Jurnal Teknologi Sistem Informasi dan Aplikasi **ISSN: 2654-3788**
Penerbit: Fakultas Ilmu Komputer Universitas Pamulang **e-ISSN:2654-4229**
Vol. 09, No. 02, April 2026 (149-159) **DOI: https://doi.org/10.32493/jtsi.v9i2.59043**

Implementasi Pipeline ETL Otomatis Berbasis Container untuk Pengelolaan Data Komentar YouTube Kasus Gagal Ginjal di Indonesia

Rafita Kusuma[1*], Tikaridha Hardiani[2]

[1 2]Program Studi Teknologi Informasi, Fakultas Sains dan Teknologi, Universitas Aisyiyah Yogyakarta
e-mail:[1]rafita0883@gmail.com*, [2]tikaridha@unisayogya.ac.id

Submitted Date: January 7th, 2026 Reviewed Date: January 30th, 2026
Revised Date: March 5th, 2026 Accepted Date: April 20th, 2026

Abstract

Data from social media platforms like YouTube has the potential to support analysis of health issues, including kidney failure, but is generally unstructured and requires data processing. This study aims to build a container-based automated Extract, Transform, Load (ETL) pipeline using Apache NiFi to manage YouTube comment data related to kidney failure. Data was collected using the YouTube API based on keywords related to kidney failure with a total of 30,000 comments. The implementation results show that the ETL pipeline is capable of running the process automatically with a throughput of 135 comments per second. In addition, data quality was improved through the removal of empty and duplicate data and an increase in data relevance to the topic of kidney failure from 64.96% to 89.81%. The results show that the container-based Apache NiFi integration is capable of producing a cleaner, more structured, and more relevant dataset to support health analysis based on social media data.

Keywords: ETL, Apache NiFi, YouTube API, containers, kidney failure.

Abstrak

Data dari media sosial seperti YouTube memiliki potensi untuk mendukung analisis isu kesehatan, termasuk gagal ginjal, namun umumnya data masih tidak terstruktur sehingga memerlukan proses pengolahan data. Penelitian ini bertujuan membangun pipeline *Extract, Transform, Load* (ETL) otomatis berbasis container menggunakan Apache NiFi untuk mengelola data komentar YouTube terkait gagal ginjal. Data dikumpulkan menggunakan API YouTube berdasarkan kata kunci terkait gagal ginjal dengan total 30.000 komentar. Hasil implementasi menunjukkan bahwa pipeline ETL mampu menjalankan proses secara otomatis dengan *throughput* 135 komentar per detik. Selain itu, kualitas data meningkat melalui penghapusan data kosong dan duplikat serta peningkatan relevansi data terhadap topik gagal ginjal dari 64,96% menjadi 89,81%. Hasil penelitian menunjukkan bahwa integrasi Apache NiFi berbasis container mampu menghasilkan dataset yang lebih bersih, terstruktur, dan relevan untuk mendukung analisis kesehatan berbasis data media sosial.

Kata kunci: ETL, Apache NiFi, API YouTube, container, gagal ginjal.

1. Pendahuluan

Perkembangan teknologi informasi telah menghasilkan volume data digital yang sangat besar dari berbagai sumber termasuk media sosial dan platform berbagi video seperti YouTube (Khoirunnisa et al., 2025). Data tersebut memiliki potensi besar untuk dianalisis guna mendukung penelitian, pengambilan keputusan, maupun pemantauan isu kesehatan masyarakat. Namun, data yang didapat dari media sosial biasanya tidak terstruktur (Walha et al., 2024). Kondisi ini menyebabkan data tidak bisa langsung digunakan untuk analisis seperti analisis sentimen atau pemetaan pendapat publik,

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2026 Rafita Kusuma, Tikaridha Hardiani

karena harus melewati proses pengolahan terlebih dahulu. Pemanfaatan teknologi informasi di sektor kesehatan terbukti dapat meningkatkan efisiensi operasional dan mutu layanan melalui pengelolaan data (Cahyani & Wijayanto, 2025).

Data Riskesdas menunjukkan prevalensi sekitar 0,38% penduduk Indonesia mengalami gagal ginjal, atau ratusan ribu orang terdampak. Jumlah pasien gagal ginjal kronik (GGK) yang menjalani hemodialisis terus bertambah, mencapai 134.057 pasien pada tahun 2024 dengan biaya pengobatan mencapai Rp 11 triliun. Besarnya jumlah penderita menunjukkan pentingnya pemanfaatan data digital sebagai sumber informasi tambahan untuk memahami persepsi publik maupun penyebaran informasi terkait penyakit tersebut (Nasution et al., 2025). Dalam pengelolaan data digital, proses utama untuk mengelola data disebut *Extract*, *Transform*, dan *Load* (ETL) (Ghani & Kurniawan, 2024).

ETL merupakan tahapan penting untuk memastikan data mentah dapat dikumpulkan, disaring, diubah, dan digabung sesuai dengan kebutuhan analisis (Singu, 2022). Apache NiFi merupakan salah satu platform *open-source* yang tanpa memerlukan pengkodean yang kompleks (Saputra et al., 2025).

Lingkungan eksekusi pipeline ETL yang melibatkan berbagai komponen memerlukan pendekatan yang mampu menjaga keseragaman lingkungan operasional. Docker container merupakan unit perangkat lunak yang mengemas kode aplikasi beserta pustaka dan dependensinya sehingga dapat dijalankan secara konsisten di berbagai lingkungan komputasi (Hutabalian, 2025). Teknologi ini memanfaatkan virtualisasi tingkat sistem operasi untuk mengisolasi proses serta mengatur penggunaan sumber daya seperti CPU dan memori, sehingga meningkatkan eksekusi. Penggunaan container menjadi penting dalam pengembangan sistem modern karena memungkinkan aplikasi dijalankan pada berbagai infrastruktur tanpa perubahan konfigurasi (Yepuri et al., 2023).

Penelitian mengenai penerapan proses ETL telah banyak dilakukan, penelitian sebelumnya menunjukkan bahwa penerapan ETL dapat memperbaiki kualitas data secara signifikan dalam meningkatkan kualitas data sebelum dianalisis (Bono et al., 2023; Sandiya et al., 2026). Seiring berkembangnya kebutuhan pengolahan data dalam skala besar, pipeline data modern juga mulai memanfaatkan platform integrasi data yang mampu mendukung pemrosesan secara otomatis. Salah satu penelitian membandingkan Apache NiFi dengan Talend dan Informatica PowerCenter pada proses otomatisasi ETL menunjukkan bahwa Apache NiFi lebih unggul dalam pengelolaan aliran data secara *real-time* dan *visual flow-based programming* (Singu, 2022).

Selanjutnya, studi menyebutkan Apache NiFi merupakan teknologi yang banyak digunakan karena mampu mengotomatisasi perpindahan data secara *real-time* dan terstruktur dalam pipeline serta mendukung antarmuka pipeline (Wakchaure, 2023). Hal ini diperkuat bahwa Apache NiFi digunakan untuk membangun sistem distribusi data berkinerja tinggi pada jaringan kompleks seperti 5G/6G, menunjukkan kemampuan untuk menangani aliran data besar secara *real-time* (Wnęk & Boryło, 2023).

Di sisi lain, pendekatan berbasis container juga mengalami perkembangan pesat. Penelitian menunjukkan bahwa container menawarkan keuntungan dalam hal efisiensi, skalabilitas, dan kemudahan penerapan dalam sistem terdistribusi (Mishra et al., 2022). Kemudian, penelitian menyebutkan bahwa penggunaan container meningkatkan efisiensi sumber daya memungkinkan isolasi layanan dan kemudahan pemindahan lintas lingkungan, sehingga mendukung pengelolaan sistem terdistribusi secara lebih efektif (Yepuri et al., 2023). Hal ini diperkuat oleh studi yang menunjukkan container seperti Docker dalam arsitektur *microservices* lebih efisien dibanding *virtual machine* karena mampu menjaga performa penggunaan CPU dan memori tetap stabil karena bersifat ringan (Liu & Guitart, 2022; Marella, 2024), Adapun pada optimasi container sistem dalam lingkungan terdistribusi mampu menangani lebih dari 2000 *concurrent requests*, menurunkan penggunaan CPU hingga lebih dari 1,5%, serta menurunkan penggunaan memori sekitar 0,6%. Selain itu, optimasi tersebut juga dapat mengurangi rata-rata waktu permintaan sebesar 7,6% dan menurunkan tingkat kegagalan request hingga 32,4% (Mondal et al., 2024).

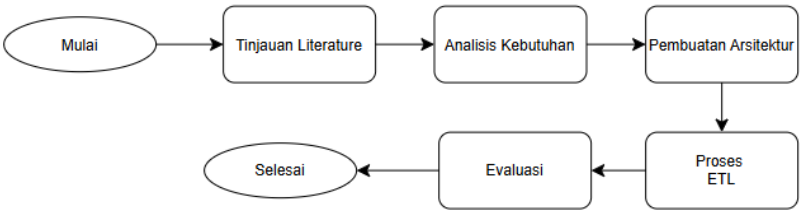
Penelitian sebelumnya hanya berfokus pada penggunaan Apache NiFi sebagai pipeline data dan penggunaan container dalam arsitektur secara terpisah. Penelitian ini berfokus pada penerapan Apache NiFi dalam lingkungan container untuk membangun pipeline ETL pengambilan data isu gagal ginjal di YouTube. Selain itu, Penelitian ini mengevaluasi kinerja pipeline berdasarkan beberapa aspek, yaitu *success rate*, waktu pemrosesan, *throughput* pipeline ETL, serta kualitas data berdasarkan relevansi terhadap topik yang dianalisis.

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2026 Rafita Kusuma, Tikaridha Hardiani

ISSN: 2654-3788
e-ISSN:2654-4229
DOI: https://doi.org/10.32493/jtsi.v9i2.59043

2. Metode Penelitian

Metode penelitian sistematis yang digunakan untuk mencapai tujuan penelitian, memastikan hasil yang diperoleh dapat dipertanggungjawabkan secara ilmiah (Nur et al., 2024). Penelitian ini menggunakan pendekatan pipeline ETL otomatis berbasis container untuk pengolahan data kesehatan dari media sosial. Tahap penelitian yang dilakukan dapat dilihat pada Gambar 1.



Gambar 1. Tahapan Penelitian

a. Tinjauan Literatur

Studi literatur dilakukan dengan membaca dan meninjau beberapa jurnal nasional dan internasional yang relevan dengan topik penelitian. Penelitian ini menggunakan pendekatan dari berbagai penelitian (Mnsen et al., 2024). Sumber literatur yang dikaji berfokus pada penelitian terbaru periode 2022–2026 terkait pemanfaatan teknologi Apache NiFi, otomatisasi pipeline, pemrosesan integrasi berbasis container, identifikasi kesamaan dan perbedaan dengan penelitian ini, serta penyusunan daftar pustaka yang menjadi landasan teori penelitian.

b. Analisis Kebutuhan

1. Kebutuhan Sistem

Kebutuhan sistem sangat penting untuk mendukung kelancaran penelitian mencakup berbagai sumber daya, baik berupa perangkat keras maupun perangkat lunak (Wijayanto & Firdonsyah, 2023). Penelitian ini menggunakan Docker container untuk menjalankan pipeline ETL. Spesifikasi perangkat yang digunakan dalam penelitian ditunjukkan pada Tabel 1,

Tabel 1. Spesifikasi Perangkat

Komponen	Spesifikasi
CPU	Intel Core i3
RAM	8 GB
SSD	512 GB

sedangkan daftar container yang digunakan ditunjukkan pada Tabel 2.

Tabel 2. Daftar container

Image	Port	Fungsi
Apache NiFi	8443	*Data ingestion* (mengambil dan mengalirkan data).
phpMyAdmin	8080	Antarmuka pengolahan basis data.
MySQL	3306	basis data.
Python_prosesor	-	*Transform* dan pengolahan data.

Apache NiFi digunakan untuk mengatur aliran data dari proses *Extract* hingga distribusi data. MySQL digunakan sebagai media untuk *Load* hasil ETL, sedangkan phpMyAdmin digunakan untuk

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2026 Rafita Kusuma, Tikaridha Hardiani

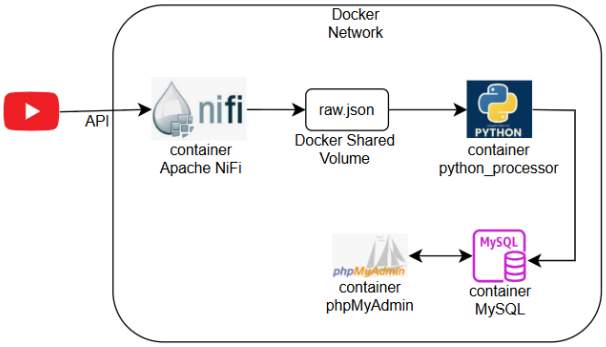
ISSN: 2654-3788
e-ISSN:2654-4229
DOI: https://doi.org/10.32493/jtsi.v9i2.59043

membantu pengelolaan basis data. Python_prosesor digunakan untuk menjalankan proses *Transform* data sebelum dimuat ke basis data.

2. Sumber Data

Sumber data yang digunakan dalam penelitian ini adalah data komentar dari video di platform YouTube yang diperoleh menggunakan YouTube Data API v3. Setiap akun API memiliki batas pengambilan 10.000 data per API, sehingga penelitian ini menggunakan tiga akun untuk memperoleh hingga 30.000 data. Data komentar dikumpulkan dari video yang diperoleh menggunakan kata kunci "gagal ginjal," "gagal ginjal pada anak," dan "gejala gagal ginjal". Data dikumpulkan dari video yang dipublikasikan pada periode 2019–2026. Setiap permintaan API memiliki batas komentar, sehingga proses pengambilan data dilakukan secara bertahap sesuai kuota API yang tersedia.

c. Pembuatan Arsitektur



Gambar 2. Arsitektur Pipeline ETL

Gambar 2 menunjukkan arsitektur pipeline ETL berbasis container yang digunakan dalam penelitian. Pipeline ETL dijalankan pada Docker *Network* yang terdiri dari container Apache NiFi, Python_prosesor, MySQL, dan phpMyAdmin. Apache NiFi digunakan pada proses *Extract* untuk mengambil data komentar YouTube melalui API dan menyimpan data dalam format *JSON* pada Docker *Shared Volume*. Data kemudian diproses oleh Python_prosesor pada proses *Transform* sebelum dimuat ke basis data MySQL pada proses *Load*. MySQL dipilih karena mudah diintegrasikan dengan teknologi container seperti Docker (Sholihah & Darujati, 2022). phpMyAdmin digunakan sebagai antarmuka berbasis web untuk membantu pengelolaan basis data MySQL.

Seluruh container dijalankan menggunakan *docker-compose* sehingga layanan dapat dikelola dalam satu lingkungan terintegrasi. Setiap container saling terhubung melalui Docker *Network* agar dapat berkomunikasi selama proses ETL berlangsung. Selain itu, Docker *Shared Volume* digunakan untuk menyimpan data secara permanen dan mendukung pertukaran file antar container (Setyaji & Wijayanto, 2025).

Apache NiFi bekerja dengan menyusun prosesor yang memungkinkan aliran data berjalan secara otomatis dalam sebuah pipeline (Wakchaure, 2023). Pada penelitian ini, Apache NiFi dijalankan sebagai container dan diintegrasikan dengan layanan lain dalam satu jaringan. Pipeline ETL dikonfigurasi menggunakan prosesor seperti *InvokeHTTP, SplitJSON, EvaluateJSONPath*, dan *PutFile* untuk mengambil, memproses, dan menyimpan data. Proses berjalan otomatis melalui hubungan antar prosesor dan pengaturan penjadwalan, serta pemantauan melalui *log* container untuk memastikan aliran data berjalan dengan baik.

d. Proses ETL

Pada proses *Extract*, Apache NiFi mengambil data komentar YouTube melalui API YouTube menggunakan prosesor seperti *InvokeHTTP* dan *EvaluateJSONPath.* Data yang diperoleh kemudian disimpan dalam format *JSON* menggunakan *PutFile.*

Proses *Transform* dilakukan pada Python_prosesor untuk membersihkan dan menyesuaikan data sebelum disimpan ke basis data MySQL. Proses *Transform* meliputi penghapusan URL, *case folding*, normalisasi teks, penghapusan data kosong, serta penanganan data duplikat berdasarkan identifier *id_str*.

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2026 Rafita Kusuma, Tikaridha Hardiani

ISSN: 2654-3788
e-ISSN:2654-4229
DOI: https://doi.org/10.32493/jtsi.v9i2.59043

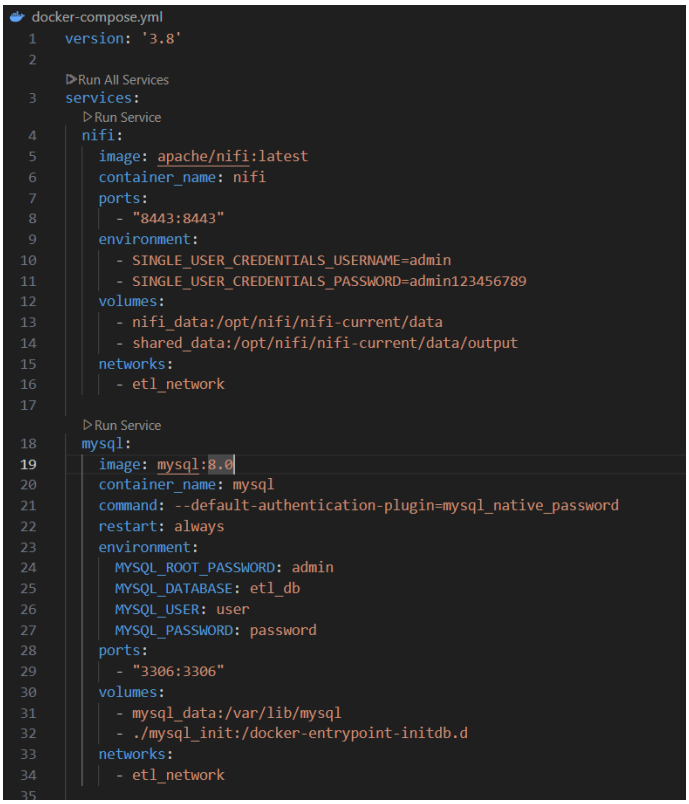
Proses *Load* dilakukan dengan memuat data hasil proses *Transform* ke basis data MySQL. Seluruh proses ETL berjalan secara otomatis melalui pengaturan aliran data pada Apache NiFi dan komunikasi antar-container dalam lingkungan container.

e. Evaluasi

Tahap evaluasi dibutuhkan untuk melihat kinerja keseluruhan untuk memastikan proses berjalan dengan baik. Evaluasi dilakukan berdasarkan keberhasilan aliran data, waktu pemrosesan, *throughput* pipeline ETL, kualitas data serta *success rate* penyimpanan data ke basis data MySQL.

3. Hasil dan Pembahasan

Hasil dari penerapan pipeline ETL otomatis dalam pengambilan data berbasis container menunjukkan Apache NiFi mampu memproses data komentar YouTube dari awal pengambilan hingga akhir penyimpanan dan berjalan dalam rangkaian proses yang saling terhubung.



Gambar 3. Konfigurasi *docker-compose*

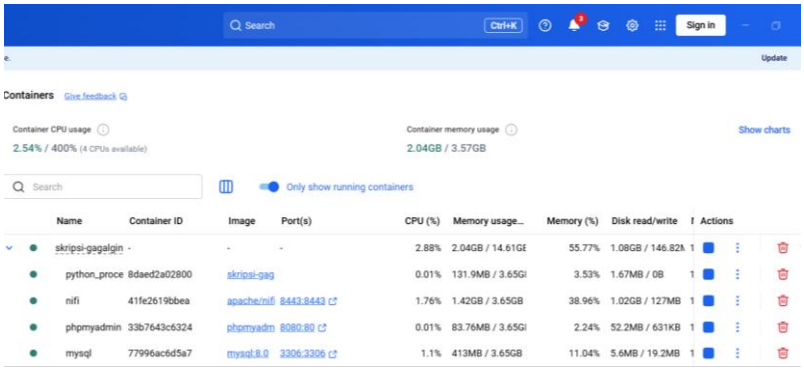
Konfigurasi *docker-compose* pada Gambar 3 mengintegrasikan Apache NiFi, MySQL, phpMyAdmin, dan Python_prosesor dalam satu lingkungan container. Seluruh layanan didefinisikan dalam satu berkas konfigurasi sehingga proses pembangunan pipeline ETL dapat dilakukan secara terstruktur dan terorkestrasi.

Keterhubungan antar layanan dimulai melalui konfigurasi jaringan *etl_network* dengan tipe *bridge network*, sehingga setiap container berada dalam satu jaringan yang sama dan dapat saling berkomunikasi menggunakan nama layanan yang telah didefinisikan. Selain itu, penggunaan *shared volume* memungkinkan pertukaran data antar-container sekaligus menjaga persistensi data selama proses ETL berlangsung.

Untuk mengatur ketergantungan antar layanan, digunakan parameter *depends_on* yang memastikan bahwa layanan yang membutuhkan basis data dijalankan setelah container MySQL aktif. Hal ini mengurangi risiko kegagalan koneksi pada tahap inisialisasi.

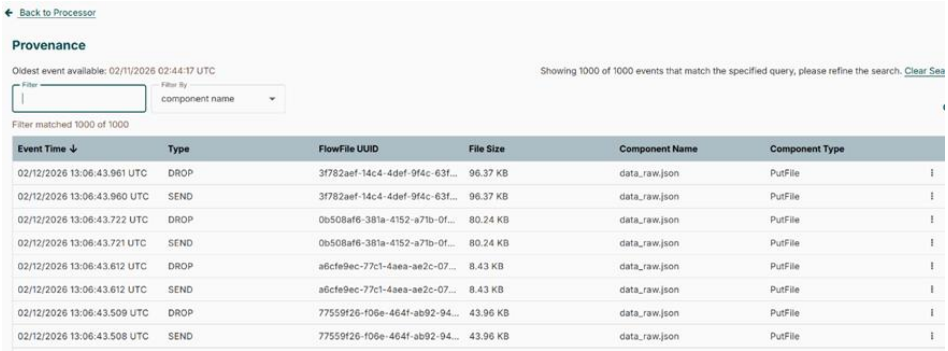
This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2026 Rafita Kusuma, Tikaridha Hardiani

Seluruh container berhasil di-*build* dan dijalankan (*running*) tanpa adanya konflik dependensi maupun kesalahan konfigurasi, sehingga lingkungan ETL berbasis container dapat beroperasi sesuai dengan rancangan yang diharapkan.



Gambar 4. Tampilan Docker

Pada Gambar 4 container dijalankan pada lingkungan Docker setelah proses konfigurasi dan *build* selesai dilakukan. Status *running* pada setiap container layanan Apache NiFi, MySQL, dan Python_prosesor dapat berjalan secara bersamaan tanpa mengalami kegagalan inisialisasi maupun konflik dependensi. Kondisi tersebut mengindikasikan bahwa arsitektur pipeline ETL yang dibangun sudah stabil pada tahap *deployment* awal, sehingga dapat mendukung integrasi antar komponen secara konsisten. Selanjutnya, dilakukan akses ke antarmuka Apache NiFi melalui *browser* untuk melakukan konfigurasi pipeline ETL, meliputi pengaturan *flow*, koneksi antar prosesor, serta integrasi dengan layanan database dan Python_prosesor dalam lingkungan container.



Gambar 5. Data *provenance* pada Apache NiFi

Implementasi Apache NiFi pada Gambar 5, Apache NiFi pada proses *Extract* dalam pipeline ETL pada tahap awal dimulai dengan penggunaan prosesor *InvokeHTTP* yang dikonfigurasi dengan API YouTube serta parameter kata kunci untuk menentukan data video yang akan diambil. Konfigurasi ini memungkinkan Apache NiFi melakukan permintaan ke API YouTube berdasarkan topik yang telah ditentukan, yaitu isu gagal ginjal.

Hasil respons dari API yang berupa data video kemudian diproses menggunakan prosesor *SplitJSON* untuk memisahkan setiap item video ke dalam struktur data yang lebih kecil. Data dari tahap ini disimpan dalam bentuk file *JSON* yang berisi daftar video yang diperoleh berdasarkan kata kunci tersebut.

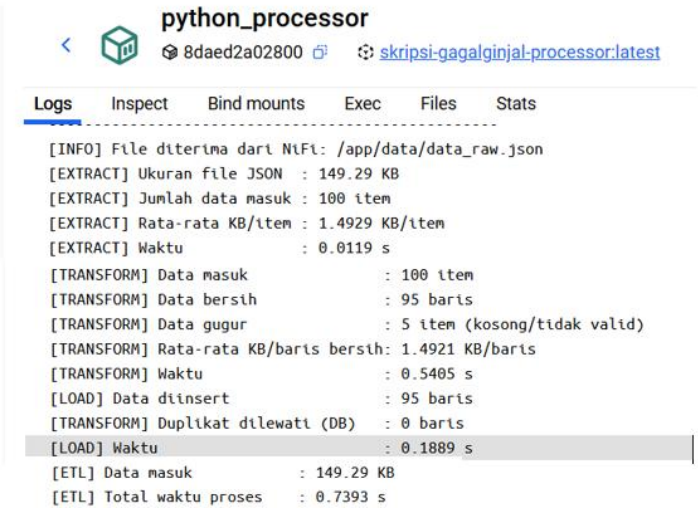
Selanjutnya, untuk setiap video yang telah diperoleh, dilakukan pengambilan data komentar YouTube berdasarkan video_id. Data komentar ini kembali diambil melalui mekanisme *request* API YouTube, kemudian diproses dan disimpan kembali dalam format *JSON*. Seluruh hasil komentar dari

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2026 Rafita Kusuma, Tikaridha Hardiani

proses ini kemudian digabungkan menjadi satu berkas utama yaitu *data_raw.json*, yang digunakan sebagai data mentah untuk tahap berikutnya dalam pipeline ETL.

Untuk memastikan bahwa seluruh data telah berhasil terkirim dan diproses, dilakukan pengecekan melalui data *provenance* Apache NiFi, yang menunjukkan bahwa file *data_raw.json* telah berhasil melewati aliran proses yang telah ditentukan. Setiap file *JSON* yang dihasilkan dibatasi maksimal 100 komentar akibat pembatasan (rate limit) dari API YouTube.

Keberhasilan tahap *Extract* karena kemampuan Apache NiFi dalam mengelola aliran data secara otomatis melalui rangkaian prosesor yang saling terhubung. Apache NiFi mampu melakukan permintaan ke API YouTube, menerima respons dalam format *JSON*, serta mengalirkannya secara terstruktur hingga menjadi dataset awal yang siap diproses lebih lanjut pada tahap *Transform* dan *Load* ke dalam basis data di container berikutnya.



Gambar 6. Tampilan *Log*

Hasil pemantauan yang masuk ke container selanjutnya dari *data_raw.json*, pipeline ETL berhasil menyimpan 30.000 komentar YouTube ke dalam MySQL diperoleh 300 *batch* file *JSON.* Hasil monitoring *log* container pada Gambar 6, setiap file *JSON* memiliki waktu pemrosesan rata-rata sebesar 0,7393 detik per *batch* mulai dari proses *Extract*, *Transform*, hingga *Load* ke basis data. Dengan waktu pemrosesan tersebut diperoleh *throughput* rata-rata sebesar 135 komentar per detik atau setara dengan 7.800 komentar per menit. Selain itu, seluruh 30.000 komentar berhasil dimuat ke basis data MySQL tanpa kegagalan, sehingga diperoleh success rate penyimpanan data sebesar [ISI: X%].

Nilai *throughput* tersebut dipengaruhi oleh karakteristik sumber data yang berasal dari API YouTube. Setiap *request* hanya dapat menghasilkan maksimal 100 komentar sehingga performa pipeline ETL tidak hanya ditentukan oleh kemampuan Apache NiFi, tetapi juga oleh batasan pengambilan data dari API.

Tabel 3. Perbandingan *Throughput* dengan Penelitian Sebelumnya

Penelitian	**Teknologi**	**Jenis Data**	***Throughput***
Perkasa & Hardiani (2026)	Apache Kafka + Spark	*Tweet* hipertensi	19,23 *tweet* per detik
Penelitian ini	Apache NiFi + Docker	Komentar YouTube gagal ginjal	135 komentar per detik

Pada Tabel 3 membandingkan *throughput* antara penelitian ini dan penelitian (Perkasa & Hardiani, 2026). Meskipun menggunakan teknologi dan jenis data yang berbeda, kedua penelitian sama-sama menerapkan pipeline ETL untuk pengolahan data media sosial secara otomatis. Hasil pengujian menunjukkan *throughput* penelitian ini mencapai 135 komentar per detik, lebih tinggi

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2026 Rafita Kusuma, Tikaridha Hardiani

dibandingkan 19,23 *tweet* per detik pada penelitian sebelumnya atau sekitar 7 kali lebih besar. Perbedaan tersebut dipengaruhi oleh karakteristik data, kompleksitas pemrosesan, dan arsitektur pipeline yang digunakan. Namun demikian, kedua penelitian menunjukkan bahwa teknologi pipeline data mampu mendukung pengolahan data media sosial secara berkelanjutan. Perlu digarisbawahi bahwa perbandingan ini melibatkan platform, jenis data, dan sumber data yang berbeda (Apache Kafka/Spark pada data tweet vs. Apache NiFi/Docker pada data komentar YouTube), sehingga angka throughput 135 komentar per detik pada penelitian ini juga dipengaruhi oleh batasan rate limit API YouTube itu sendiri dan tidak semata-mata mencerminkan keunggulan arsitektur; perbandingan ini bersifat indikatif, bukan perbandingan apple-to-apple yang terkontrol.

Tabel 4. Pengujian *latency* terhadap Ukuran Data Komentar

Jumlah Komentar	Ukuran Data (kb)	*latency* (detik)
100	116,26	0,61
200	205,61	1,28
300	309,54	1,38
400	322,89	1,50

Pada Tabel 4 *latency* proses pipeline ETL meningkat seiring bertambahnya ukuran data yang diproses. Pada pengujian dengan ukuran data 116,26 KB, *latency* yang dihasilkan sebesar 0,61 detik. Ketika ukuran data meningkat menjadi 322,89 KB, *latency* juga meningkat menjadi 1,50 detik.

Peningkatan *latency* ini terjadi karena pipeline ETL harus memproses jumlah data yang lebih banyak pada setiap proses, mulai dari proses *Extract* data, *Transform*, hingga *Load* ke tujuan akhir. Semakin besar ukuran data yang diproses, semakin banyak pula sumber daya dan waktu yang dibutuhkan untuk menyelesaikan seluruh aliran data dalam pipeline.

Peningkatan ukuran data diikuti oleh peningkatan *latency* dengan pola yang cenderung bertahap pada rentang data yang diuji. Hasil ini menunjukan volume data memengaruhi waktu pemrosesan pipeline ETL dan menunjukkan hubungan antara ukuran data dan *latency* pada implementasi Apache NiFi yang digunakan dalam penelitian ini. Pengujian ini hanya menggunakan empat titik ukuran data (100–400 komentar), sehingga pola hubungan tersebut masih bersifat indikatif; pengujian pada rentang data yang lebih luas serta analisis statistik (misalnya regresi) diperlukan untuk mengonfirmasi pola ini secara lebih kuat.

Tabel 5. Perbandingan Kualitas Dataset Sebelum dan Sesudah ETL

Metrik	Tanpa Pipeline ETL	Dengan Pipeline ETL
Total Komentar	30.000	30.000
Data kosong	8304	0
Data duplikat	9998	0
Relevansi topik gagal ginjal	64.96%	89.81%

Evaluasi kualitas data pada Tabel 5 dilakukan dengan membandingkan dataset mentah dan dataset hasil pipeline ETL yang berasal dari jumlah data yang sama, yaitu 30.000 komentar YouTube terkait gagal ginjal. Pada dataset mentah tanpa proses pipeline ETL ditemukan 8.304 data kosong dan 9.998 data duplikat. Setelah melalui proses *transform* pada pipeline ETL, seluruh data kosong dan data duplikat berhasil dihilangkan sehingga kualitas data menjadi lebih baik.

Selain itu, relevansi komentar terhadap topik gagal ginjal meningkat dari 64,96% menjadi 89,81%. Peningkatan sebesar 24,85% menunjukkan bahwa proses ETL tidak hanya berfungsi untuk mengotomatisasi perpindahan data, tetapi juga meningkatkan kualitas dan kesesuaian dataset terhadap

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2026 Rafita Kusuma, Tikaridha Hardiani

topik penelitian. Hasil ini karena mekanisme penyaringan data yang diterapkan mampu menghasilkan dataset yang relevan untuk analisis isu kesehatan terkait gagal ginjal di Indonesia. Relevansi topik dihitung berdasarkan kemunculan kata kunci terkait gagal ginjal seperti "gagal ginjal", "ginjal", "cuci darah", "hemodialisis", dan istilah terkait lainnya pada data hasil ETL. Perlu dicatat bahwa pengukuran relevansi ini bersifat otomatis berdasarkan kecocokan kata kunci dan belum divalidasi melalui anotasi manual atau uji reliabilitas independen, sehingga besaran peningkatan relevansi perlu dimaknai sebagai indikasi awal dan bukan ukuran yang telah tervalidasi sepenuhnya.

e: etl_db » Table: etl_results

SQL | Search | Insert | Export | Import | Operations | Triggers

id	id_str	author_id	created_at	komentar	processed_at
29861	UgwMSPHXiW1TF7Hulml4AaABAg	UCP6GEa3CTvf6Qbu4U7VbNJg	2023-09-07 14:17:31	batu ginjal tidak minuman berwarna	2026-02-13 02:15:45
29862	UgwUzH0v9klIh7ksT1d4AaABAg	UCYwtlsHcPHUssjAqYKl5W5Q	2023-08-21 04:48:21	kalo nutrisari bagaimana dokter aman gaksoalnya sa...	2026-02-13 02:15:45
29863	UgyElopD88vv_o8yk914AaABAg	UCDimNAwjD-aJGOjQKtGeDcQ	2023-07-07 11:40:15	semua kalo diminum sesuai takarannya enggak masala...	2026-02-13 02:15:45
29864	UgzWRpvJ3KT8rHKOoTF4AaABAg	UC8blApPokik_2cbfums_heg	2023-06-13 19:30:16	ya sudah pasti kena gagal ginjal karena pola hidup...	2026-02-13 02:15:45
29865	UgxlwAKlS0bUjLiBOw54AaABAg	UCbqwmsLxfYCkDM79KkkcQvA	2023-03-25 19:55:20	dulu aku sering minum florida yang ada butiran jer...	2026-02-13 02:15:45
29866	UgxncHnKRsdqqQuQzL54AaABAg	UC8XoY1eCk5Vd1DzzYdWen-Q	2023-03-24 16:42:33	gagal ginjal lebih baik dari pada gagal cinta	2026-02-13 02:15:45
29867	UgwORovP8vU5LSgF3kh4AaABAg	UChrSsNG3El1Hug5R9v_jq1g	2022-09-26 09:50:11	coba dibaca sampai habis karena ada sepotong infor...	2026-02-13 02:15:45
29868	Ugx6zNEmEuVDeXR3i4t4AaABAg	UC5AFdqYkVBJu9XNTqAenajg	2022-09-16 13:36:35	cuaca yang pnas rasanya pngin mnum yang dingin	2026-02-13 02:15:45
29869	Ugw9Ep-p1XkgfBkhpa14AaABAg	UCvdNjY4KuLbeDpXJVwERA8A	2022-09-10 05:58:04	sebagai orang yang sering minumminuman manis kemas...	2026-02-13 02:15:45
29870	Ugxry8Hm_27iE1qZ4Bx4AaABAg	UCC9RbJzXredxWsYfRmyRqnQ	2022-06-04 09:50:08	air mineral aqua boleh mineral dll juga termasuk a...	2026-02-13 02:15:45
29871	UgwsZGJ73ZvDkGpndWF4AaABAg	UCMUPw9gcghbdPaAjTCakCKQ	2022-06-04 06:50:01	paman saya sopir meninggal gagal ginjal karena ser...	2026-02-13 02:15:45
29872	UgxKzJB0dwWRiFupocZ4AaABAg	UC7u79f_yeaVgAbVxYGT7tcw	2022-06-04 02:18:07	betul sekali	2026-02-13 02:15:45
29873	Ugx8FM45vehxCHBq0iF4AaABAg	UCAybr_EaR2CffPwnnV2OlRg	2022-06-03 14:39:06	betol sekali suhu kalau enggak salh tubuh kita ber...	2026-02-13 02:15:45
29874	UgxtHLWr-ep27eBkPvN4AaABAg	UCB7qtHLg6TfyYK2Te4Dm2oQ	2022-06-03 14:28:58	teman saya seorang dokter dari waktu kuliah sampai...	2026-02-13 02:15:45
29875	Ugz8_02UtPr8ZO1bHYF4AaABAg	UCuH8HsVkeJW3SFd_T3Fimmg	2022-06-03 12:40:19	terima kasih informasinya	2026-02-13 02:15:45
29876	UgyubOFo8Kfecvk2MuZ4AaABAg	UCp6kKvYciyt1kVRobEPx-pg	2022-06-03 12:11:17	cara mengetahui ada gejala gangguan ginjal bagaima...	2026-02-13 02:15:45
29877	Ugxuxx4-DowbQGv12WB4AaABAg	UCWXM8w9FF0JCh58nY1JEpdw	2022-06-03 12:09:10	astaghfirullah thanks dokter	2026-02-13 02:15:45
29915	UgzQGlTuruh_SIlHCTp4AaABAg	UCNoTJm9o-TXQqNF9lDHDlqw	2025-06-09 05:32:44	kok aneh sih kalo dari dokter sppd malah lebih bag...	2026-02-13 02:25:16
29916	Ugwlpl3p8wE0BEke4VZ4AaABAg	UCQZKvfNC7Lomn3QqciPr40A	2025-01-27 20:20:14	alhamdulillah terima kasih telah berbagi ilmu	2026-02-13 02:25:16
29917	UgzvPosTml9r_qE6pJR4AaABAg	UCfymX3fGnp7czgdM-plCRGQ	2025-01-17 15:38:11	semua penyakit ada penawar nya dan saya yakin akhi...	2026-02-13 02:25:16
29918	UgzYRU-NNz20P6a7CdB4AaABAg	UC6jhUJbXHi_UxxWBzSsbHLg	2024-07-02 20:55:21	ini bagi yang sudah cuci darah bagi yang belum bat...	2026-02-13 02:25:16
29919	UgwA3lfKXjlwpK7630Z4AaABAg	UCqPrTWOc375qoHUpcvEhgjA	2024-06-26 10:57:23	ini untuk stadium stadium	2026-02-13 02:25:16

Gambar 7. Kualitas Data Hasil Pipeline ETL

Gambar 7 adalah data hasil pipeline ETL yang telah tersimpan pada MySQL. Hasil ini mendukung evaluasi kualitas data pada Tabel 5 bahwa seluruh data yang tersimpan tidak mengandung data kosong maupun duplikasi.

Selain meningkatkan kualitas data, hasil observasi terhadap data menunjukkan adanya kata-kata yang berkaitan dengan isu gagal ginjal, data yang dikumpulkan memiliki keterkaitan dengan topik kesehatan yang menjadi fokus penelitian.

Namun, pada kalimatnya pada kolom komentar masih terdapat kata yang kurang baku terkait normalisasi. Hasil normalisasi masih terlihat bahwa pipeline ETL belum sepenuhnya mampu menangani bahasa tidak formal atau bahasa gaul. Beberapa komentar masih mengandung singkatan atau kata-kata tidak baku, sehingga hasil normalisasinya belum benar-benar bersih secara bahasa.

4. Kesimpulan dan Saran

Penelitian ini menunjukkan bahwa implementasi Apache NiFi dan teknologi container dapat dimanfaatkan untuk mengelola data komentar YouTube terkait gagal ginjal secara otomatis dan terintegrasi sehingga menghasilkan data yang lebih siap digunakan untuk analisis. Hasil penelitian ini memperkuat potensi pemanfaatan data media sosial sebagai sumber informasi tambahan untuk memahami perhatian, pengalaman, dan persepsi masyarakat terhadap isu gagal ginjal, yang selama ini umumnya masih bersifat tidak terstruktur dan memiliki kualitas yang beragam. Penelitian ini memberikan kerangka kerja pipeline ETL berbasis container untuk pengelolaan data kesehatan dari media sosial sekaligus menunjukkan peran proses ETL dalam meningkatkan kualitas data sebelum digunakan pada tahap analisis lanjutan. Dengan demikian, penelitian ini dapat menjadi landasan bagi pengembangan pipeline ETL pemantauan dan analisis isu kesehatan berbasis media sosial di Indonesia. Namun, penelitian ini masih memiliki keterbatasan dalam menangani bahasa tidak formal serta hanya menggunakan sumber data dari YouTube, sehingga penelitian selanjutnya dapat mengembangkan metode normalisasi yang lebih kontekstual dan memanfaatkan berbagai platform

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2026 Rafita Kusuma, Tikaridha Hardiani

media sosial lainnya. Keterbatasan lain penelitian ini meliputi: pengujian kinerja pipeline yang belum mencakup skenario data dalam skala yang jauh lebih besar sehingga skalabilitas jangka panjang belum sepenuhnya teruji; pengukuran relevansi topik yang masih bersifat otomatis berbasis kata kunci tanpa validasi manual independen; serta belum dievaluasinya penggunaan sumber daya (CPU dan memori) secara terukur pada lingkungan container yang dibangun. Penelitian ini juga menggunakan data komentar publik YouTube tanpa proses anonimisasi eksplisit maupun kajian etika terkait privasi pengguna, sehingga aspek ini perlu menjadi perhatian pada penelitian selanjutnya, khususnya karena topik yang dibahas berkaitan dengan isu kesehatan yang bersifat sensitif.

Referensi

Afrizal, H., & Prihanto, A. (2022). Analisis Kebutuhan Resource Dan Independensi Antara Teknologi Single Server, Virtualisasi Dan Container. *Journal of Informatics and Computer Science*, *4(1)*. https://doi.org/10.26740/jinacs.v4n01.p26-33

Bono, C. A., Cappiello, C., Pernici, B., Ramalli, E., & Vitali, M. (2023). Pipeline Design for Data Preparation for Social Media Analysis. *Journal of Data and Information Quality*, *15*(4). https://doi.org/10.1145/3597305

Cahyani, P. I., & Wijayanto, D. (2025). Implementation of a Web-Based BPJS Bridging API Using the Waterfall Method. *Jurnal Sistem Informasi Dan Bisnis Cerdas*, 18(1). https://doi.org/10.33005/sibc.v18i1.426

Ghani, R. A., & Kurniawan, R. (2024). Implementasi Extract, Transform, Load Process pada Perancangan Data Warehouse Terkait Kualitas Pendidikan di Kabupaten Serang. *Jurnal Mahasiswa Teknik Informatika*, 8(2). https://doi.org/10.36040/jati.v8i2.9081

Hutabalian, P. (2025). Implementasi Pipeline ETL/ELT dan Model Dimensional untuk Analisis Penjualan Shopee Menggunakan PostgreSQL, Docker, dan Apache Superset. *Jurnal Informatika Dan Teknik Elektro Terapan*, *13*(3S1). https://doi.org/10.23960/jitet.v13i3S1.8093

Khoirunnisa, S., Winangsih, R., & Sagita, D. (2025). Analisis Big Data Konsumsi Informasi Kesehatan Masyarakat Banten Pasca Turunnya Kasus Covid-19 Pada Media Sosial YouTube. *Jurnal Riset Komunikasi*, *16*(1). https://doi.org/10.62870/jrk.v1i1.32945

Liu, P., & Guitart, J. (2022). Performance characterization of containerization for HPC workloads on InfiniBand clusters: an empirical study. *Cluster Computing*, *25*(2), 847–868. https://doi.org/10.1007/s10586-021-03460-8

Marella, V. (2024). Comparative Analysis of Container Orchestration Platforms: Kubernetes vs. Docker Swarm. *International Journal of Scientific Research in Science and Technology*, *11*(5), 526–543. https://doi.org/10.32628/ijsrst24105254

Mishra, A. K., Pilli, E. S., & Govil, M. C. (2022). CONTAIN4n6: a systematic evaluation of container artifacts. *Journal of Cloud Computing*, *11*(1). https://doi.org/10.1186/s13677-022-00303-8

Mnsen, H. K. I., Purnomosidi, B., Kartadie, R., & Kurnaedi, R. (2024). Data Pipeline Architecture for Academic Information System at Akademi Teknik Biak. *Journal of Intelligent Software Systems*, *3*(1), 1. https://doi.org/10.26798/jiss.v3i1.1335

Mondal, S. K., Zheng, Z., & Cheng, Y. (2024). On the Optimization of Kubernetes toward the Enhancement of Cloud Computing. *Mathematics*, *12*(16). https://doi.org/10.3390/math12162476

Nasution, M. Z., Sikumbang, E. S., & Gurning, F. P. (2025). Analisis Tren Penyakit Gagal Ginjal Kronik Peserta BPJS dan Dampaknya pada Pembiayaan Kesehatan Indonesia. *Jurnal Kolaboratif Sains*, *8*(7), 4308–4317. https://doi.org/10.22146/jkki.107401

Nur, Z., Sulaiman, U., & Rahman, U. (2024). Metodologi Penelitian: Analisis Konseptual untuk Memahami Hakikat, Tujuan, Prosedur, dan Klasifikasi Penelitian. *PEDAGOGIC: Indonesian Journal of Science Education and Technology*, *4*(1), 34–45. https://doi.org/10.54373/ijset.v4i1.1395

Perkasa, L., & Hardiani, T. (2026). Penerapan Arsitektur Kappa dengan Kafka dan Spark untuk Pemrosesan Data Hipertensi di Media Sosial X. *Jurnal Teknologi Sistem Informasi Dan Aplikasi*, *9*(1), 13–22. https://doi.org/10.32493/jtsi.v9i1.57819

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2026 Rafita Kusuma, Tikaridha Hardiani

Sandiya, I. W., Sudiarsa, I. W., Putra, I. G. A. N. P. M., Arta, I. W. W., & Prayogi, K. D. (2026). *Implementasi Pipeline Etl Dan Algoritma K-Means Untuk Segmentasi Geografis Penjualan E-Commerce*. https://doi.org/10.65094/jcs.v1i1.115

Saputra, A. K., Laksitowening, K. A., & Herdiani, A. (2025). Performance Analysis of Extract, Transform, and Load (ETL) Methods for Business Intelligence in E-Learning Systems Using Pentaho Data Integration. *Jurnal Teknik Informatika (Jutif)*, *6*(1), 75–86. https://doi.org/10.52436/1.jutif.2025.6.1.4173

Setyaji, D. R., & Wijayanto, D. (2025). Perancangan Arsitektur Sistem Antrian Online Terintegrasi BPJS di RSU Queen Latifa Berbasis Docker. *Jurnal Teknik Mesin, Industri, Elektro Dan Informatika*, *4*(2), 32–40. https://doi.org/10.55606/jtmei.v4i2.4890

Sholihah, I., & Darujati, C. (2022). Sistem Replikasi Basis Data berdasarkan MySQL menggunakan Container Docker. *Majalah Ilmiah Teknologi Elektro*, *21*(2), 209. https://doi.org/10.24843/MITE.2022.v21i02.P08

Singu, S. K. (2022). ETL Process Automation: Tools and Techniques. *ESP Journal of Engineering & Technology Advancements*, *2*(1 February 2022), 74–85. https://doi.org/10.56472/25832646/ESP-V2I1P110

Wakchaure, S. L. (2023). Overview of Apache NiFi. *International Journal of Advanced Research in Science, Communication and Technology*, 523–530. https://doi.org/10.48175/ijarsct-14362

Walha, A., Ghozzi, F., & Gargouri, F. (2024). Extract-Transform-Load Process for Recognizing Sentiment from User-Generated Text on Social Media. *International Conference on Evaluation of Novel Approaches to Software Engineering, ENASE - Proceedings*, 641–648. https://doi.org/10.5220/0012706100003687

Wijayanto, D., & Firdonsyah, A. (2023). Implementation of Continous Delivery using Jenkins And Kubernetes with Docker Local Images. *Sinkron*, *8*(4), 2226–2235. https://doi.org/10.33395/sinkron.v8i4.12624

Wnęk, K., & Boryło, P. (2023). A Data Processing and Distribution System Based on Apache Nifi. *Photonics*, *10*(2). https://doi.org/10.3390/photonics10020210

Yepuri, V. K., Polamarasetty, V. K., Donthi, S., & Gondi, A. K. R. (2023). *Containerization of a polyglot microservice application using Docker and Kubernetes*. https://doi.org/10.48550/arXiv.2305.00600

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License
Copyright © 2026 Rafita Kusuma, Tikaridha Hardiani