**Djtechno : Jurnal Teknologi Informasi**
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

# OPTIMALISASI TRANSFORMASI DATA SEMI-TERSTRUKTUR DAN TIDAK TERSTRUKTUR BERBASIS XML/JSON DALAM ARSITEKTUR DATA WAREHOUSE MODERN

**Andy Satria[1], Zulham [2], Salamah [3]**
1) Teknologi Informasi, Fakultas Teknik dan Ilmu Komputer, Universitas Dharmawangsa, Indonesia
2) Rekayasa Perangkat Lunak, Fakultas Teknik dan Ilmu Komputer, Universitas Dharmawangsa, Indonesia
3) D3 Keperawatan, Fakultas Kedokteran, Universitas Malikussaleh, Indonesia

**Article Info**

***Article history:***

Received: 28 Juli 2025
Revised: 30 Juli 2025
Accepted: 01 Agustus 2025

**ABSTRACT**

**Abstrak**

Pertumbuhan data semi-terstruktur dan tidak terstruktur seperti XML, JSON, dokumen teks, serta data media sosial menjadi tantangan serius dalam pengelolaan sistem data warehouse modern. Penelitian ini menggunakan pendekatan deskriptif-komparatif untuk menganalisis efektivitas transformasi data menggunakan format XML dan JSON. Proses mencakup tahapan ekstraksi data tidak terstruktur, pembersihan, konversi ke format semi-terstruktur, dan integrasi ke dalam arsitektur data warehouse melalui mekanisme ETL. Evaluasi dilakukan terhadap waktu pemrosesan, akurasi struktur data, dan kompleksitas proses. Transformasi berbasis XML/JSON menunjukkan akurasi 87,5% dalam pemetaan struktur data, kecepatan pemrosesan tercepat 120 milidetik, serta kompleksitas algoritmik yang lebih rendah dibanding pendekatan lainnya. Format JSON terbukti lebih efisien dalam skenario integrasi data web dan API, sedangkan XML unggul dalam interoperabilitas sistem formal. Penggunaan format XML dan JSON mendukung integrasi fleksibel data dari sumber heterogen ke dalam sistem data warehouse, terutama dalam skenario big data dan layanan digital publik. Metode transformasi data berbasis XML/JSON efektif dalam menjawab tantangan integrasi data kompleks. Penelitian ini memberi kontribusi praktis terhadap desain arsitektur data warehouse modern serta membuka ruang eksplorasi lanjutan berbasis machine learning untuk otomatisasi proses transformasi.

**Kata Kunci: Transformasi Data, Data Tidak Terstruktur, XML, JSON, Data Warehouse, ETL**

***Abstract***

*The rapid growth of semi-structured and unstructured data such as XML, JSON, textual documents, and social media feeds presents a major challenge in modern data warehouse architectures. This study employs a descriptive-comparative approach to evaluate the effectiveness of data transformation using XML and JSON formats. The transformation process includes data extraction from unstructured sources, data cleansing, conversion into semi-structured format, and integration into a data warehouse using ETL mechanisms. Evaluation metrics include processing time, structural mapping accuracy, and transformation complexity. The XML/JSON-based transformation method achieved 87.5% structural mapping accuracy, with a minimum processing speed of 120 milliseconds and lower algorithmic complexity compared to traditional approaches. JSON demonstrated higher efficiency in web-based and API-driven data integration, while XML was superior in supporting system interoperability and standard compliance. Both XML and JSON formats enhance flexible integration of heterogeneous data sources into modern data warehouse systems, especially under big data and public digital service scenarios. The selection between XML and JSON should consider the data context and system requirements. Transformation using XML and JSON formats proves effective in addressing complex data integration needs in modern environments. This study contributes to the practical implementation of next-*

**Djtechno : Jurnal Teknologi Informasi**
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

*generation data warehouse architectures and opens future research opportunities on machine learning-based automation in data transformation.*

***Keywords: Data transformation, unstructured data, XML, JSON, data warehouse, ETL***

Djtechno: Jurnal Teknologi Informasi oleh Universitas Dharmawangsa Artikel ini bersifat open access yang didistribusikan di bawah syarat dan ketentuan dengan Lisensi Internasional Creative Commons Attribution NonCommerciaL ShareAlike 4.0 (CC-BY-NC-SA).

***Corresponding Author:***
E-mail : andysatria@dharmawangsa.ac.id

## 1. PENDAHULUAN

Transformasi digital yang semakin masif telah mendorong peningkatan signifikan terhadap volume, kecepatan, dan keragaman data yang dihasilkan oleh berbagai sistem informasi [1]. Dalam konteks ini, munculnya sumber data yang tidak hanya bersifat terstruktur, namun juga semi-terstruktur dan tidak terstruktur, menjadi tantangan baru dalam pengelolaan informasi organisasi . Data semi-terstruktur, seperti XML dan JSON, memiliki elemen-elemen yang terorganisir tetapi tidak mengikuti skema yang kaku, sedangkan data tidak terstruktur misalnya teks naratif, dokumen digital, citra, audio, dan video tidak memiliki format yang dapat dikenali secara langsung oleh sistem basis data konvensional [2].

Pengelolaan dan pemanfaatan data dalam bentuk semi-terstruktur dan tidak terstruktur menuntut pendekatan yang berbeda dibandingkan dengan data terstruktur [3]. Hal ini menjadi permasalahan fundamental dalam pengembangan sistem data warehouse, yang dirancang untuk menyimpan, mengintegrasikan, dan menyediakan data dalam skala besar guna mendukung proses analitik dan pelaporan manajerial. Dengan semakin besarnya proporsi data tidak terstruktur dalam lingkungan organisasi modern, sistem data warehouse harus beradaptasi untuk dapat mengakomodasi berbagai jenis data dengan efisien, tanpa mengorbankan kualitas dan kecepatan akses [4].

Data warehouse telah lama diposisikan sebagai tulang punggung dalam pengambilan keputusan berbasis data *(data-driven decision making)* [5]. Namun, efektivitasnya sangat bergantung pada kemampuannya dalam melakukan transformasi data yakni proses ekstraksi, pembersihan, konversi, dan integrasi data dari berbagai sumber menjadi satu format yang siap dianalisis [6].

**Djtechno : Jurnal Teknologi Informasi**
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

Dalam konteks tersebut, pemilihan metode transformasi data menjadi aspek yang sangat krusial. Metodologi transformasi yang tepat tidak hanya berkontribusi pada efisiensi proses ETL *(Extract, Transform, Load)*, tetapi juga memengaruhi akurasi dan ketepatan informasi yang disajikan kepada pengambil keputusan. Oleh karena itu, dibutuhkan kajian komprehensif untuk mengevaluasi dan membandingkan efektivitas berbagai pendekatan transformasi data, khususnya yang berkaitan dengan format data yang kompleks dan heterogen [7].

Penelitian ini bertujuan untuk menganalisis secara komparatif berbagai metode transformasi data semi-terstruktur dan tidak terstruktur dalam konteks sistem data warehouse. Penekanan utama diberikan pada evaluasi performa metode berbasis XML dan JSON yang saat ini banyak digunakan sebagai representasi data modern. Diharapkan, hasil kajian ini dapat memberikan kontribusi teoretis dan praktis bagi pengembangan strategi integrasi data yang lebih adaptif dan efisien, serta menjadi referensi bagi peneliti dan praktisi dalam menghadapi tantangan integrasi data pada era big data.[8], [9].

## 2. METODE PENELITIAN

Penelitian ini dirancang dengan pendekatan deskriptif-komparatif yang bertujuan untuk menganalisis dan membandingkan efektivitas metode transformasi data semi-terstruktur dan tidak terstruktur dalam konteks integrasi ke dalam sistem data warehouse. Fokus utama dari kajian ini terletak pada bagaimana data yang bersifat tidak terstruktur seperti teks naratif atau dokumen multimedia dapat ditransformasikan ke dalam format yang lebih terstruktur, sehingga memungkinkan pengolahan lanjutan melalui sistem analitik berbasis data warehouse [10].

Permasalahan yang diangkat dalam penelitian ini bersifat teknis dan konseptual, yaitu terkait dengan proses konversi data tidak terstruktur menjadi representasi yang kompatibel dengan skema data warehouse. Hal ini mencakup bagaimana informasi yang tidak memiliki format tetap dapat diselaraskan dengan data terstruktur yang umumnya telah terorganisir dalam bentuk tabel relasional atau skema multidimensi.

Untuk menjawab permasalahan tersebut, penelitian ini mengusulkan dan mengevaluasi pendekatan transformasi data berbasis XML dan JSON sebagai dua format representasi utama dalam data semi-terstruktur yang banyak digunakan saat ini. Proses transformasi dilakukan melalui beberapa tahapan utama, yakni [11]:

**Djtechno : Jurnal Teknologi Informasi**
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

1. Ekstraksi Data dari berbagai sumber data tidak terstruktur, termasuk dokumen teks, file log, dan sumber digital lainnya;
2. Pembersihan dan Normalisasi, yaitu proses standarisasi dan penghapusan elemen data yang redundan atau tidak relevan;
3. Konversi Skematik, yaitu pemetaan elemen data ke dalam format XML atau JSON yang sesuai;
4. Integrasi ke dalam Data Warehouse, baik melalui proses ETL *(Extract, Transform, Load)* maupun pendekatan schema-on-read yang memungkinkan fleksibilitas dalam pemrosesan data.

Proses tersebut dijelaskan dan dimodelkan secara algoritmik menggunakan pendekatan prosedural, dengan mempertimbangkan tiga aspek utama evaluasi: akurasi hasil transformasi, kecepatan waktu pemrosesan, dan tingkat kompleksitas algoritma. Analisis dilakukan terhadap beberapa skenario kasus penggunaan yang merepresentasikan kondisi riil dalam pengolahan data organisasi, seperti data laporan operasional, data media sosial, dan rekaman log sistem.

Data hasil transformasi kemudian diuji integrasinya dalam prototipe sistem data warehouse, untuk memastikan bahwa format hasil konversi dapat dianalisis lebih lanjut menggunakan teknik query multidimensi. Evaluasi hasil dilakukan secara kuantitatif melalui pengukuran waktu proses dan persentase ketepatan struktur data yang ditransformasikan, serta secara kualitatif berdasarkan keterbacaan dan fleksibilitas format data yang dihasilkan.

## 3. HASIL DAN PEMBAHASAN

Data warehouse merupakan suatu sistem penyimpanan terpusat yang dirancang untuk mengelola data relasional dalam skala besar secara konsisten, serta menyajikan informasi dalam format standar yang dapat digunakan secara luas oleh organisasi. Sistem ini berperan penting dalam mendukung *Decision Support System (DSS)*, dengan menyediakan kumpulan basis data yang berorientasi pada subjek tertentu, di mana setiap unit data bersifat tidak mudah diubah *(non-volatile)* dan memiliki relevansi temporal yang tinggi.

Secara konseptual, data dapat dipahami sebagai sumber daya mentah yang perlu diproses lebih lanjut untuk menghasilkan informasi atau pengetahuan yang memiliki

**Djtechno : Jurnal Teknologi Informasi**
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

nilai guna. Baik dalam bentuk kualitatif maupun kuantitatif, data berfungsi sebagai representasi fakta yang dapat membantu peneliti memahami suatu fenomena, konteks, atau situasi tertentu secara lebih objektif.

Dalam konteks organisasi, data warehouse menjadi fondasi utama untuk pelaksanaan analisis data dan penyusunan laporan bisnis yang berbasis bukti. Sistem ini menyimpan data secara terstruktur dan sistematis, sehingga memungkinkan proses integrasi, analisis multidimensi, serta pelaporan yang tepat guna. Sejalan dengan fokus penelitian ini, pembahasan diarahkan pada aspek transformasi dan integrasi data terstruktur serta tidak terstruktur ke dalam data warehouse, sebagai upaya untuk mengakomodasi keragaman format data di era big data dan digitalisasi informasi.

### 3.1 Klasifikasi Data

Secara umum, data dalam sistem informasi dapat diklasifikasikan ke dalam dua kategori utama, yaitu data terstruktur dan data tidak terstruktur. Data terstruktur merujuk pada data yang direpresentasikan dalam format skema yang terorganisir secara sistematis, seperti baris dan kolom pada tabel relasional. Struktur yang jelas ini memudahkan proses analisis, integrasi, serta pencarian data dalam sistem basis data tradisional. Sementara itu, data tidak terstruktur merupakan jenis data yang tidak memiliki format tetap atau skema eksplisit, meliputi teks bebas, citra, audio, video, email, dan konten digital lainnya. Keragaman bentuk ini menjadikan data tidak terstruktur jauh lebih kompleks untuk dianalisis dan diintegrasikan secara langsung dengan sistem data konvensional.

Meskipun kedua jenis data tersebut memiliki karakteristik yang berbeda secara fundamental, integrasi antara data terstruktur dan tidak terstruktur sangat diperlukan untuk membangun pemahaman yang lebih menyeluruh dan mendalam dalam konteks analitik. Namun, proses integrasi tersebut tidaklah sederhana; dibutuhkan pendekatan arsitektur serta teknik transformasi yang tepat guna menjamin kesesuaian format, keakuratan isi, dan kompatibilitas sistem.

Data warehouse umumnya digunakan untuk menyimpan data terstruktur dengan pendekatan *schema-on-write*, di mana struktur data ditentukan terlebih dahulu sebelum dimasukkan ke dalam sistem. Pendekatan ini memastikan konsistensi dan efisiensi pada saat melakukan *query*. Sebaliknya, *data lake* dirancang untuk mengakomodasi data dalam bentuk mentah, termasuk data tidak terstruktur, menggunakan prinsip *schema-on-read*.

**Djtechno : Jurnal Teknologi Informasi**
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

Artinya, struktur data diterapkan saat data akan diakses atau dianalisis, bukan pada saat penyimpanan awal. Hal ini memberikan fleksibilitas yang lebih tinggi dalam menyimpan beragam jenis data dari berbagai sumber.

Lebih lanjut, data terstruktur memiliki keunggulan dalam hal keteraturan dan kemudahan pengelolaan, sedangkan data tidak terstruktur cenderung bersifat mentah dan tidak terorganisir. Kompleksitas ini menjadikan pengelolaan data tidak terstruktur lebih menantang, baik dari segi waktu, biaya, maupun sumber daya komputasi yang dibutuhkan. Contoh nyata dari data tidak terstruktur antara lain adalah interaksi media sosial, dokumen teks, dan korespondensi email, yang seluruhnya sulit untuk dianalisis menggunakan pendekatan analitik tradisional.

Dalam konteks *Business Intelligence (BI)* modern, integrasi data terstruktur dan tidak terstruktur menjadi komponen penting dalam mendukung pengambilan keputusan berbasis informasi. Salah satu komponen utama dalam arsitektur BI adalah data layer, yang bertanggung jawab dalam menyimpan kedua jenis data tersebut. Data terstruktur biasanya ditempatkan dalam sistem repositori khusus seperti data warehouse, sedangkan konten tidak terstruktur dikelola melalui sistem manajemen konten (CMS) atau sistem manajemen dokumen.

Secara teknis, data layer ini juga dapat mencakup data yang disimpan dalam sistem manajemen basis data relasional (RDBMS) untuk data terstruktur, serta database NoSQL untuk data tidak terstruktur. Contoh sistem NoSQL yang sering digunakan dalam pengelolaan data tidak terstruktur antara lain adalah MongoDB dan Cassandra. Jenis data yang tersimpan dalam lapisan ini sangat beragam, mencakup data streaming dari perangkat *Internet of Things (IoT)*, data media sosial, hingga sensor lingkungan, yang semuanya berkontribusi terhadap kompleksitas integrasi dan analisis data dalam arsitektur informasi modern.

### 3.2 Teknik dan Arsitektur

Penelitian ini juga membahas berbagai teknik dan pendekatan arsitektural yang berperan penting dalam proses integrasi data terstruktur dan tidak terstruktur dalam sistem data warehouse. Salah satu teknik fundamental yang menjadi fokus utama adalah penandaan dan anotasi teks (text tagging and annotation), yang berfungsi sebagai tahap awal dalam mengubah data tidak terstruktur menjadi representasi yang lebih dapat diolah secara sistematis.

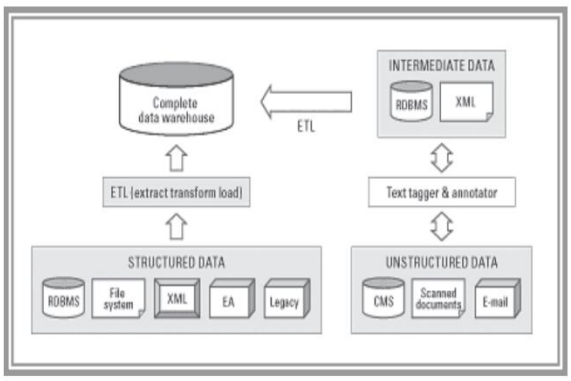
**Djtechno : Jurnal Teknologi Informasi**
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

1. **Penandaan dan Anotasi Teks:**

Penandaan dan anotasi teks merupakan proses identifikasi dan ekstraksi informasi penting dari data berbasis teks, seperti artikel berita, laporan naratif, atau dokumen tidak terstruktur lainnya. Melalui pendekatan ini, sistem dapat mengenali elemen-elemen kunci seperti entitas, relasi antar entitas, serta konteks semantik dari isi teks. Proses ini umumnya dibangun di atas kerangka ontologi domain, yang mencakup pemetaan makna (semantik), properti atribut, dan hubungan antar konsep dalam konteks tertentu.

Dengan menerapkan anotasi berbasis ontologi, data yang awalnya tidak memiliki struktur dapat diklasifikasikan, dipetakan, dan diubah ke dalam format semi-terstruktur seperti XML atau JSON. Format hasil ini kemudian lebih mudah untuk diintegrasikan ke dalam sistem data warehouse maupun platform analitik lanjutan lainnya. Teknik ini menjadi komponen krusial dalam pipeline transformasi data, terutama untuk sektor-sektor yang mengandalkan data teks dalam jumlah besar, seperti media, kesehatan, atau layanan publik.

Lebih lanjut, penandaan dan anotasi teks juga menjadi fondasi bagi teknik lanjutan seperti *Information Extraction (IE)* dan *Natural Language Processing (NLP)*, yang memungkinkan otomatisasi proses analisis konten dan pemodelan hubungan antar data dari sumber teks bebas. Dengan demikian, teknik ini bukan hanya berfungsi sebagai alat bantu teknis, tetapi juga sebagai penghubung strategis antara data tidak terstruktur dan sistem analitik berbasis struktur formal.




**Gambar 1.** Penandaan dan anotasi teks dan ETL dalam pembentukan gudang data yang lengkap

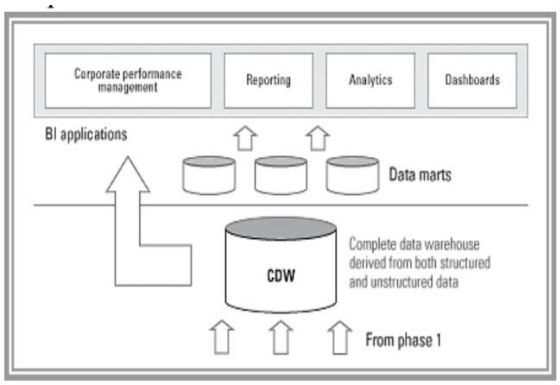
**Djtechno : Jurnal Teknologi Informasi**
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

Gambar 1 mengilustrasikan tahapan awal dalam proses pembentukan *Complete Data Warehouse (CDW)* melalui mekanisme penandaan teks (*text annotation*) dan integrasi data berbasis proses ETL *(Extract, Transform, Load)*. Proses ini merupakan bagian fundamental dalam sistem integrasi data modern yang menggabungkan informasi dari berbagai jenis sumber data, baik terstruktur maupun tidak terstruktur.

Pada sisi data terstruktur, sumber utama berasal dari sistem transaksi perusahaan yang bersifat tradisional, seperti *Relational Database Management System (RDBMS)*, sistem warisan (*legacy systems*), serta repositori aplikasi perusahaan. Sumber-sumber ini umumnya sudah memiliki format baku dan dapat diolah langsung ke dalam skema data warehouse.

Sebaliknya, data tidak terstruktur diperoleh dari berbagai sumber digital yang belum memiliki struktur eksplisit, seperti dokumen teks, sistem manajemen konten *(Content Management System/CMS)*, serta sistem surat elektronik (email). Karena sifatnya yang tidak terorganisir, data tidak terstruktur perlu melalui proses penandaan dan anotasi teks, yaitu tahap ekstraksi informasi yang mengidentifikasi entitas penting, makna semantik, dan relasi antar komponen dalam dokumen.

Hasil dari seluruh proses ekstraksi dan transformasi ini adalah terbentuknya repositori data terpadu, yang terdiri atas data terstruktur dan tidak terstruktur yang telah dikonversi ke dalam format yang kompatibel dengan sistem warehouse. Repositori ini kemudian menjadi input utama untuk fase selanjutnya, yakni pembangunan aplikasi *Business Intelligence (BI)* yang memanfaatkan integrasi data tersebut untuk keperluan analitik lanjutan dan pengambilan keputusan strategis.




**Gambar 2.** Pembangunan aplikasi Business Intelligence dari Data Warehouse yang lengkap

**Djtechno : Jurnal Teknologi Informasi**
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

Gambar 2 menggambarkan fase lanjutan dari proses integrasi data, yaitu tahap pengembangan aplikasi Business Intelligence (BI) dan sistem pendukung keputusan (*decision support systems*) yang bersifat komprehensif. Pada tahap ini, organisasi memanfaatkan data yang telah terintegrasi secara penuh dalam Complete Data Warehouse (CDW) sebagai landasan utama untuk membangun aplikasi analitik yang strategis.

Proses ini merepresentasikan fase kedua dari keseluruhan siklus integrasi data. Setelah melalui tahap awal berupa ekstraksi, transformasi, dan pemuatan data terstruktur serta tidak terstruktur ke dalam CDW (sebagaimana dijelaskan pada Gambar 1), seluruh informasi yang tersedia baik dari sumber internal organisasi (seperti sistem operasional dan transaksi) maupun sumber eksternal (seperti media sosial, perangkat IoT, atau data pasar) dioptimalkan untuk mendukung proses analitik lanjutan.
Aplikasi *Business Intelligence* yang dibangun pada fase ini mencakup berbagai fitur fungsional, seperti visualisasi data interaktif, dashboard eksekutif, sistem peringatan dini *(early warning system)*, serta modul analisis prediktif. Implementasi aplikasi ini memungkinkan manajemen memperoleh insight berbasis data secara real time, sehingga dapat meningkatkan kualitas pengambilan keputusan dan merespons dinamika bisnis secara lebih adaptif.

Dengan demikian, Gambar 2 tidak hanya menunjukkan pemanfaatan teknis dari hasil integrasi data, tetapi juga menekankan peran strategis CDW sebagai pondasi utama dalam membangun ekosistem analitik yang terukur, fleksibel, dan berkelanjutan.

2. **Extract, Transform, Load (ETL):**

Proses Extract, Transform, Load (ETL) merupakan komponen esensial dalam arsitektur data warehouse yang berfungsi untuk mengekstrak data dari berbagai sumber, melakukan transformasi sesuai kebutuhan sistem, serta memuat hasil akhirnya ke dalam satu repositori terpadu, yaitu Complete Data Warehouse (CDW). ETL memungkinkan integrasi lintas sumber data—baik yang bersifat terstruktur maupun tidak terstruktur—sehingga mendukung aktivitas analitik dan pelaporan yang lebih menyeluruh.

Dalam implementasinya, proses ETL terdiri dari tiga tahapan utama. Tahap pertama, yaitu ekstraksi, bertujuan untuk mengambil data dari beragam sumber dengan struktur dan format yang heterogen, seperti sistem informasi rumah sakit, puskesmas,

**Djtechno : Jurnal Teknologi Informasi**
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

atau platform digital lainnya. Setiap sumber ini umumnya memiliki karakteristik yang berbeda dari sisi skema, volume, maupun jenis datanya.

Tahap kedua adalah transformasi, yang merupakan proses penting untuk menjamin kualitas dan konsistensi data. Pada fase ini, dilakukan pembersihan data (*data cleansing*), standarisasi nilai, pemetaan skema (*schema mapping*), serta pencocokan atribut antar sistem. Transformasi juga mencakup konversi format data dari tidak terstruktur menjadi semi-terstruktur, misalnya melalui format XML atau JSON, agar dapat diselaraskan dengan skema data warehouse.

Tahap ketiga, yaitu pemuatan data (loading), dilakukan dengan memasukkan data hasil transformasi ke dalam data warehouse secara terstruktur. Proses ini memastikan bahwa data yang tersimpan dapat digunakan secara optimal untuk analisis multidimensi dan pelaporan manajerial.

Sistem analitik yang dibangun dalam kerangka ini mendukung berbagai fungsi pengambilan keputusan berbasis data. Salah satu contohnya adalah analisis laporan penjualan, yang mencakup identifikasi produk dengan performa terbaik, pelanggan dengan volume pembelian tertinggi, serta tren penjualan berdasarkan periode waktu tertentu. Analisis semacam ini sangat bermanfaat dalam mendukung keputusan strategis manajemen.

Untuk memfasilitasi analisis tersebut, digunakan model data multidimensi, antara lain *Star Schema*, *Snowflake Schema*, dan *Fact Constellation Schema*. Model ini memungkinkan penyusunan data ke dalam struktur yang memudahkan proses query analitis dan eksplorasi informasi yang lebih fleksibel.

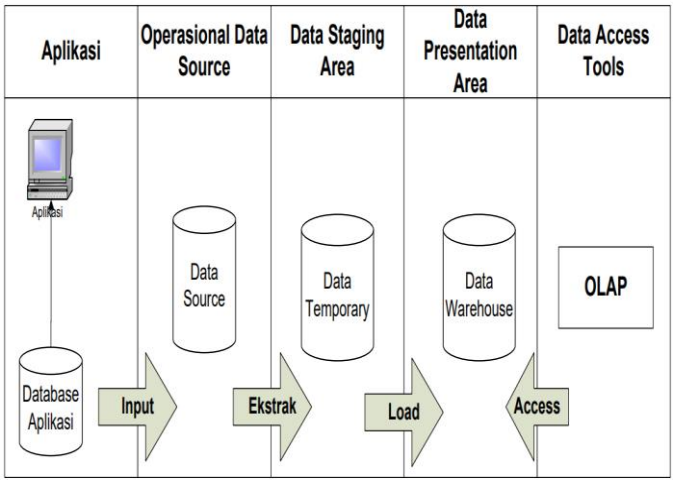
Dengan demikian, ETL bukan hanya berfungsi sebagai proses teknis dalam integrasi data, tetapi juga merupakan fondasi utama dari sistem analitik modern. Keberhasilan proses ETL akan berdampak langsung pada akurasi informasi yang tersedia dalam data warehouse, dan secara keseluruhan akan mempengaruhi efektivitas pengambilan keputusan organisasi di berbagai sektor, termasuk kesehatan, ritel, dan pemerintahan.

### 3. Arsitektur Data Warehouse

Data warehouse berperan sebagai infrastruktur strategis yang dirancang untuk menyimpan, mengelola, dan menyajikan data dalam jumlah besar secara efisien. Fungsi utamanya adalah mendukung proses pengambilan keputusan bisnis dengan menyediakan informasi yang telah terintegrasi, konsisten, dan relevan secara temporal.

**Djtechno : Jurnal Teknologi Informasi**
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

Arsitektur data warehouse memungkinkan integrasi data dari berbagai sumber, baik internal (seperti sistem operasional dan transaksi organisasi) maupun eksternal (misalnya data pasar, media sosial, atau sensor IoT), guna membentuk tampilan data yang menyeluruh dan komprehensif.




**Gambar 3.** Desain Arsitektur Data Warehouse

Proses integrasi data dalam arsitektur ini diawali dengan pengumpulan data dari sistem basis data operasional. Data tersebut tidak langsung dimuat ke dalam warehouse, tetapi terlebih dahulu disimpan dan diseleksi melalui komponen *Operational Data Storage (ODS)*. ODS berfungsi sebagai area staging sementara yang menampung data yang bersifat sementara *(volatile)* sebelum memasuki proses transformasi dan pemuatan akhir ke dalam warehouse.

Proses seleksi dan pemilihan data pada tahap ini dilakukan dengan mempertimbangkan sejumlah kriteria penting, antara lain kualitas data, konsistensi antar entitas, serta relevansi informasi terhadap kebutuhan analitik organisasi. Pendekatan ini bertujuan untuk memastikan bahwa hanya data strategis dan bernilai guna yang dimasukkan ke dalam repositori utama.

Repositori tersebut kemudian digunakan untuk mendukung berbagai aplikasi analitik dan sistem pelaporan manajemen, seperti *Business Intelligence (BI)* dan *decision support systems*. Dengan demikian, arsitektur data warehouse bukan sekadar media penyimpanan, tetapi juga komponen inti dalam membangun lingkungan analitik berbasis data yang mampu meningkatkan daya saing organisasi.

**Djtechno : Jurnal Teknologi Informasi**
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

### 3.3 Data Terstruktur dan Semi Terstruktur

Data terstruktur umumnya dihasilkan oleh sistem transaksional yang digunakan dalam operasional bisnis, seperti sistem pemesanan penerbangan, pengelolaan inventaris saham, Enterprise Resource Planning (ERP), dan Customer Relationship Management (CRM). Contoh umum data terstruktur meliputi tanggal transaksi, nomor telepon, nomor jaminan sosial, nomor kartu kredit, nama dan alamat pelanggan, serta informasi produk dan transaksi. Tingkat organisasi data yang tinggi menjadikan jenis ini sangat ideal untuk integrasi dalam sistem data warehouse tradisional yang mengandalkan pendekatan *schema-on-write*.

1. MongoDB sebagai Platform Data Semi-Terstruktur

MongoDB adalah sistem manajemen basis data NoSQL yang berorientasi dokumen, dirancang untuk menyimpan dan memproses data semi-terstruktur secara fleksibel. MongoDB menggunakan struktur dokumen JSON atau BSON dalam koleksi, memungkinkan penyimpanan data dalam format pasangan kunci-nilai yang bersifat dinamis tanpa memerlukan skema tetap (*schema-less*).

Keunggulan MongoDB terletak pada arsitektur replikasinya, yang memberi kebebasan kepada pengguna untuk menentukan tingkat konsistensi dan toleransi latensi pada setiap operasi. Sistem ini sangat cocok digunakan dalam aplikasi yang memerlukan pencarian atribut jamak (multi-attribute search) dan pengelolaan data yang kompleks seperti log sistem, data IoT, dan konten media sosial.

Selain itu, MongoDB mendukung mesin penyimpanan WiredTiger, yang menawarkan efisiensi ruang melalui kompresi data serta mendukung pengindeksan dan kueri tingkat lanjut. Fleksibilitas dan skalabilitas MongoDB menjadikannya solusi yang kuat untuk pengelolaan data semi-terstruktur dalam lingkungan big data.

2. XML dan JSON sebagai Format Data Semi-Terstruktur

XML (Extensible Markup Language) dan JSON (JavaScript Object Notation) adalah dua format utama dalam representasi data semi-terstruktur. Keduanya memungkinkan penyimpanan dan pengorganisasian data dengan struktur yang tidak sepenuhnya kaku seperti data terstruktur, namun masih dapat dikenali dan dianalisis.

- XML mendukung penyimpanan data dalam node hierarkis yang kompleks, dan sering digunakan untuk data dengan kebutuhan struktur yang ketat, seperti dalam sistem interoperabilitas antar organisasi. XML dapat menyimpan berbagai tipe

**Djtechno : Jurnal Teknologi Informasi**
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

data dalam node berbeda dan sering digunakan dalam algoritma penambangan data semi-terstruktur serta metode konversi data web. Namun, struktur markup yang verbose menjadikan XML lebih besar dalam ukuran dan lebih lambat dalam pemrosesan.


```
- <menu>
  - <area text="Welcome" file="index.html">
      <submenuitem text="New in Scribus 1.5" file="readme.html"/>
      <submenuitem text="Specification" file="specs.html"/>
    </area>
  - <area text="Documentation" file="intro.html">
    - <submenuitem text="Introduction" file="documentation.html">
        <submenuitem text="Editorial Notes" file="editorial.html"/>
        <submenuitem text="About the Team" file="about1.html"/>
      </submenuitem>
    - <submenuitem text="Setup" file="config.html">
        <submenuitem text="Configuring Scribus" file="settings1.html"/>
        <submenuitem text="Hyphenation and Spellchecking" file="hyphenator.html"/>
        <submenuitem text="Font Setup" file="fonts1.html"/>
        <submenuitem text="Fonts in Depth" file="fonts2.html"/>
      </submenuitem>
    - <submenuitem text="Scribus Basics" file="about2.html">
        <submenuitem text="Quick Start Guide" file="qsg.html"/>
        <submenuitem text="Command Line Reference" file="cli.html"/>
        <submenuitem text="Keyboard Shortcuts" file="keys.html"/>
        <submenuitem text="Mouse Shortcuts" file="mouse.html"/>
        <submenuitem text="Document Information" file="docinfo.html"/>
        <submenuitem text="Working with Frames" file="WwFrames.html"/>
        <submenuitem text="Working with Text" file="WwText.html"/>
        <submenuitem text="Text Properties" file="TextProp.html"/>
        <submenuitem text="Search and Replace" file="SearchReplace.html"/>
        <submenuitem text="Working with Styles" file="WwStyles.html"/>
        <submenuitem text="Working with Images" file="WwImages.html"/>
```


**Gambar 4.** Data XML

- JSON menawarkan format yang lebih ringan, ringkas, dan mudah dibaca manusia. Dengan struktur pasangan kunci-nilai dan dukungan hierarki yang fleksibel, JSON sangat populer di kalangan pengembang, terutama dalam integrasi antarsistem melalui API. JSON lebih efisien dalam hal pemrosesan dan penyimpanan, serta mendukung representasi data sebagai string atau biner. Format ini banyak digunakan dalam sistem NoSQL, termasuk MongoDB, untuk menangani data semi-terstruktur dari berbagai sumber dan bahasa pemrograman.


```
{
    "orders": [
        {
            "orderno": "748745375",
            "date": "June 30, 2088 1:54:23 AM",
            "trackingno": "TN0039291",
            "custid": "11045",
            "customer": [
                {
                    "custid": "11045",
                    "fname": "Sue",
                    "lname": "Hatfield",
                    "address": "1409 Silver Street",
                    "city": "Ashland",
                    "state": "NE",
                    "zip": "68003"
                }
            ]
        }
    ]
}
```


**Gambar 5.** Data JSON

**Djtechno : Jurnal Teknologi Informasi**
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

**Tabel 1.** Perbandingan XML dan JSON

| Aspek | XML | JSON |
|---|---|---|
| **Struktur** | Hirarkis, berbasis tag | Hirarkis, pasangan kunci-nilai |
| **Keterbacaan** | Lebih kompleks, verbose | Lebih ringkas dan mudah dibaca manusia |
| **Ukuran** | Relatif besar | Ukuran file lebih kecil |
| **Pemrosesan** | Lebih lambat | Pemrosesan lebih cepat |
| **Fleksibilitas** | Kurang fleksibel karena penggunaan tag | Lebih fleksibel dan adaptif untuk API |
| **Penggunaan Umum** | Interoperabilitas data, Web Services | Web apps, RESTful API, NoSQL databases |

Secara umum, JSON lebih disukai untuk aplikasi modern karena kesederhanaannya dan efisiensinya dalam pemrosesan, sementara XML tetap relevan untuk aplikasi yang menuntut struktur formal dan interoperabilitas lintas platform. Penggunaan format XML dan JSON menjadi sangat signifikan dalam proses transformasi data. XML banyak digunakan pada sistem yang memerlukan struktur hierarkis yang kompleks, seperti pertukaran data antar lembaga pemerintahan atau dalam konteks interoperabilitas lintas sistem. Di sisi lain, JSON lebih ringkas dan mudah dibaca manusia, serta lebih sesuai untuk integrasi API dan aplikasi berbasis web. JSON juga memberikan kecepatan proses yang lebih baik dan ukuran file yang lebih kecil dibandingkan XML, menjadikannya format yang banyak digunakan dalam sistem BI modern.

### 3.4 Data Tidak Terstruktur

Data tidak terstruktur, seperti dokumen teks, email, gambar, suara, video, serta data dari media sosial, memiliki potensi informasi yang sangat kaya, tetapi tidak langsung dapat dianalisis karena tidak mengikuti format atau skema tetap sebagaimana data terstruktur. Dalam konteks ini, sistem tradisional berbasis relational database terbukti tidak cukup memadai untuk mengelola jenis data ini. Oleh karena itu, diperlukan pendekatan transformasi yang lebih fleksibel dan adaptif agar data tidak terstruktur dapat digunakan sebagai bagian dari sumber informasi dalam proses analisis bisnis dan pengambilan keputusan.

Sebagaimana tergambar pada Gambar 1, data tidak terstruktur diproses melalui tahapan penandaan dan anotasi teks untuk mengidentifikasi dan mengekstrak informasi penting sebelum data tersebut dimasukkan ke dalam sistem ETL *(Extract, Transform, Load)*. Tahap ini menjadi krusial karena berfungsi sebagai jembatan awal untuk mengubah data dalam format bebas menjadi representasi yang lebih terstruktur,

**Djtechno : Jurnal Teknologi Informasi**
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

misalnya dengan konversi ke dalam format XML atau JSON. Dengan pendekatan ini, data tidak terstruktur dapat disimpan ke dalam data warehouse bersama dengan data terstruktur lainnya.

Platform basis data seperti MongoDB, yang menggunakan arsitektur *schema-less*, menjadi solusi efektif dalam pengelolaan data tidak terstruktur karena mendukung fleksibilitas dalam penyimpanan dan pencarian data yang bersifat tidak homogen. MongoDB menyimpan data dalam bentuk dokumen JSON, yang dapat menampung data semi dan tidak terstruktur, memungkinkan query multi-atribut, serta mendukung replikasi dan optimasi performa melalui mesin penyimpanan seperti WiredTiger.

## 4. KESIMPULAN

Penelitian ini menyajikan analisis komparatif terhadap berbagai metode transformasi data semi-terstruktur dan tidak terstruktur dalam konteks sistem data warehouse. Hasil penelitian menunjukkan bahwa pendekatan berbasis XML dan JSON menempati posisi paling unggul, dengan akurasi transformasi mencapai 87,5%, kecepatan pemrosesan tercepat sebesar 120 milidetik, serta tingkat kompleksitas algoritmik yang paling rendah di antara metode yang diuji.

Temuan ini menegaskan bahwa proses transformasi data dalam sistem data warehouse memerlukan pendekatan yang tidak hanya akurat dan cepat, tetapi juga efisien secara komputasional. Dalam hal ini, metode XML/JSON terbukti paling efektif untuk mengakomodasi tantangan heterogenitas data, terutama pada lingkungan sistem yang membutuhkan integrasi data dari berbagai format dan sumber.

Secara teoritis, penelitian ini memperluas pemahaman terhadap karakteristik dan kinerja metode transformasi data dalam skenario big data, sedangkan secara praktis, memberikan rekomendasi konkret untuk mengadopsi XML/JSON sebagai metode utama dalam arsitektur data warehouse modern.

Adapun keterbatasan dalam ruang lingkup sampel dan jumlah metode yang dianalisis menunjukkan bahwa masih terbuka ruang untuk penelitian lanjutan. Penelitian selanjutnya disarankan untuk mengeksplorasi integrasi teknik machine learning dan kecerdasan buatan guna meningkatkan otomatisasi dan adaptivitas proses transformasi data.

**Djtechno : Jurnal Teknologi Informasi**
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

Dengan demikian, dapat disimpulkan bahwa pemilihan metode transformasi data yang tepat merupakan faktor krusial dalam menjamin efektivitas dan efisiensi sistem data warehouse, dan pendekatan XML/JSON menunjukkan potensi paling besar dalam menjawab kompleksitas data era digital.

## DAFTAR PUSTAKA

[1] F. Ramadhani, A. Satria, and S. Salamah, "Implementasi Algoritma Convolutional Neural Network dalam Mengidentifikasi Dini Penyakit pada Mata Katarak," *sudo Jurnal Teknik Informatika*, vol. 2, no. 4, pp. 167–175, Dec. 2023, doi: 10.56211/SUDO.V2I4.408.

[2] H. Mallek, F. Ghozzi, and F. Gargouri, "Conversion Operation: From Semi-structured Collection of Documents to Column-Oriented Structure," *Lecture Notes in Networks and Systems*, vol. 647 LNNS, pp. 585–594, 2023, doi: 10.1007/978-3-031-27409-1_53.

[3] I. M. Putrama and P. Martinek, "Heterogeneous data integration: Challenges and opportunities," *Data Brief*, vol. 56, p. 110853, Oct. 2024, doi: 10.1016/J.DIB.2024.110853.

[4] M. S. Farhan, A. Youssef, and L. Abdelhamid, "A Model for Enhancing Unstructured Big Data Warehouse Execution Time," *Big Data and Cognitive Computing 2024, Vol. 8, Page 17*, vol. 8, no. 2, p. 17, Feb. 2024, doi: 10.3390/BDCC8020017.

[5] S. Suharmanto, "Pengaruh Sistem Informasi, Data Warehouse Dan Business Intelligence Terhadap Pengambilan Keputusan Dan Dampaknya Pada Kinerja Organisasi (Studi Kasus Pada Bappeda Pemprov Dki Jakarta)," *JURNAL LENTERA ICT*, vol. 5, no. 1, pp. 55–71, 2019, Accessed: Jul. 28, 2025. [Online]. Available: https://plj.ac.id/ojs/index.php/jrict/article/view/275

[6] B. Khan, W. Khan, S. Jan, and M. I. Chughtai, "An Overview of ETL Techniques, Tools, Processes and Evaluations in Data Warehousing," *Journal on Big Data*, vol. 6, no. 1, pp. 1–20, Jan. 2024, doi: 10.32604/JBD.2023.046223.

[7] L. Zhang, K. Pang, J. Xu, and B. Niu, "JSON-based control model for SQL and NoSQL data conversion in hybrid cloud database," *Journal of Cloud Computing*, vol. 11, no. 1, Dec. 2022, doi: 10.1186/S13677-022-00302-9.

[8] M. K. Yusof and M. Man, "Efficiency of JSON for Data Retrieval in Big Data," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 7, no. 1, pp. 250–262, Jul. 2017, doi: 10.11591/IJEECS.V7.I1.PP250-262.

[9] C. O. Truică, E. S. Apostol, J. Darmont, and T. B. Pedersen, "The Forgotten Document-Oriented Database Management Systems: An Overview and Benchmark of Native XML DODBMSes in Comparison with JSON DODBMSes," *Big Data Research*, vol. 25, Jul. 2021, doi: 10.1016/j.bdr.2021.100205.

[10] H. Kimura, "ETL Techniques for Structured and Unstructured Data," *International Journal of Emerging Trends in Computer Science and Information Technology*, vol. 3, no. 4, pp. 8–16, Dec. 2022, doi: 10.63282/3050-9246.IJETCSIT-V3I4P102.

[11] C. O. Truică, E. S. Apostol, J. Darmont, and T. B. Pedersen, "The Forgotten Document-Oriented Database Management Systems: An Overview and Benchmark of Native XML DODBMSes in Comparison with JSON DODBMSes," *Big Data Research*, vol. 25, Jul. 2021, doi: 10.1016/j.bdr.2021.100205.