu, pendekatan yang digunakan adalah menggabungkan informasi dari berbagai sumber serta menyebutkan keterbatasan dalam hasil analisis.

3. Rentang Waktu dan Wilayah

   Data yang digunakan mencakup periode lima tahun, mulai dari tahun 2019 hingga 2023. Rentang waktu lima tahun ini dipilih berdasarkan beberapa pertimbangan teknis dana analitis, yaitu :

   1. Ketersediaan dan kelengkapan data

      Tahun 2023 merupakan periode terbaru dengan data statistik resmi yang telah sepenuhnya dipublikasikan oleh KKP dan BPS. Sedangkan data tahun 2024 masih belum tersedia secara lengkap sehingga belum layak untuk digunakan sebagai dasar analisis yang akurat.
   2. Konsistensi dan kestabilan rentang waktu

      Lima tahun merupakan periode umum yang digunakan dalam analisis statistik berdasarkan rentang waktu, dimana jangka waktunya tidak terlalu panjang maupun terlalu pendek. Rentang waktu 2019-2023 juga relatif stabil secara administrasi dan kebijakan

   Dengan begitu, data yang digunakan ini cukup representatif untuk analisis tren dan perubahan produksi dalam jangka menengah dan relevan terhadap kondisi administrasi dan kebijakan saat ini.

   Dataset pada analisis ini juga mencakup seluruh provinsi di Pulau Sumatera, meliputi Aceh, Sumatera Utara, Sumatera Barat, Sumatera Selatan, Riau, Kepulauan Riau, Jambi, Bengkulu, Kepulauan Bangka Belitung, dan Lampung. Dengan cakupan wilayah ini, analisis dapat dilakukan pada skala provinsi maupun agregat regional untuk melihat pola distribusi spasial produksi perikanan di Sumatera.

E. Tahapan Kegiatan

Minggu Ke -	Fokus Kegiatan	Aktivitas Utama	Keluaran yang Diharapkan	Potensi Masalah & Mitigasi
1	Persiapan dan Perencanaan Awal	- Studi literatur (Hadoop dan kasus serupa). - Desain arsitektur (HDFS, MapReduce, Hive, HBase) - Instalasi WSL2, Docker, image Hadoop	- Lingkungan Hadoop lokal berjalan. - Dokumen desain arsitektur awal	- Konflik versi Java. - Kesalahan konfigurasi Docker. Mitigasi: Ikuti dokumentasi resmi & konsistensi versi
2	Persiapan Data	- Pemeriksaan dan pembersihan data CSV produksi perikanan. - Unggah ke HDFS (hdfs dfs -put)	- Data bersih tersimpan di HDFS	- Format tidak standar - Ukuran file besar Mitigasi: Praproses data & cek kestabilan koneksi
3	Pengembangan MapReduce	- Pembuatan skrip MapReduce (Java/Python Streaming) - Uji coba subset data - Eksekusi pada HDFS	- Skrip MapReduce & output agregasi di HDFS	- Bug kode - Kurang memori Mitigasi: Cek log error & uji skala kecil

4	Penerapan Hive	- Konfigurasi Hive - Definisi tabel sesuai skema - Query analitis (agregasi, tren, rata-rata)	- Tabel Hive terisi - Hasil analisis awal	- Skema salah - Query lambat Mitigasi: Revisi skema, partisi & indeks
5	Pengembangan HBase	- Instalasi HBase & desain tabel key-value. - Load data agregasi ke HBase - Uji integrasi Hive–HBase	- Tabel HBase terisi - Hive dapat query ke HBase	- Konfigurasi HBase (Zookeeper, heap) Mitigasi: Review config & uji baca/tulis
6	Analisis Lanjutan	- Visualisasi tren - Pembuatan model prediksi (Regresi, ARIMA, ML) - Evaluasi model	- Model prediksi awal - Matrik evaluasi (akurasi, MSE)	- Beban komputasi besar Mitigasi: Gunakan sampel, optimasi, atau tambahan node/cloud
7	Finalisasi & Dokumentasi	- Pengujian akhir pipeline lengkap - Perbaikan bug - Penyusunan laporan & presentasi	- Pipeline Hadoop lengkap - Laporan akhir & presentasi	- Waktu terbatas Mitigasi: Buffer waktu & alternatif komputasi tambahan

F. Target Output

Proyek ini ditujukan untuk menghasilkan sistem pipeline analitik produksi perikanan di Sumatera berbasis Hadoop. Target output utama proyek meliputi:

1. Lingkungan Hadoop Lokal Terintegrasi
   Cluster pseudo-distributed berbasis Docker dan WSL2 dapat menjalankan seluruh komponen Hadoop (HDFS, MapReduce, Hive, HBase). Berhasil melakukan Konfigurasi dan menyediakan dokumentasi lengkap untuk instalasi dan replikasi lingkungan.

2. Pipeline Pemrosesan Data Produksi Perikanan
   Dilakukan proses ingestion (input data) dan data CSV disimpan ke dalam HDFS. Dilanjutkan dengan pengolahan data berupa transformasi dan agregasi data menggunakan MapReduce atau Spark serta melakukan analisis deskriptif menggunakan HiveQL untuk memperoleh ringkasan statistik.

3. Integrasi Sistem Analitik
   Menyimpan hasil analitik dan ringkasan data secara kolom dan real-time menggunakan Apache HBase. Terbentuk integrasi Hive dengan HBase untuk kueri cepat dan fleksibel.

4. Model Prediksi Produksi Perikanan
   Menghasilkan model prediktif berbasis time-series atau regresi (misalnya ARIMA, Linear Regression) untuk memproyeksikan volume produksi masa depan berdasarkan data historis. Mengevaluasi performa model dengan metrik statistik seperti MAE atau MSE.

5. Dashboard dan Visualisasi Interaktif
   Menghasilkan dashboard visualisasi interaktif terkait tren produksi perikanan, perbandingan antar provinsi dan proyeksi masa depan.

6. Dokumentasi Proyek
   Laporan akhir yang berisi dokumentasi keseluruhan proses dan hasil proyek, serta rekomendasi yang dapat digunakan sebagai dasar pengambilan keputusan di sektor perikanan

G. Tools dan Teknologi Apache

Analisis produksi perikanan Sumatera memerlukan rangkaian teknologi terpadu yang andal menangani data berukuran besar. Serangkaian solusi open-source, khususnya dalam ekosistem Apache, dipilih agar tiap fase pipeline berjalan efisien, skalabel, dan mudah dipelihara. Rincian tiap kebutuhan beserta perangkat pendukungnya tersaji pada tabel berikut:

Lapisan/Tahap	Tools & Teknologi	Peran
Kontainer & Koordinasi	Docker, Docker Compose	Menyediakan lingkungan terisolasi dan portabel dan Docker Compose mempermudah penskalaan node worker Hadoop.
Penyimpanan Terdistribusi	Apache HDFS	Menyimpan dataset produksi perikanan dalam blok‑blok terdistribusi; menjamin throughput tinggi dan redundansi.
Pemrosesan Paralel	Apache MapReduce atau Apache Spark	Memecah data besar menjadi tugas kecil untuk diproses serempak, Spark dipertimbangkan jika analisis interaktif dibutuhkan.
Warehouse & Query SQL-like	Apache Hive (HiveQL)	Menyederhanakan analisis deskriptif melalui sintaks mirip SQL tanpa perlu menulis kode MapReduce mentah.
Penyimpanan NoSQL	Apache HBase	Menyimpan hasil agregasi secara acak (random read/write) dan real-time; terhubung langsung dengan Hive.
Machine Learning & Time Series	Apache Mahout, atau Python libraries (Scikit-Learn, Prophet)	Membangun model regresi/ARIMA untuk prediksi produksi; pilihan disesuaikan dengan kompleksitas dan ekosistem tim.
Koordinasi Workflow & Automasi	Apache Airflow	Menjadwalkan tahap ingestion -> MapReduce/Spark -> Hive -> HBase -> ML secara teratur dan terdokumentasi.
Integrasi & Ingestion Data	Python ETL scripts, Apache Sqoop/Flume	Memproses file CSV, membersihkan, dan memuat ke HDFS; Sqoop untuk migrasi SQL DB, Flume untuk stream real-time.
Visualisasi & BI	Apache Superset, Metabase, atau Grafana	Menampilkan dashboard tren produksi, heatmap provinsi-komoditas, maupun proyeksi model secara interaktif.

Versi & Kolaborasi	Git + GitHub/GitLab	Melacak perubahan skrip MapReduce/Spark, definisi Hive, notebook ML, serta file Airflow DAG.
Monitoring & Logging	Hadoop Metrics atau Prometheus + Grafana	Memantau NameNode/DataNode, penggunaan memori, serta performa job MapReduce/Spark.
Local Development (Windows)	WSL 2 + Docker Desktop	Menjalankan kluster pseudo-distributed secara lokal sebelum deployment ke server/VM berbasis Linux.

H. Anggota Kelompok dan Pembagian Tugas

No	Nama Anggota	Tugas
1	Dinda Joycehana (Infrastruktur & Koordinasi)	1. Menyiapkan kluster pseudo-distributed (WSL 2, Docker, Compose) dan YARN 2. Mengelola resource & monitoring 3. Optimasi performa node selama proyek
2	Elia Meylani Simanjuntak (Data Engineer)	1. Pra-pemrosesan & validasi file CSV kemudian ingestion ke HDFS 2. Menulis & menjalankan skrip MapReduce / Spark untuk agregasi awal
3	Presilia (Warehouse & NoSQL)	1. Mendesain skema Hive, partisi, indeks; membuat tabel eksternal 2. Instalasi & desain tabel HBase, integrasi Hive↔HBase
4	Randa Andriana Putra (Analitik & Workflow)	1. Membuat DAG Airflow: ingestion -> MR/Spark -> Hive -> HBase -> ML 2. Membangun model prediksi (ARIMA/Regresi) & visualisasi 3. Pengujian akhir pipeline

REFERENCES

Adnan, A. D. I., & Hasana, S. (2021). Implementasi Konsep Blue Economy di Indonesia dengan Memanfaatkan Teknologi Big Data. *SENSISTEK*, *4*, 25-33.

Hari Darmica. (2023, Januari 31). KELAUTAN BERBASIS BIG DATA DALAM MENGHADAPI ERA INDUSTRI 4.0 MARITIME-BASED BIG DATA IN FACING THE INDUSTRY 4.0 ERA. *Jurnal Kelautan dan Perikanan Terapan, Edisi Khusus 2023*, 81-85.

Kamaluddin, A., & Rachman, H. (n.d.). Analisis Tren Produksi Perikanan Berbasis Big Data di Sumatera. *Jurnal Data dan Sistem Informasi*, *6*, 87-95.

Smith, T., & Zhang, Y. (n.d.). A Comparative Study of Hadoop Ecosystem for Fisheries Data Management. *International Journal of Big Data Analytics*, *9*, 55-70.

Sri Endang Rahayu. (2022, November 30). Perkembangan Produksi Subsektor Perikanan di Sumatera Utara. *Jurnal Ilmu Ekonomi dan Studi Pembangunan*, *22*, 178-189.

Sugiarto, E., & Hartono, R. (2023). Optimalisasi Produksi Perikanan dengan Penerapan Teknologi Big Data. *Jurnal Teknologi dan Manajemen Kelautan*, *10*, 102-114.