Seminar Nasional Sains Data 2025 (SENADA 2025)
UPN "Veteran" Jawa Timur
E-ISSN 2808-5841
P-ISSN 2808-7283

Implementasi Ekosistem Hadoop untuk Analisis Segmentasi Pelanggan E-commerce di Pulau Sumatera

Khoirul Mizan abdullah[1], Kharisa Harvanny[2], Feryadi Yulius[3], Nabila Zakiyah Zahra[4], Ardika Satria[5], Luluk Muthoharoh[6], Vidia[7]

[1, 2,3,4,5,6,7]*Program Studi Sains Data, Fakultas Sains, Institut Teknologi Sumatera*
[1]khoirul.122450010@student.itera.ac.id [2]kharisa.122450061@student.itera.ac.id [3]feryadi.122450087@student.itera.ac.id
[4]nabila.122450139@student.itera.ac.id [5]ardika.satria@sd.itera.ac.id [6]luluk.muthoharoh@sd.itera.ac.id [7]vidia@sd.itera.ac.id

Corresponding author email: ardika.satria@sd.itera.ac.id

Abstract: *This study develops a big data solution based on the Hadoop ecosystem for customer segmentation analysis in e-commerce across the Sumatra region. A three-layer medallion architecture (bronze, silver, gold) is implemented, leveraging Sqoop for data integration, Spark SQL for data transformation, and MLlib for predictive modeling. In the bronze layer, raw data is stored in Parquet format on HDFS, then processed in the silver layer through data cleansing and RFM (Recency, Frequency, Monetary Value) feature extraction. In the gold layer, the K-Means algorithm is optimized using a combination of the Elbow Method and Silhouette Score to determine the optimal number of clusters, resulting in four distinct customer segments. The segmentation results are visualized using Apache Superset, providing an interactive dashboard for business analytics. The entire workflow is automated using Apache Oozie, with Apache Atlas for metadata management, and integration with Apache Ambari and ZooKeeper for real-time cluster monitoring. The findings demonstrate the system's capability to address large-scale e-commerce data processing challenges in Sumatra while establishing a solid foundation for developing more effective and data-driven marketing strategies.*
Keywords: *Big Data, Customer Segmentation, E-commerce Hadoop, K-Means, Medallion Architecture, RFM Analysis,*

Abstrak: Studi ini mengembangkan solusi big data berbasis ekosistem Hadoop untuk analisis segmentasi pelanggan *e-commerce* di wilayah Sumatera. Pendekatan arsitektur medallion tiga lapis (*bronze, silver, gold*) diimplementasikan dengan memanfaatkan teknologi *Sqoop* untuk integrasi data, Spark SQL untuk transformasi, dan MLlib untuk pemodelan prediktif. Pada lapisan *bronze*, data mentah disimpan dalam format *Parquet* di HDFS, kemudian diproses di lapisan silver melalui tahap pembersihan data dan ekstraksi fitur RFM (*Recency, Frequency, Monetary Value*). Pada lapisan gold, algoritma *K-Means* dioptimalkan menggunakan kombinasi Metode *Elbow* dan *Silhouette Score* untuk menentukan jumlah cluster optimal, menghasilkan empat segmen pelanggan yang berbeda. Visualisasi hasil segmentasi dikembangkan menggunakan *Apache* Superset, menyediakan dashboard interaktif untuk analisis bisnis. Seluruh alur kerja diotomatisasi melalui Apache *Oozie*, dengan dukungan Apache *Atlas* untuk manajemen metadata dan integrasi Apache Ambari serta *ZooKeeper* untuk pemantauan kluster secara real-time. Temuan penelitian membuktikan kemampuan sistem dalam mengatasi tantangan pengolahan data *e-commerce* skala besar di Sumatera, sekaligus menyediakan landasan yang kuat untuk pengembangan strategi pemasaran berbasis data yang lebih efektif dan terukur.
Kata kunci: Analisis RFM, Arsitektur *Medallion,* Big Data, *E-commerce,* Hadoop, *K-Means,* Segmentasi Pelanggan

I. PENDAHULUAN

Pertumbuhan *e-commerce* di Indonesia, terutama di wilayah Sumatera menjadikan kebutuhan atas pemanfaatan teknologi berskala besar dalam pengambilan keputusan yang tepat. Tantangan yang sering dihadapi oleh pelaku *e-commerce* adalah memahami karakteristik pelanggan [1]. Berdasarkan karakteristik potensi pelanggan disegmentasikan kedalam beberapa kelompok tertentu. Menggunakan *Silhouette Score* yang dihitung berdasarkan data, hingga didapatkan nilai segmentasi terbaik [2].

Merancang sistem big data berbasis ekosistem *Hadoop* untuk menganalisis segmentasi pelanggan *e-commerce* di Sumatera melalui pengumpulan data transaksi dan demografi [3], pembangunan *pipeline* ETL (*Extract, Transform, Load*), serta implementasi algoritma clustering dan visualisasi hasil guna mendukung rekomendasi bisnis berbasis data. Fokus utama meliputi optimasi penyimpanan data dalam HDFS dengan format *Parquet*, pemrosesan menggunakan *Spark SQL* dan *MLlib*, serta integrasi tools seperti *Spark, Sqoop,* dan *Hive* untuk membangun alur kerja terotomasi dari akuisisi data hingga analisis [4]. Sistem ini mencakup pengolahan data *batch* transaksi dan demografi pelanggan dengan arsitektur terdistribusi *Hadoop*, meliputi: lapisan akuisisi data menggunakan *Sqoop*, penyimpanan efisien di HDFS berformat *Parquet*, pemrosesan ETL dan segmentasi pelanggan dengan Spark dengan *K-Means* dan PCA, serta visualisasi hasil melalui dashboard analitik [5]. Seluruh alur kerja di otomatisasi dengan *Oozie*, didukung manajemen metadata *Apache Atlas* dan monitoring cluster dengan Ambari dan *Zookeeper*, memastikan skalabilitas dan *traceability* untuk kebutuhan bisnis *e-commerce* di wilayah Sumatera. Dengan menggunakan arsitektur data *medallion* yang terdiri dari 3 layer yaitu *bronze, silver, dan gold.*

Dilakukan segmentasi pelanggan pada pulau sumatera untuk mengelompokkan jenis-jenis pelanggan berdasarkan karakteristik demografisnya, mengoptimalisasi strategi pemasaran yang sesuai pada setiap segmen-segmen yang ada, dan meningkatkan customer experience berupa rekomendasi produk juga layanan terkait [6]. Disebabkan pada distribusi populasi yang belum merata di Pulau Sumatera, masih ada keterbatasan mengenai akses logistik, dan belum terlihat peluang pasar yang ter-eksplorasi secara penuh. Dibandingkan dengan pulau-pulau lain seperti Pulau Jawa dan Bali dengan segmentasi yang fokusnya pada *high-frequency shoppers* karena infrastruktur digital dan logistik sudah matang. *Ekosistem Hadoop* dipilih karena dataset transaksi *e-commerce* di Pulau Sumatera mencakup ribuan bahkan jutaan record per tahun, variabelnya yang kompleks, dan kecepatan akuisisinya yang tinggi dimana semua itu memenuhi syarat pemrosesan big data [7]. Dengan keunggulan skalabilitas untuk pemrosesan paralel di *cluster*, menjadi solusi open source yang mengurangi biaya infrastruktur, dan tersedianya integrasi alat ETL (*Extract, Transform, Load*) yang mempermudah ekstraksi data dari banyak sumber.

II. METODE PENELITIAN

Penelitian ini menggunakan beberapa metode untuk arsitektur, *pre-processing*, dan pengolahan data hingga akhirnya membentuk 4 segmentasi.

II.1 Data

Data yang digunakan merupakan data historis yang di-*generate* melalui *script* python yang telah disesuaikan dengan atribut yang diperlukan dan pola-pola aktivitas nyata yang terjadi dalam sistem *e-commerce* seperti pemesanan produk, ulasan pelanggan, dan interaksi antara penjual dan pembeli untuk mendukung analisis nantinya. Atribut yang digunakan terdiri dari beberapa tabel terpisah yang saling terkait melalui relasi kunci (*key*).

Tabel 1. Deskripsi Data

Nama Tabel	Atribut	Deskripsi
Data Pelanggan (*customers*)	id_pelanggan, kota_kabupaten_pelanggan, provinsi_pelanggan	Data ini penting untuk menganalisis distribusi geografis pelanggan serta mengaitkan perilaku pembelian dengan asal wilayah.
Data Penjual (*sellers*)	id_seller, kota_kabupaten_seller, provinsi_seller	Data ini digunakan untuk mengevaluasi kontribusi penjual terhadap volume penjualan berdasarkan lokasi mereka.
Data Produk (*product*)	id_produk, id_seller, kategori_produk, harga	Data ini digunakan untuk menganalisis popularitas produk berdasarkan kategori dan jenis.
Data Pesanan (*order_items*)	id_order, id_produk, harga, jumlah, total_item	Data ini digunakan untuk menganalisis volume penjualan dan perilaku pembelian.
Data Transaksi (*transaction*)	id_order, id_pelanggan, metode_pembayaran, banyak_cicilan, total_pembayaran, status_order, timestamp_pembelian, timestamp_persetujuan_toko, timestamp_pengiriman_ke_pelanggan, estimasi_sampai	Data ini digunakan untuk memahami preferensi metode pembayaran.
Data Ulasan (*reviews*)	id_pelanggan, id_produk, nilai_rating_produk, tanggal_review	Data ini digunakan untuk mengukur kepuasan pelanggan dan kualitas layanan.

II.2 Desain Sistem

Pengimplementasian arsitektur *medallion* merupakan pendekatan dalam pengelolaan data yang membagi alur pemrosesan data menjadi tiga lapisan utama:

- *Bronze* layer : menyimpan data mentah yang diimpor menggunakan *Sqoop* dan disimpan dalam HDFS dengan format *Parquet*.
- *Silver* layer : melakukan pembersihan data (*handling missing values*, deduplikasi) dan ditransformasi menggunakan *Spark* SQL termasuk fitur RFM untuk analisis perilaku belanja
- *Gold* layer : pemodelan data dengan algoritma *K-Means* dari *MLlib Spark* dengan optimasi *cluster* menggunakan *Elbow Method* dan *Silhouette Score*.

II.3 Proses Analisis Data

- *Pre-processing*
  - Standarisasi format data
  - Pembuatan fitur RFM (*Recency, Frequency, Monetary*)
- Pemodelan:
  - Implementasi algoritma *K-Means*.
  - Penentuan jumlah *cluster* optimal menggunakan *Elbow Method* dan *Silhouette Score*.

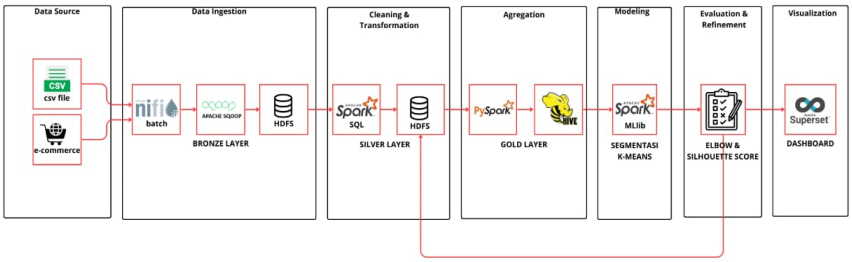
- Visualisasi
  - Pembuatan *dashboard* interaktif menggunakan Superset.

II.4. Implementasi Teknis

- Infrastruktur
  - *Cluster Hadoop*
  - *Konfigurasi Spark*
- Otomasi
  - Penggunaan *Apache Oozie* untuk *Workflow Management.*
  - *Apache Atlas* untuk metadata *management.*

II.5 Validasi

- Evaluasi kualitas *cluster* menggunakan *Silhouette Score*
- Validasi bisnis melalui interpretasi karakteristik tiap *cluster*



Gambar 1. *Pipeline* Alur Pengerjaan

II.6 Perhitungan dalam penggunaan *K-Means* dalam *MLlib.*

- Menghitung jarak *Euclidean*

$$d(x_i, C_j) = \sqrt{\sum_{m=1}^{k} (x_{i,m} - C_{j,m})^2} \quad \textbf{(1)}$$

- Menetapkan titik data ke *cluster* terdekat

$$C_j^{\text{new}} = \{x_i : d(x_i, C_j) \le d(x_i, C_m) \text{ untuk semua } m \ne j\} \quad \textbf{(2)}$$

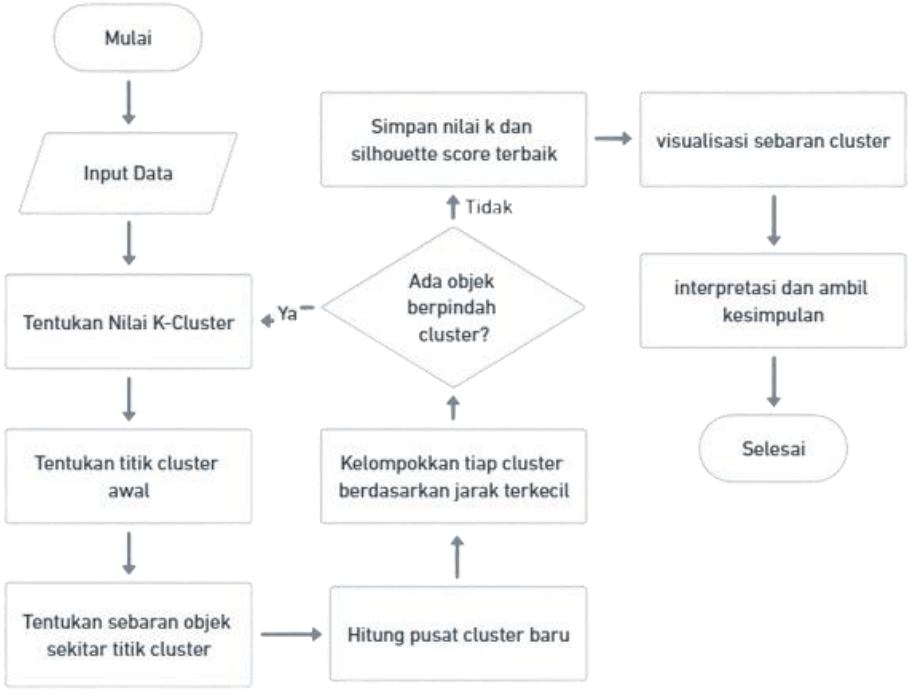
- Mengupdate pusat *cluster*

$$C_j^{\text{new}} = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i \quad \textbf{(3)}$$

- Konvergensi

$$\left\| C_j^{\text{new}} - C_j^{\text{old}} \right\| < \varepsilon \quad \textbf{(4)}$$

II.7 Diagram Alir



Gambar 2. Diagram Alir

III. HASIL DAN PEMBAHASAN

III.1 Implementasi *Pipeline Big* Data dengan Arsitektur Medallion

Penerapan sistem arsitektur *medallion* dengan 3 lapisan utama, bronze, silver, dan gold dalam sistem *Apache Hadoop.* Setelah dikumpulkan secara *batch* dari *file* .csv dan API sistem *e-commerce* data masuk ke layer berikut.

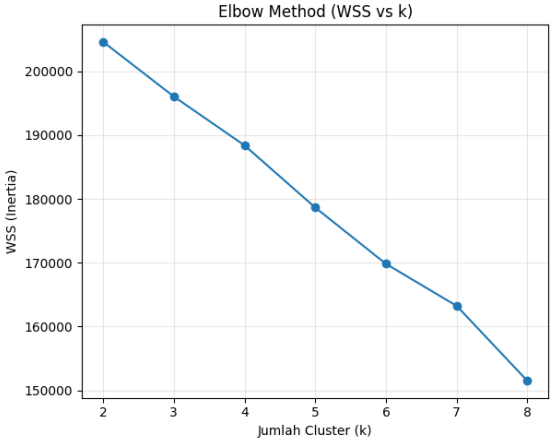
- *Bronze* *Layer*
  Data mentah dimuat ke HDFS dalam format .csv di akomodasi dari sumber heterogen.
- *Silver* *Layer*
  Menggunakan *Spark* SQL dihasilkan data berformat *parquet* sebagai bentuk efisiensi. Data dinormalisasi, dihapus apabila terdapat data duplikat, dan dibuat metrik RFM (*Recency, Frequency, Monetary*).
- *Gold* *Layer*
  Dilakukan *aggregate* lalu disimpan kedalam tabel Hive, digunakan query untuk selanjutnya dianalisis.

III.2 Segmentasi Pelanggan dengan Spark MLlib

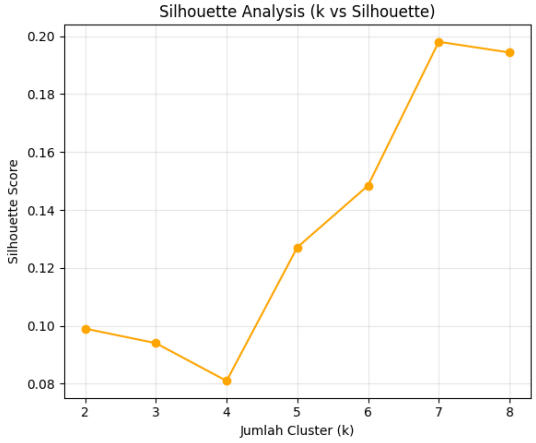
Dengan *K-Means Clustering* dari *Spark MLlib* dan penentuan jumlah *cluster* optimal dengan metode analisis *silhouette score* dan grafik elbow *Within-Cluster Sum of Squares* (WSS). Dengan nilai tertinggi K = 7 namun setelah dianalisis kembali didapatkan K = 4 untuk hasil segmentasi paling representatif secara visual dan interpretatif dalam konteks bisnis.

E-ISSN 2808-5841
P-ISSN 2808-7283

Setelah dilakukan pemilihan jumlah cluster optimal dilakukan menggunakan Elbow Method dan Silhouette Analysis. Berdasarkan *Elbow Method* pada Gambar 3a. terlihat titik siku pada K = 4 kemudian dilanjut K = 5, yang menunjukkan penurunan WSS mulai melambat, menandakan bahwa penambahan klaster setelah nilai tersebut kurang memberikan manfaat signifikan. Sementara itu, analisis *Siluet Score* pada Gambar 3b. menunjukkan skor tertinggi pada K = 7 dan K = 8. Namun, untuk menjaga keseimbangan antara interpretabilitas dan pemisahan klaster, nilai K = 4 dipilih sebagai jumlah *cluster* optimal.



Gambar 3a. Elbow Method untuk Memilih k

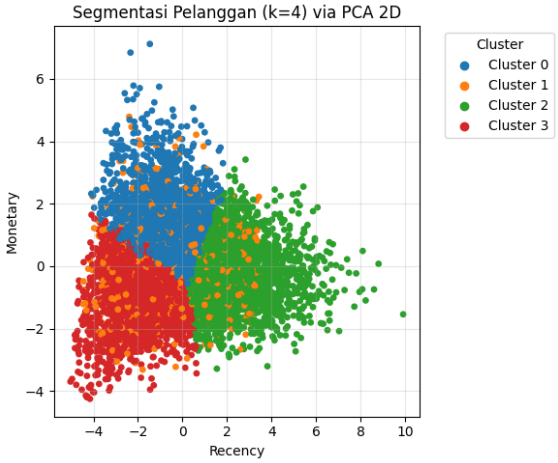


Gambar 3b. Silhouette Analysis untuk Memilih k

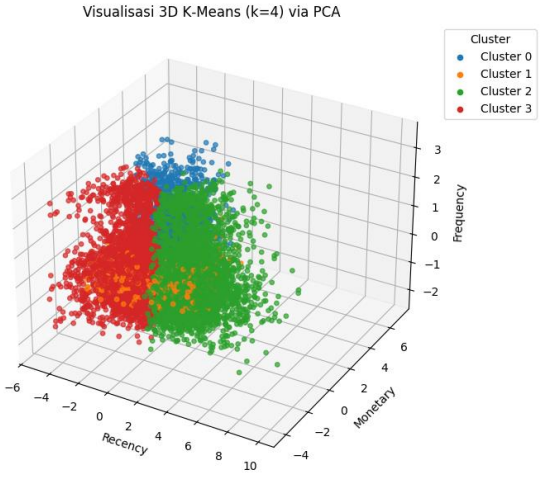
Lalu pada Gambar 4a. dan Gambar 4b. memvisualisasikan segmentasi pelanggan menggunakan algoritma *K-Means* dengan jumlah *cluster* sebanyak 4. Visualisasi ini dilakukan dengan Principal Component Analysis (PCA) untuk mereduksi dimensi data menjadi dua komponen utama untuk 2D dan tiga komponen utama untuk 3D, sehingga data yang semula berdimensi tinggi dapat divisualisasikan dalam bidang dua dimensi dan tiga dimensi. Sumbu X mewakili komponen yang berkaitan dengan *recency* (seberapa baru pelanggan melakukan transaksi), sumbu Y mewakili komponen yang berkaitan dengan *monetary* (total pembayaran oleh pelanggan), dan sumbu Z mewakili komponen yang berkaitan dengan *frequency* (banyak barang yang dibeli). Titik - titik pada plot mewakili masing-masing pelanggan yang telah dikelompokkan ke dalam empat segmen berbeda dan diwarnai berdasarkan label clusternya.

Cluster biru tua menunjukkan bahwa kelompok pelanggan yang baru saja melakukan pembelian dan juga mengeluarkan uang dalam jumlah besar. Mereka cenderung sering belanja dan berkontribusi besar terhadap pendapatan perusahaan. Untuk *Cluster* merah mereka juga baru-baru ini melakukan pembelian, tetapi pengeluarannya tidak terlalu besar, kemungkinan besar mereka adalah pembeli baru. *Cluster* hijau merupakan kelompok dengan *recency* tinggi namun *monetary* sedang hingga tinggi. Mereka dulunya pernah melakukan pembelian besar, namun cukup lama tidak aktif, ini menandakan adanya pelanggan bernilai tinggi yang mulai tidak aktif. Sedangkan *cluster* orange terdiri dari pelanggan dengan recency dan monetary normal, mereka adalah segmen yang memiliki perilaku belanja paling bervariasi. Pelanggan seperti ini termasuk dalam kategori pelanggan tetap.

E-ISSN 2808-5841
P-ISSN 2808-7283

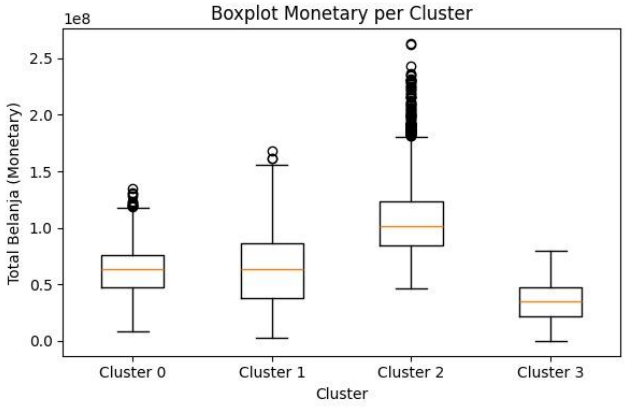


Gambar 4a. Segmentasi Pelanggan 4 cluster: PCA 2D



Gambar 4b. Segmentasi Pelanggan 4 cluster: PCA 3D

Pada Gambar 5. Boxplot *Monetary* per *Cluster* memberikan gambaran tentang perbedaan total belanja pelanggan di tiap klaster hasil segmentasi. Dari grafik terlihat bahwa cluster 2 memiliki median total belanja tertinggi dan persebaran yang luas, bahkan banyak pelanggan dalam cluster ini merupakan big spender dengan total belanja sangat tinggi. Sementara itu, cluster 0 dan cluster 1 memiliki median yang relatif serupa pada kisaran menengah, namun cluster 1 memiliki sebaran yang lebih besar menunjukkan adanya pelanggan dengan perilaku belanja yang lebih bervariasi. Keduanya merupakan target yang baik untuk strategi penguatan loyalitas. *Cluster* 3 merupakan kelompok dengan total belanja paling rendah baik dari sisi median maupun sebaran datanya. Menunjukkan bahwa pelanggan di *Cluster* 3 belum terlalu aktif atau loyal, dan bisa menjadi sasaran untuk kampanye promosi atau edukasi agar lebih sering dan lebih besar dalam bertransaksi.

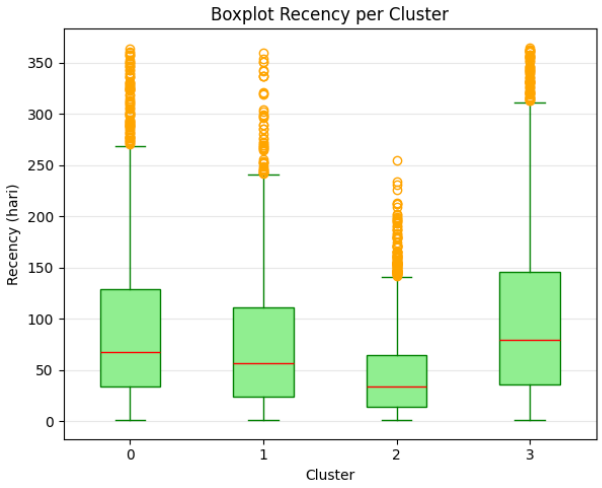


Gambar 5. Boxplot Monetary per Cluster

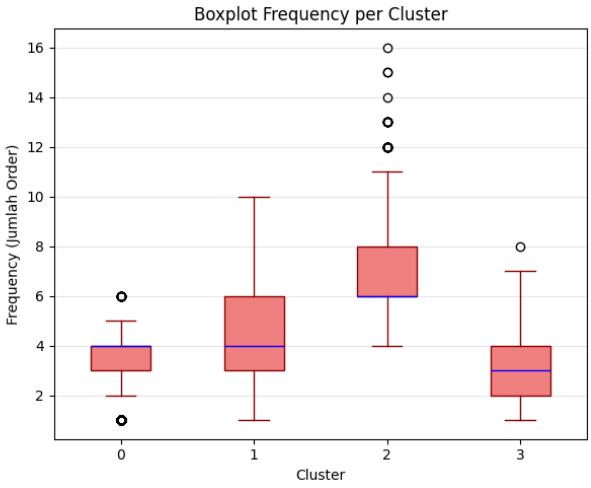
Pada Gambar 6. Serangkaian boxplot yang menggambarkan distribusi variabel *frequency* dan *recency* dalam satuan hari untuk setiap klaster yang dianalisis. Pada boxplot Recency, terlihat variasi yang signifikan antar kluster. *Cluster* 1 menunjukkan nilai median *recency* terendah (5 hari) dengan rentang interkuartil (IQR) yang relatif sempit, mengindikasikan bahwa sebagian besar anggota cluster ini memiliki waktu interaksi terakhir yang baru dan konsisten. Sebaliknya, kluster 3 menampilkan median *recency* tertinggi (30 hari) dengan IQR yang lebih lebar, menunjukkan variasi yang besar dalam waktu interaksi terakhir dan kecenderungan umum yang kurang aktif. Namun pada *recency* tinggi di

kluster 2 dan 3 terdapat beberapa outlier yang merepresentasikan kasus khusus dengan penanganan tersendiri. Boxplot *frequency* mengungkapkan pola kluster 1 tidak hanya unggul dalam *recency* tetapi juga menunjukkan nilai median *frequency* tertinggi (15 interaksi), dengan distribusi yang cenderung simetris. Kluster 2 menampilkan karakteristik menengah dengan median *frequency* 8 interaksi dan IQR moderat. Sementara itu, kluster 3 memiliki performa terendah dengan median *frequency* hanya 2 interaksi dan adanya beberapa outlier di ujung atas, yang mungkin mengindikasikan sub-kelompok dengan pola penggunaan yang tidak biasa.

Analisis komparatif mengenai kedua boxplot menyatakan adanya korelasi negatif antara *recency* dan *frequency*. Kluster dengan *recency* rendah cenderung memiliki *frequency* tinggi, dan sebaliknya. Temuan ini konsisten dengan teori perilaku konsumen dimana pengguna yang lebih aktif (*frequency* tinggi) secara alami akan memiliki *recency* yang lebih baru. Perbedaan karakteristik antar kluster ini membuktikan efektivitas metode *clustering* yang digunakan dalam membedakan kelompok berdasarkan pola interaksi. Perbedaan *recency* dan *frequency* menunjukkan bahwa segmentasi ini efektif dalam membedakan kelompok berdasarkan tingkat keterlibatan. Hasil ini dapat digunakan untuk menyusun strategi yang lebih terarah, seperti memberikan promosi khusus kepada *cluster* aktif atau program loyalitas untuk *cluster* yang kurang terlibat.



Gambar 6a. Boxplot Recency per Cluster



Gambar 6b. Boxplot Frequency per Cluster

IV. KESIMPULAN

Studi ini berhasil mengimplementasikan ekosistem *Hadoop* untuk membangun suatu pipeline analitik terdistribusi yang skalabel guna melakukan segmentasi pelanggan *e-commerce* di wilayah Sumatera. Arsitektur berbasis *medallion* dengan teknologi inti seperti HDFS untuk penyimpanan, *Spark* untuk pemrosesan, *Hive* untuk analitik, dan *NiFi* untuk integrasi data, telah terbukti mampu menangani pemrosesan data batch dalam skala besar secara efisien. Pada tahap pemodelan, penerapan algoritma *K-Means* yang dioptimasi dengan *Silhouette Score* berhasil mengidentifikasi segmen pelanggan yang signifikan secara statistik dan relevan secara bisnis berdasarkan pendekatan RFM (*Recency, Frequency, Monetary*). Hasil visualisasi tidak hanya memvalidasi temuan analitik tetapi juga memberikan landasan empiris yang kuat untuk pengambilan keputusan strategis di tingkat manajerial.

Implementasi sistem terotomasi ini menunjukkan efektivitas yang tinggi dalam mengatasi tantangan kompleks industri *e-commerce* di wilayah dengan karakteristik geografis dan infrastruktur

yang beragam seperti Sumatera. Selain memberikan wawasan bisnis yang *actionable*, solusi yang dikembangkan juga memiliki fleksibilitas tinggi untuk diaplikasikan pada berbagai wilayah geografis dan domain bisnis lainnya.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada Bapak Ardika Satria, S.Si., M.Si., , Ibu Luluk Muthoharoh, S.Si., M.Si., dan Ibu Vidia, M.Si., yang telah membimbing serta memberikan pengetahuan dan arahan selama perkuliahan mata kuliah Analisis Big Data, yang menjadi dasar penting dalam penyusunan paper ini.

Ucapan terima kasih juga penulis sampaikan kepada teman-teman yang telah memberikan semangat dan dukungan selama proses penulisan. Secara khusus, penulis mengapresiasi tim SENADA yang telah meluangkan waktu untuk menyelenggarakan kegiatan ini dan memberikan kesempatan serta masukan berharga dalam proses review agar paper ini menjadi lebih baik.

REFERENSI

[1] D. V. N. Hasibuan dan M. I. P. Nasution, "Penerapan Big Data dalam Pemasaran Digital: Studi Kasus pada Industri E-commerce di Indonesia," *Jurnal Ilmiah Nusantara (JINU)*, vol. 1, no. 4, pp. 776–783, Jul. 2024, doi: 10.61722/jinu.v1i4.1913.

[2] S. Septiani, M. A. Akhyar, dan P. Seviawani, "Penggunaan Big Data untuk Personalisasi Layanan dalam Bisnis E-Commerce," *Prosiding Penelitian Nasional Teknologi Informasi dan Komunikasi*, 2024.

[3] S. S. Vamsi, K. Deeksha, S. A. Varshini, J. Srija, S. J. Kanna, P. Srihari, dan U. Archana, "E-Commerce Data Analysis Using Hadoop," *International Journal of Research Publication and Reviews*, vol. 3, no. 4, pp. 2235–2240, Apr. 2022.

[4] M. Sewak dan S. Singh, "A Reference Architecture and Road map for Enabling E-commerce on Apache Spark," *Communications on Applied Electronics (CAE)*, vol. 2, no. 1, pp. 37–42, Jun. 2015. [Online]. Available: www.caeaccess.org

[5] T. F. Febriana dan M. I. Rosadi, "Analisis Segmentasi Pelanggan Toko Sepatu di E-Commerce Shopee Berdasarkan Model FRM (Recency, Frequency, Monetary) Menggunakan Algoritma K-Means Klastering," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 6, pp. 12069–12072, Des. 2024.

[6] F. H. Pasaribu, "Studi Kasus: Penggunaan Teknologi Big Data dalam Analisis Pemodelan Perilaku Konsumen," *Jurnal Teknik Universitas Medan Area*, 2024.

[7] F. A. Lubis dan M. I. P. Nasution, "Penggunaan Teknologi Big Data untuk Analisis Prediksi Bisnis," *Jurnal Ilmiah Nusantara (JINU)*, vol. 1, no. 4, pp. 667–672, Jul. 2024, doi: 10.61722/jinu.v1i4.1882.