DISERTASI

MODEL INTEGRASI DATA DENGAN CLUSTERING DALAM BUSINESS INTELLIGENCE

Nama	**: Antonius Wahyu Sudrajat**
NIM	**: 03013622025008**
BKU	**: Teknik Informatika**
Promotor	**: Prof. Dr. Ermatita, M. Kom**
Ko - Promotor	**: Samsuryadi, S.Si., M.Kom., Ph.D.**

PROGRAM STUDI DOKTOR ILMU TEKNIK
FAKULTAS TEKNIK
UNIVERSITAS SRIWIJAYA
2025

HALAMAN PENGESAHAN

DISERTASI

(TKT7105)

MODEL INTEGRASI DATA DENGAN CLUSTERING DALAM BUSINESS INTELLIGENCE

Oleh:

Antonius Wahyu Sudrajat

NIM: 03013622025008

Telah disetujui

Pada Tanggal, April 2025

Promotor	Ko-Promotor
Prof. Dr. Ermatita, M.Kom.	**Samsuryadi, S.Si., M.Kom., Ph.D.**
NIP. 196709132006042001	**NIP. 197102041997021003**

Mengetahui,

Dekan Fakultas Teknik,	Koordinator Program Studi
Dr. Ir. Bhakti Yudho Suprapto, S.T., M.T., IPM.	**Prof. Dr. Ir. Nukman, M.T.**
NIP. 197502112003121002	**NIP. 195903211987031001**

HALAMAN PERSETUJUAN

Disertasi berjudul "Model Integrasi Data Dengan Clustering Dalam Business Intelligence" telah dipresentasikan dihadapan Tim Penguji Disertasi pada Program Studi Doktor Ilmu Teknik Fakultas Teknik Universitas Sriwijaya pada Hari Rabu, Tanggal 30 April 2025.

Palembang, 30 April 2025
Tim Penguji Disertasi berupa Disertasi:

Ketua Tim Penguji:

Dr. Ir. Bhakti Yudho Suprapto, S.T., M.T., IPM ( )
NIP : 97502112003121002

Anggota Tim Penguji:

1. **Nama :** Dr. Fathoni, S.T., M.MSI. ( )
   NIP : 197210182008121001
2. **Nama :** Dr. Ali Ibrahim, S.Kom., M.T.
   NIP : 198407212019031004
3. **Nama :** Dr. Sanmorino, S.Kom., M.Kom. ( )
   NIDN : 0221118302

Mengetahui,

Dekan Fakultas Teknik,	**Koordinator Program Studi**
Dr. Ir. Bhakti Yudho Suprapto, S.T., M.T., IPM.	**Prof. Dr. Ir. Nukman, M.T.**
NIP. 197502112003121002	**NIP. 195903211987031001**

SURAT PERNYATAAN

Saya yang bertanda tangan di bawah ini:

Nama	: Antonius Wahyu Sudrajat
NIM	: 03013622025008
Program Studi	: Doktor Ilmu Teknik
BKU	: Teknik Informatika

Dengan ini menyatakan bahwa disertasi saya dengan judul "Model Integrasi Data Dengan Clustering Dalam Business Intelligence"**,** bebas dari fabrikasi, falsifikasi, plagiat, kepengarangan yang tidak sah dan konflik kepentingan dan pengajuan jamak, seperti yang tercantum dalam Peraturan Menteri Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia Nomor 39 Tahun 2021.

Bilamana ditemukan ketidak sesuaian dengan hal-hal di atas, maka saya bersedia menerima sanksi akademik dari Universitas Sriwijaya sesuai dengan aturan yang berlaku.

Demikian pernyataan ini dibuat dengan sesungguhnya dan dengan sebenar-benarnya.

Palembang, 6 April 2025

Yang menyatakan

Antonius Wahyu Sudrajat

NIM. 03013622025008

KATA PENGANTAR

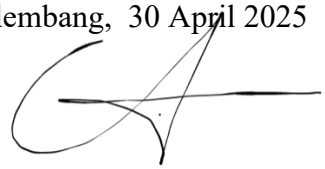
Puji syukur penulis panjatkan ke hadirat Allah SWT. Tuhan Yang Maha Esa, atas segala rahmat dan karunia yang diberikan sehingga Saya dapat menyelesaikan penelitian Disertasi ini, sebagai salah satu syarat untuk memperoleh gelar Doktor pada Fakultas Teknik Universitas Sriwijaya. Saya menyadari bahwa penulis tidak dapat sendiri dalam penyelesaian Disertasi ini, melalui kesempatan ini Saya menyampaikan terima kasih kepada :

1. Keluarga tercinta yang selalui mendoakan dan mendukung proses penelitian Disertasi ini.
2. Ibu Prof. Dr. Ermatita., M.Kom selaku promotor, yang telah meluangkan waktu, memberikan arahan, saran dan masukan dalam penelitian Disertasi ini.
3. Syamsuryadi, S. Si., M. Kom., Ph.D selaku ko-promotor, yang telah meluangkan waktu untuk memberikan arahan terkait data penelitian tanaman sorghum.
4. Dosen-dosen di Program Studi Ilmu Teknik (S3) BKU Teknik Informatika beserta staf yang telah membimbing dan juga membantu dalam proses perkuliahan.
5. Bapak Alexander Kurniawan dan Bapak James Alexander selaku Ketua dan pihak Yayasan Multi Data Palembang, yang telah memberikan dukungan untuk menyelesaikan studi Doktor (S3) ini.
6. Bapak Dr. Johannes Petrus, S.Kom., M.T.I selaku Rektor Universitas Multi Data Palembang, yang memberikan dukungan bagi penulis untuk dapat menyelsaikan studi Doktor (S3) ini.
7. Para mentor Doctoral Reseacrh Universitas MDP yang telah sangat membantu mengarahkan dan memotivasi dalam proses penyelesaian Disertasi saya ini.
8. Rekan-rekan Universitas MDP, Bapak/Ibu Wakil Rektor, Bapak Dekan FIKR dan FIB serta Bapak/Ibu Kaprodi yang telah sangat membantu dalam mendukung penyelesaian pendidikan Doktor (S-3) saya ini.
9. Semua teman-teman yang telah memberikan dukungannya kepada saya selama ini.

10. Keluarga besar saya yang terus memberikan dukungan, do'a dan semangat dalam menyelsaikan studi Doktor (S3) ini.
11. Pihak-pihak lain yang tidak dapat disebutkan satu persatu.

Penulis menyadari bahwa dalam penyusunan Disertasi ini masih jauh dari kesempurnaan, oleh karena itu kritik dan saran yang membangun dalam penyempurnaan penulisan sangat penulis harapkan dan dapat menjadi acuan yang berguna dalam penelitian-penelitian selanjutnya. Dan Penulis juga mohon maaf apabila dalam penulisan Disertasi ini masih banyak kekurangan dan kesalahan.

Palembang, 30 April 2025



Antonius Wahyu Sudrajat
NIM : 03013622025008

RINGKASAN

MODEL INTEGRASI DATA DENGAN CLUSTERING DALAM BUSINESS INTELLIGENCE

Karya Tulis Ilmiah berupa Disertasi, 30 April 2025

Antonius Wahyu Sudrajat; dibimbing oleh Prof. Dr. Ermatita, M. Kom dan Samsuryadi S.Si., M.Kom., Ph.D.

Program Studi Doktor Ilmu Teknik, Fakultas Teknik, Universitas Sriwijaya

Data yang berkualitas sangat penting untuk mendukung pengelolaan dan pengembangan Usaha Mikro, Kecil, dan Menengah (UMKM) yang dilakukan oleh pemerintah. Namun, keterbatasan kemampuan pelaku UMKM dalam menyediakan data yang lengkap sering kali menjadi hambatan, sehingga data yang dikumpulkan mengandung banyak nilai hilang (*missing values*). Kondisi ini menimbulkan tantangan serius dalam proses analisis dan pengambilan keputusan berbasis data. Untuk mengatasi permasalahan nilai hilang tersebut, penelitian ini mengusulkan model baru dalam imputasi data hilang, yaitu *Clustering and Normalization-based GAIN* (CN-GAIN), yang merupakan pengembangan dari metode *Generative Adversarial Imputation Network* (GAIN). Model ini mengintegrasikan dua tahap pra-pemrosesan penting, yaitu klasifikasi data berbasis klaster dan normalisasi-denormalisasi sebelum proses imputasi dilakukan oleh model GAIN. Penelitian ini mensimulasikan tiga jenis pola nilai hilang yang berbeda, yaitu: MAR (*Missing At Random*), MCAR (*Missing Completely At Random*), dan MNAR (*Missing Not At Random*). Setiap pola diuji menggunakan dua model, yaitu CN-GAIN dan GAIN sebagai baseline. Evaluasi dilakukan menggunakan empat metrik utama: *Root Mean Squared Error* (RMSE),

Mean Squared Error (MSE), *Mean Absolute Error* (MAE) dan Akurasi. Hasil eksperimen menunjukkan bahwa model CN-GAIN menghasilkan performa yang lebih baik dibandingkan GAIN di seluruh kategori *missing value*. Beberapa temuan utama antara lain: Untuk kategori MNAR, CN-GAIN menurunkan nilai RMSE sebesar 48,78% dibanding GAIN, menandakan kemampuan adaptifnya dalam menghadapi data hilang yang tidak acak. Pada kategori MAR, model CN-GAIN mencatat penurunan MSE sebesar 99,60% dibandingkan baseline. Untuk metrik MAE, CN-GAIN mencatat penurunan error hingga 70% pada skenario MNAR, menunjukkan efisiensi dalam estimasi nilai hilang dengan akurasi tinggi. Pada kategori MCAR, CN-GAIN mencapai tingkat akurasi sangat tinggi hingga 1.0000 (100%), dibandingkan GAIN yang mencapai 0.9992. CN-GAIN terbukti unggul dalam mengimputasi data hilang pada berbagai pola missing value. Integrasi klasifikasi dan normalisasi sebagai langkah pra-pemrosesan mampu meningkatkan akurasi dan menurunkan tingkat kesalahan secara signifikan. Model ini sangat potensial untuk diterapkan dalam pengelolaan data UMKM yang tidak lengkap dan mendukung sistem pengambilan keputusan berbasis data di sektor publik maupun swasta.

Kata Kunci: Missing values; GAIN method; normalization denormalization; imputation; UMKM data

SUMMARY

DATA INTEGRATION MODEL WITH CLUSTERING IN BUSINESS INTELLIGENCE

Quality data is very important to support the management and development of Micro, Small, and Medium Enterprises (MSMEs) carried out by the government. However, the limited ability of MSME actors to provide complete data is often an obstacle, so the data collected contains a lot of *missing values*. This condition poses serious challenges in the process of analysis and data-driven decision-making. To overcome the problem of lost value, this study proposes a new model in the imputation of lost data, namely Clustering and Normalization-based GAIN (CN-GAIN), which is a development of the Generative Adversarial Imputation Network (GAIN) method. This model integrates two important pre-processing stages, namely cluster-based data classification and normalization-denormalization before the imputation process is carried out by the GAIN model. This study simulates three different types of missing value patterns, namely: MAR (Missing At Random), MCAR (Missing Completely At Random), and MNAR (Missing Not At Random). Each pattern was tested using two models, namely CN-GAIN and GAIN as baseline. The evaluation was conducted using four main metrics: Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Error (MAE) and Accuracy. The results of the experiment showed that the CN-GAIN model performed better than GAIN in all missing value categories. Some of the key findings include: CN-GAIN lowered the RMSE value by 48.78% compared to GAIN, indicating its adaptive ability in dealing with non-random lost data. In the MAR category, the CN-GAIN model recorded a decrease in MSE of 99.60% compared to the baseline. For the MAE metric, CN-GAIN recorded a 70% reduction in errors in the MNAR scenario, showing efficiency in estimating lost values with high accuracy. In the MCAR category, CN-GAIN achieved a very high accuracy level of

up to 1.0000 (100%), compared to a GAIN of 0.9992. For the MNAR category CN-GAIN has proven to be superior in imputing lost data on various missing value patterns. The integration of classification and normalization as a pre-processing step is able to improve accuracy and lower the error rate significantly. This model has great potential to be applied in the management of incomplete MSME data and support data-driven decision-making systems in both the public and private sectors.

Kata Kunci: Missing values; GAIN method; normalization denormalization; imputation; UMKM data

PUBLIKASI

Pubikasi Pendahuluan 1

Nama Konferensi	:	2nd International Conference of Health, Science and Technology (ICOHETECH) 2021
Judul	:	Application of the Apriori Algorithm and FP-Growth to find out the Association Rule between Gender, Education level on wages of SMEs workers in Palembang City
Tahun	:	2021
ISBN	:	978-623-92207-1-6
Penerbit	:	LPPM Universitas Duta Bangsa Surakarta Indonesia

Pubikasi Pendahuluan 2

Nama Konferensi	:	Electrical Engineering, Computer Science and Informatics (EECSI 2023),
Judul	:	Extending The Data Integration Model As The Foundation Of Business Intelligence: A Systematic Literature Review
Tahun	:	2023
DOI	:	10.1109/EECSI59885.2023.10295685
Penerbit	:	IEEE Xplore / Scopus

Pubikasi Utama

Judul Artikel #1	:	CN-GAIN: Classification and Normalization-Denormalization-based Generative Adversarial Imputation Network for Missing SMES Data Imputation
DOI	:	10.14569/IJACSA.2025.0160131
Nama Jurnal	:	**International Journal of Advanced Computer Science and Applications (IJACSA)**
ISSN	:	21565570, 2158107X
Negara	:	United Kingdom (UK)
Terindeks	:	Scopus Q3 (Elsevier)
SJR	:	0,26

Printed Media / Buku Cetak

Judul Buku	:	Pengantar Sistem Informasi dan Knowledge Manajement
Tahun	:	2025
ISBN	:	978-634-229-033-0
ISBN		978-634-229-034-7 (PDF)
Penerbit	:	CV Jejak, anggota IKAPI

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I
PENDAHULUAN

Pendahuluan merupakan bab awal dari laporan disertasi. Bab ini berisikan pendahuluan yang terdiri dari latar belakang, tujuan penelitian, rumusan masalah, batasan masalah dan sistematika penulisan. Permasalahan dan kontribusi penelitian dijelaskan pada latar belakang.

1.1 LATAR BELAKANG

Membuat keputusan strategis dalam lingkungan bisnis yang dinamis merupakan tantangan yang dihadapi oleh banyak organisasi saat ini(Richards et al., 2019). Pengambilan keputusan secara tepat dan cepat berdasarkan fakta menjadi hal penting untuk dapat bertahan. Penggunaan teknologi informasi menjadi hal penting dalam menyediakan informasi bisnis. Dalam lingkungan bisnis yang semakin kompleks saat ini, organisasi memerlukan sistem informasi manajemen yang secara khusus untuk merespon cepat setiap perubahan pasar. Pada pertangahan tahun 90-an, *Business Intelligence* (BI) muncul sebagai tanggapan terhadap perubahan lingkungan kompetitif yang luar biasa, pertumbuhan teknologi yang cepat, peningkatan dukungan IT untuk implementasi proses bisnis dan penyebaran internet di seluruh dunia(Jahantigh et al., 2019). BI berakar pada sistem pendukung keputusan (SPK) dan telah mengalami perkembangan yang signifikan selama decade terakhir.

UKM (Usaha Mikro, Kecil, dan Menengah) Indonesia sangat penting dalam meningkatkan pertumbuhan ekonomi dan pendapatan daerah. Hal ini mendorong pemerintah Indonesia untuk terus mengembangkan UKM melalui beberapa skema, antara lain penyediaan modal usaha, peningkatan kapasitas usaha melalui pelatihan, dan sebagainya. Sebagai dasar pengembangan UKM, pemerintah membutuhkan data karakteristik UKM sebagai dasar pengambilan keputusan.

Business Intelligence adalah teknologi untuk mendukung pekerjaan pemerintah dalam mengelola data UKM.

Business Intelligence (BI) adalah konsep pemanfaatan sejumlah besar data perusahaan yang diproses sedemikian rupa untuk menghasilkan informasi yang bermanfaat(Runtuwene et al., 2018). Unsur penting dalam pengambilan keputusan dalam *Business Intelligence* adalah bagaimana data dikumpulkan dari berbagai sumber dan bentuk ke dalam data warehouse, dimana proses ini selanjutnya disebut integrasi data. Salah satu komponen penting untuk mengembangkan kerangka intelijen bisnis adalah integrasi data(Rodzi et al., 2016). Integrasi data memiliki tujuan untuk menyediakan akses terpadu ke data yang berada di beberapa sumber data otonom(X. L. Dong & Srivastava, 2015). Integrasi data merupakan fondasi dari data warehouse(Sherman, 2014), yang pada gilirannya menjadi fondasi dari BI(Sherman, 2015). Proses integrasi data yang mendukung BI memiliki keunikan masing-masing(Dayal et al., 2009). Beberapa alasan yang melatarbelakangi pentingnya fase integrasi data dalam sistem keputusan adalah: format heterogen, format data sulit untuk ditafsirkan atau ambigu, *database* tidak relevan dan struktur sumber data berubah seiring waktu(Souibgui et al., 2019a). Data warehouse adalah kumpulan *database* terintegrasi, berorientasi subjek yang ditunjuk untuk mendukung proses pengambilan keputusan(El Akkaoui et al., 2011). Laporan bisnis yang cepat dan efisien dari berbagai sumber dapat disedikan di data warehouse. ETL (*Extract Transform Loading*) adalah komponen yang memasok data warehouse dengan semua data yang diperlukan. Pertumbuhan, pemeliharaan, evolusi dan kualitas data yang ada di *Data Warehouse* (DW) sangat bergantung pada semua operasi yang dilakukan di ETL(Vassiliadis, n.d.). Gambar 1.1 merupakan model proses ETL.

Gambar 1.1 Ekstraksi Transform Load (ETL)

ETL merupakan proses mengidentifikasi dan mengekstrak data dari berbagai sumber, menyaring dan menyesuaikan data tersebut sesuai dengan format yang diperlukan, akhirnya mengintegrasikan dan memperbaharuinya ke dalam data warehouse (DW) (Neepa Biswas, Samiran Chattopadhyay, Gautam Mahapatra, Santanu Chatterjee, 2017). Sekitar 70% sumber daya dalam mengimplementasikan data warehouse dihabiskan untuk proses ETL(Liu et al., 2014) (Sun, 2017), karena proses ini sangat penting dan memakan waktu serta kesalahan pada proses ETL akan menyebabkan pengambilan keputusan yang salah(Trujillo & Luján-Mora, 2003). Data bisnis biasanya heterogen dan disimpan dalam format terstruktur, semi-terstruktur atau tidak terstruktur. Mengintegrasikan keragaman dan kuantitas data secara efektif dan efisien merupakan tantangan(Kathiravelu et al., 2018). Proses ETL harus mampu menyediakan data untuk DW, dimana Sumber datanya berasal dari data yang heterogen dan terdistribusi, serta dapat memastikan kualitas data yang diharapkan tersimpan pada DW(El Akkaoui et al., 2011).

Proses ETL yang tepat akan mempengaruhi hasil data yang tersimpan pada DW sehingga dapat menyediakan informasi yang tepat kepada orang yang tepat dan waktu yang tepat. ETL telah menarik perhatian para peneliti, khususnya pada awal tahun 2000-an. Beberapa peneliti telah melakukan penelitian dengan tujuan pengembangan model ETL, beberapa diantaranya (Vassiliadis et al., 2002) (Trujillo & Luján-Mora, 2003) (Vassiliadis et al., 2003) (Vassiliadis et al., 2005) (Simitsis et al., 2005) (El Akkaoui & Zimányi, 2009) (Vincenzo Deufemia*,†, Massimiliano Giordano, 2009). Terdapat beberapaa penelitian yang fokus pada semantic di model pendekatan ETL diantaranya (Simitsis, 2005) (Simitsis & Vassiliadis, 2008) (Skoutas & Simitsis, 2006). Proses ETL saat ini (Badiuzzaman

Biplob & Mokammel Haque, 2022; El-Sappagh et al., 2011) tidak efektif lagi untuk menghasilkan data warehouse, terlebih di era Big Data. Data digital dibuat ketika organisasi atau instansi mengubah data mereka dari analog ke digital, dimana data yang dihasilkan dapat mencapai lebih dari exabytes ≈ 1016(Pavan Kumar & Dhinesh Babu, 2019). BI berfokus pada data terstruktur dan internal perusahaan. Akibatnya banyak informasi berharga yang tertanam dalam data eksternal dan tidak terstruktur tetap tersembunyi, yang berpotensi menyebabkan pandangan yang tidak lengkap dan menghasilkan pengambilan keputusan yang bias(Ram et al., 2016a). Penggunaan big data dengan baik dapat meningkatkan 60% margin operasi dengan memperoleh pangsa pasar atas pesaingnya dan mengeksploitasi data konsumen dengan lebih rinci(Ram et al., 2016b). Di era Big Data proses ETL menjadi lebih menantang karena keragaman sumber data, volume, kecepatan, dan kompleksitas data. ETL bukan tahapan yang mudah dalam teknologi klasik, terlebih di era big data proses ETL menajdi lebih kompleks(Sirin & Karacan, 2017). Informasi yang diambil di Big Data dalam berbagai jenis dan berbagai sumber diintegrasikan dengan *database* sistem informasi untuk menghasilkan informasi yang ringkas dan prediktif bagi manajemen dalam pengambilan keputusan(Chittayasothorn, 2019).

Berbagai upaya pengembangan dilakukan terhadap proses ETL yang semakin kompleks data dan tugas ETL. Saat ini ETL dianggap sebagai isu terpenting di bidang *Business Intelligence(Bala et al., 2016)*. Beberapa model pendekatan untuk proses ETL yang sudah mengakomodir big data dalam prosesnya diantaranya adalah (Liu & Thomsen, 2014) (Bala et al., 2016; Liu et al., 2013). Berbagai pendekatan yang telah dilakukan bertujuan untuk memaksimalkan proses integrasi data sehingga dihasilkan data yang memenuhi persyaratan dan kebutuhan. Pendekatan integrasi data yang digunakan dalam kerangka kerja *Business Intelligence* yang sudah ada masih perlu dilakukan peningkatan kinerja, efisiensi waktu dan peningkatan kualitas data. Dalam proses integrasi data, banyak masalah yang akan mempengaruhi kualitas data (Souibgui et al., 2019b). Kualitas data yang mendasari sangat menentukan kualitas pengetahuan yang diekstraksi. Oleh karena

itu, kualitas data menjadi perhatian yang signifikan dalam analisis data, dan kualitas data merupakan prasyarat untuk memperoleh pengetahuan yang berkualitas. Masalah nilai hilang terjadi karena nilai yang hilang dari atribut yang disebabkan oleh kesalahan saat mengumpulkan data, kesalahan sistem ((Fernando et al., 2021; D. Li et al., 2022)), kesalahan dalam entri data, penolakan atau ketidakmampuan responden untuk memberikan jawaban yang akurat (Doquire & Verleysen, 2012) dan penggabungan data yang tidak terkait (Emmanuel et al., 2021a). Nilai yang hilang adalah masalah mendasar dalam ilmu data (Chen et al., 2023).

Dalam beberapa aplikasi, nilai yang hilang tidak dapat ditoleransi dan harus diganti dengan nilai konkret (Shahbazian & Trubitsyna, 2023a). Studi terkait telah menunjukkan bahwa imputasi nilai yang hilang bermanfaat dan merupakan pilihan yang lebih baik daripada penghapusan data (Huang et al., 2016). Imputasi data yang hilang berarti mengganti atau mengoreksi data yang hilang dengan nilai yang wajar untuk mencapai kelengkapan (Thomas & Rajabi, 2021). Imputasi data yang hilang sangat penting karena kesalahan pengambilan keputusan akan terjadi ketika kumpulan data yang tidak lengkap didukung (Ismail et al., 2022). Beberapa dampak penting dari penanganan data yang hilang antara lain keakuratan analisis statistik, interpretasi yang lebih baik, pengurangan bias, dan peningkatan kualitas data ((D. Li et al., 2022; Setiawan et al., 2023)).

Pendekatan imputasi nilai yang hilang dapat dikategorikan secara luas ke dalam metode tradisional dan metode algoritma berbasis Machine Learning (ML). Metode tradisional termasuk rata-rata(Yulian Pamuji et al., 2024), median, linear regression(Karmitsa et al., 2022), dan mode. Beberapa metode berbasis ML meliputi Algorithms Clustering(Dubey & Rasool, 2020), K-Narest Neighbor (KNN) (Sudrajat & Cholid, 2023), Support Vector Machine (SVM) (Syarif et al., n.d.-a), Decision Trees (DT)(Nikfalazar et al., 2020; Syarif et al., n.d.-b), Random Forest (RF)(Alsaber et al., 2021) dan Generative Adversarial Networks (GAN) ((W. Dong et al., 2021; Gao et al., 2023; *Mixed Data Imputation Using Generative Adversarial*

Networks, n.d.; Ou et al., 2024; Shahbazian & Greco, 2023)). Kemampuan untuk mengoptimalkan dan mengekstrak hubungan antar titik data merupakan keuntungan dari metode berbasis pembelajaran mesin (Shahbazian & Trubitsyna, 2023a). GAN adalah metode ML yang telah menarik perhatian peneliti dalam beberapa tahun terakhir. Nilai yang hilang adalah masalah yang signifikan dalam penambangan data, analisis big data, dan alur pengambilan keputusan berbasis ML, karena hasil penambangan atau analisis akhir dapat terpengaruh secara negatif ketika data yang tidak lengkap tidak diperhitungkan dengan benar (Hasan et al., 2021). Upaya perbaikan telah dilakukan dalam beberapa penelitian yang mendasari metode GAN, termasuk penelitian yang disajikan dalam (J. Yoon et al., 2018a) mengusulkan perbaikan dalam metode baru, yaitu Generative Adversarial Imputation Nets (GAIN) (J. Yoon et al., 2018b). Dalam metode ini, generator secara akurat memperhitungkan data yang hilang, dan diskriminator bertujuan untuk membedakan antara komponen yang diamati dan diperhitungkan. Perbaikan lebih lanjut terhadap GAIN dilakukan dengan penelitian (Shahbazian & Trubitsyna, 2023), di mana idenya adalah menggunakan dekonvolusi pada generator dan diskriminator (DEGAIN). Metode ini melakukan perbaikan dengan menambahkan dekonvolusi untuk menghilangkan korelasi antar data. Perbaikan metode imputasi dilakukan berdasarkan karakteristik struktur data ((Rosado-Galindo & Dávila-Padilla, 2020; Sefidian & Daneshpour, 2019) dan karakteristik nilai data. Pada saat yang sama, penelitian yang berfokus pada karakteristik nilai data masih jarang dilakukan. Karakteristik nilai data merupakan langkah awal yang penting untuk melakukan imputasi yang akurat. Perbedaan nilai data yang tinggi akan mengakibatkan hasil yang tidak akurat dalam pemrosesan data.

Dalam penelitian ini, kami mengoptimalkan metode GAIN (J. Yoon et al., 2018b), algoritma berbasis GAN, dengan mengembangkan versi yang disempurnakan yang disebut sebagai CN-GAIN. Metode CN-GAIN meningkatkan GAIN dengan memasukkan tugas prapemrosesan data sebagai langkah awal

sebelum imputasi, dengan mempertimbangkan karakteristik data yang ada. Langkah-langkah prapemrosesan ini meliputi klasifikasi data menggunakan metode k-means dan normalisasi/denormalisasi menggunakan scaler yang kuat. Tujuan klasifikasi data adalah untuk mengkategorikan data berdasarkan karakteristik yang melekat (El-Bakry et al., n.d.). Sementara itu, normalisasi dan denormalisasi memastikan bahwa tidak ada nilai data yang mendominasi kumpulan data secara tidak proporsional. Kami mengevaluasi kinerja metode yang kami usulkan menggunakan kumpulan data UKM dari sebuah kota di Provinsi Sumatera Selatan. Evaluasi tersebut meliputi pengukuran akurasi dan beberapa metrik kesalahan seperti *Root Mean Square Error* (RMSE), *Mean Absolute Error* (MAE), dan *Mean Squared Error* (MSE).

Penelitian ini dimaksudkan untuk meningkatkan kinerja dan efisiensi waktu pada model integrasi data, sehingga dapat meningkatan kualitas data yang tersimpan di *data warehouse* dengan menggunakan metode-metode data mining. Dengan proses integrasi data yang baik dan data yang berkualitas akan mendukung *Business Intelligence* dalam ketepatan pengambilan keputusan bisnis, khususnya dalam pengelolaan dan pengembangan UMKM.

1.2 RUMUSAN MASALAH

Proses integrasi data merupakan aspek penting dalam *business intelligence*, dimana proses integrasi data akan menyediakan data pada data warehouse yang digunakan dalam pengambilan keputusan. Kualitas data yang tersimpan di data warehouse sangat dipengaruhi oleh model pendekatan dalam integrasi data. Dengan demikian masalah-masalah berikut ini perlu dikaji lebih lanjut dan mendalam guna menyelesaikan permasalah diatas, yaitu:

1. Bagaimana meningkatkan kualitas data dalam proses integrasi dengan penedekatan ETL, khususnya pada proses ekstrak data sebagai gerbang masuknya ada di *business intelligence*?

2. Bagaimana algoritma k-Means Clustering dapat mengelompokkan dataset sebelum dilakukan imputisasi data?
3. Bagaimana algoritma robus sceler dapat menormalisasi dan mendenormalisasi data untuk mengurangi outleyer dataset sebelum dilakukan imputisasi data.
4. Bagaiman mengoptimalkan model GAIN dalam meningkatkan kualitas data dari data yang tidak lengkap sebelum dilakukan proses transformasi data?
5. Bagaiaman rancangan kerangka kerja integrasi data dalam *business intelligence*?

1.3 TUJUAN PENELITIAN

Penelitian ini membuat model konseptual integrasi data dengan memperhatikan karakteristik big data. Dengan demikian tujuan penelitian yang akan dicapai dalam disertasi ini adalah sebagai berikut:

1. Melakukan perbaikan proses ETL khususnya ekstrak dan transformasi sebelum disimpan pada data warehouse.
2. Meningkatkan kualitas data dengan melakukan imputisasi data pada dataset dengan nilai null.
3. Merancang model integrasi data untuk *business intelligence* yang menyediakan informasi berkualitas dalam pengambilan keputusan.

1.4 MANFAAT PENELITIAN

Hasil penelitian ini diharapkan dapat bermanfaat bagi berbagai pihak untuk berbagai hal sebagai berikut:

1. Menghasilkan model pendekatan integrasi data yang tepat, khususnya pada pengelolaan dan pengembangan UMKM pada dinas pemerintah.
2. Menyediakan data berkualitas yang dapat digunakan dalam *business intelligence* guna mendukung pengambilan keputusan bisnis.
3. Menghasilkan kerangka kerja *business intelligence* yang sesuai dengan kebutuhan khususnya dalam pengelolaan dan pengembangan UMKM.

1.5 BATASAN PENELITIAN

Penelitian ini dibatasi pada elemen integrasi data pada *business intelligence*, khususnya pada ekstrak dan transform, untuk load tidak dibahas dalam penelitian ini. Sumber data yang digunakan dalam penelitian ini adalah data sekunder yang berasal dari hasil pengumpulan data yang dilakukan oleh dinas koperasi dan UMKM Provinsi Sumatera Selatan dalam bentuk excel. Optimalisasi metode yang digunakan adalah dari metode deep learning.

1.6 SISTEMATIKA LAPORAN

Pada bagian ini menjelaskan sistematika laporan penelitian secara garis besar, yang dijelaskan menjadi 5 (lima) bab sebagai berikut:

BAB 1 PENDAHULUAN

Bab ini berisikan uraian mengenai elemen pada *business intelligence* khususnya pada integrasi data dan permasalahannya sehingga menjadi alasan dilakukannya pada penelitian Disertasi ini. Bab ini juga menyajikan permasalahan yang timbul, tujuan penelitian dan manfaat yang diperoleh dari penelitian ini dengan batasan penelitian.

BAB 2 LANDASAN TEORI

Bab ini berisi toeri-teori yang terkait dengan penulisan Disertasi ini serta menguraikan hasil-hasil penelitian yang telah dilakukan oleh peneliti sebelumnya yang juga berkenaan dengan penelitian ini. Bab ini merupakan rujukan dasar untuk membahas objek yang akan diteliti.

BAB 3 METODOLOGI PENELITIAN

Bab ini berisi metodologi yang menggambarkan langkah-langkah sistematis yang dilakukan oleh peneliti guna mencapai tujuan penelitian. Langkah-langkah sistematis tersebut diarahkan untuk menghasilan suatu

pengembangan atau penciptaan baru dari metode sebelumnnya yang sesuai dengan topik penelitian.

BAB 4 HASIL DAN PEMBAHASAN

Bab ini berisi uraian dari hasil penelitian dari penerapan kerangka model integrasi yang dilakukan melalui serangkaian penggunaan model. Bab ini juga memaparkan hasil dari proses kerangka integrasi data yang dilakukan dan juga hasil uji yang dilakukan oleh peneliti.

BAB 5 PENUTUP

Bab ini berisi kesimpulan yang secara singkat menjelaskan pokok-pokok hasil penelitian dan saran untuk pengembangan penelitian pada masa yang akan dating.

DAFTAR PUSTAKA

LAMPIRAN

DAFTAR PUSTAKA

Alsaber, A. R., Pan, J., & Al-Hurban, A. (2021). Handling complex missing data using random forest approach for an air quality monitoring dataset: A case study of kuwait environmental data (2012 to 2018). *International Journal of Environmental Research and Public Health*, *18*(3), 1–26. https://doi.org/10.3390/ijerph18031333

Andresvelosacuneduco, A. V., Quijano, A., Nydiamartinezcuneduco, C. M., Gustavopaboncuneduco, G. P., & Jorgeportellacuneduco, J. P. (2021). *BUSINESS INTELLIGENCE AND ITS BIG EVOLUTION*. *March*, 1–13.

Awad, F. H., & Hamad, M. M. (2022). Improved k-Means Clustering Algorithm for Big Data Based on Distributed SmartphoneNeural Engine Processor. *Electronics (Switzerland)*, *11*(6). https://doi.org/10.3390/electronics11060883

Baars, H., & Kemper, H. G. (2008). Management support with structured and unstructured data - An integrated business intelligence framework. *Information Systems Management*, *25*(2), 132–148. https://doi.org/10.1080/10580530801941058

Badiuzzaman Biplob, M., & Mokammel Haque, M. (2022). Development of an Efficient ETL Technique for Data Warehouses. *Lecture Notes on Data Engineering and Communications Technologies*, *95*(January), 243–255. https://doi.org/10.1007/978-981-16-6636-0_20

Bala, M., Boussaid, O., & Alimazighi, Z. (2016). Extracting-transforming-loading modeling approach for big data analytics. *International Journal of Decision Support System Technology*, *8*(4), 50–69. https://doi.org/10.4018/IJDSST.2016100104

Batra, S. (n.d.). *A Combined Method to Impute Missing Data and Predict Accurate Value of a Target Variable in Supervised Machine Learning*. www.ijcset.net

Berliantara, A. Y., Wicaksono, S. A., & Pinandito, A. (2017). Optimasi Scheduling untuk Proses Extract , Transform , Load (ETL) pada Data Warehouse Menggunakan Metode Round Robin Data Partitioning (Studi Kasus : Universitas XYZ). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIIK) Universitas Brawijaya*, *1*(11), 1358–1366.

Bordeleau, F. E., Mosconi, E., & de Santa-Eulalia, L. A. (2018). Business intelligence in Industry 4.0: State of the art and research opportunities. *Proceedings of the Annual Hawaii International Conference on System Sciences*, *2018-Janua*, 3944–3953. https://doi.org/10.24251/hicss.2018.495

Camargo-Perez, J. A., Puentes-Velasquez, A. M., & Sanchez-Perilla, A. L. (2019). Integration of big data in small and medium organizations: Business intelligence and cloud computing. *Journal of Physics: Conference Series*, *1388*(1). https://doi.org/10.1088/1742-6596/1388/1/012029

Chang, B. J. (2018). Agile business intelligence: Combining big data and business intelligence to responsive decision model. *Journal of Internet Technology*, *19*, 1699–1706. https://doi.org/10.3966/160792642018111906007

Chen, Z., Tan, S., Chajewska, U., Rudin, C., & Caruana, R. (2023). *Missing Values and Imputation in Healthcare Data: Can Interpretable Machine Learning Help?*

Chittayasothorn, S. (2019). Some key issues in information systems, databases and big data integration. *Proceeding - 5th International Conference on Engineering, Applied Sciences and Technology, ICEAST 2019*, 1–4. https://doi.org/10.1109/ICEAST.2019.8802557

Cristobal-Salas, A., Tchernykh, A., Nesmachnow, S., García-Morales, C. Y., Santiago-Vicente, B., Herrera-Vargas, J. E., Solís-Maldonado, C., & Luna-Sánchez, R. A. (2019). ETL Processing in Business Intelligence Projects for Public Transportation Systems. *Communications in Computer and Information Science*, *1151 CCIS*, 42–50. https://doi.org/10.1007/978-3-030-38043-4_4

Dayal, U., Castellanos, M., Simitsis, A., & Wilkinson, K. (2009). Data integration flows for Business Intelligence. *Proceedings of the 12th International Conference*

on Extending Database Technology: Advances in Database Technology, EDBT'09, 1–11. https://doi.org/10.1145/1516360.1516362

Dharshinni, N. P., Azmi, F., Fawwaz, I., Husein, A. M., & Siregar, S. D. (2019). Analysis of Accuracy K-Means and Apriori Algorithms for Patient Data Clusters. *Journal of Physics: Conference Series*, *1230*(1). https://doi.org/10.1088/1742-6596/1230/1/012020

Dong, W., Fong, D. Y. T., Yoon, J. sun, Wan, E. Y. F., Bedford, L. E., Tang, E. H. M., & Lam, C. L. K. (2021). Generative adversarial networks for imputing missing data for big data clinical research. *BMC Medical Research Methodology*, *21*(1). https://doi.org/10.1186/s12874-021-01272-3

Dong, X. L., & Srivastava, D. (2015). *Big Big Data Data Integration Integration*.

Doquire, G., & Verleysen, M. (2012). Feature selection with missing data using mutual information estimators. *Neurocomputing*, *90*, 3–11. https://doi.org/10.1016/j.neucom.2012.02.031

Dr. Matthew North. (2013). *Data Mining for the Masses* (Vol. 53, Issue 9).

Dubey, A., & Rasool, A. (2020). Clustering-Based Hybrid Approach for Multivariate Missing Data Imputation. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 11, Issue 11). www.ijacsa.thesai.org

Dzulkalnine, M. F., & Sallehuddin, R. (2019). Missing data imputation with fuzzy feature selection for diabetes dataset. *SN Applied Sciences*, *1*(4). https://doi.org/10.1007/s42452-019-0383-x

El Akkaoui, Z., & Zimányi, E. (2009). Defining ETL worfklows using BPMN and BPEL. *International Conference on Information and Knowledge Management, Proceedings*, *January 2015*, 41–48. https://doi.org/10.1145/1651291.1651299

El Akkaoui, Z., Zimányi, E., Mazón, J. N., & Trujillo, J. (2011). A model-driven framework for ETL process development. *International Conference on Information and Knowledge Management, Proceedings*, *October*, 45–52. https://doi.org/10.1145/2064676.2064685

El-Bakry, M., El-Kilany, A., Mazen, S., & Ali, F. (n.d.). Fuzzy based Techniques for Handling Missing Values. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 12, Issue 3). www.ijacsa.thesai.org

El-Sappagh, S. H. A., Hendawi, A. M. A., & El Bastawissy, A. H. (2011). A proposed model for data warehouse ETL processes. *Journal of King Saud University - Computer and Information Sciences*, *23*(2), 91–104. https://doi.org/10.1016/j.jksuci.2011.05.005

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021a). A survey on missing data in machine learning. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00516-9

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021b). A survey on missing data in machine learning. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00516-9

Fernando, M. P., Cèsar, F., David, N., & José, H. O. (2021). Missing the missing values: The ugly duckling of fairness in machine learning. *International Journal of Intelligent Systems*, *36*(7), 3217–3258. https://doi.org/10.1002/int.22415

Gad, I., Hosahalli, D., Manjunatha, B. R., & Ghoneim, O. A. (2021). A robust deep learning model for missing value imputation in big NCDC dataset. *Iran Journal of Computer Science*, *4*(2), 67–84. https://doi.org/10.1007/s42044-020-00065-z

Gao, J., Cai, Z., Sun, W., & Jiao, Y. (2023). A Novel Method for Imputing Missing Values in Ship Static Data Based on Generative Adversarial Networks. *Journal of Marine Science and Engineering*, *11*(4). https://doi.org/10.3390/jmse11040806

Gunady, M. K., Kancherla, J., Corrada Bravo, H., & Feizi, S. (n.d.). *scGAIN: Single Cell RNA-seq Data Imputation using Generative Adversarial Networks*. https://doi.org/10.1101/837302

Gupta, G., Kumar, N., & Chhabra, I. (2020). Optimised Transformation Algorithm For Hadoop Data Loading in Web ETL Framework. *EAI Endorsed Transactions on Scalable Information Systems*, *7*(25), 1–8. https://doi.org/10.4108/eai.13-7-2018.160600

Hasan, M. K., Alam, M. A., Roy, S., Dutta, A., Jawad, M. T., & Das, S. (2021). Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). In *Informatics in Medicine Unlocked* (Vol. 27). Elsevier Ltd. https://doi.org/10.1016/j.imu.2021.100799

Hawking, P., & Sellitto, C. (2010). Business Intelligence (BI) critical success factors. *ACIS 2010 Proceedings - 21st Australasian Conference on Information Systems*.

Heidari, S., Alborzi, M., Radfar, R., Afsharkazemi, M. A., & Rajabzadeh Ghatari, A. (2019). Big data clustering with varied density based on MapReduce. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0236-x

Huang, M. W., Lin, W. C., Chen, C. W., Ke, S. W., Tsai, C. F., & Eberle, W. (2016). Data preprocessing issues for incomplete medical datasets. *Expert Systems*, *33*(5), 432–438. https://doi.org/10.1111/exsy.12155

Hussien, H. H., Elssayad, O. M., & El-Zoghabi, A. A. (2019). Improving MapReduce Based k-Means Algorithm using Intelligent Technique. *Asian Journal of Information Technology*, *18*(5), 150–159. https://doi.org/10.36478/ajit.2019.150.159

Hwang, J., & Suh, D. (2024). CC-GAIN: Clustering and classification-based generative adversarial imputation network for missing electricity consumption data imputation. *Expert Systems with Applications*, *255*. https://doi.org/10.1016/j.eswa.2024.124507

Imane, L., & Youness, T. (2017). State of the art in MapReduce: Issues and approaches. *ACM International Conference Proceeding Series*, *Part F1294*. https://doi.org/10.1145/3090354.3090397

Ismail, A. R., Abidin, N. Z., & Maen, M. K. (2022). Systematic Review on Missing Data Imputation Techniques with Machine Learning Algorithms for Healthcare. In *Journal of Robotics and Control (JRC)* (Vol. 3, Issue 2, pp. 143–152). Department of Electrical Engineering, Universitas Muhammadiyah Yogyakarta. https://doi.org/10.18196/jrc.v3i2.13133

Jahantigh, F. F., Habibi, A., & Sarafrazi, A. (2019). A conceptual framework for business intelligence critical success factors. *International Journal of Business Information Systems*, *30*(1), 109–123. https://doi.org/10.1504/IJBIS.2019.097058

Jain, M., & Verma, C. (2014). Adapting k-means for Clustering in Big Data. *International Journal of Computer Applications*, *101*(1), 19–24. https://doi.org/10.5120/17652-8457

Jiawei Han Micheline Kamber. (2006). *DataMining Concepts and Techniques*.

Karagiannis, A., Vassiliadis, P., & Simitsis, A. (2012). *Scheduling Strategies for Efficient ETL Execution*.

Karmitsa, N., Taheri, S., Bagirov, A., & Makinen, P. (2022). Missing Value Imputation via Clusterwise Linear Regression. *IEEE Transactions on Knowledge and Data Engineering*, *34*(4), 1889–1901. https://doi.org/10.1109/TKDE.2020.3001694

Kathiravelu, P., Sharma, A., Galhardas, H., Roy, P. Van, & Veiga, L. (2018). On-demand big data integration A hybrid ETL approach for reproducible scientific research. *Distributed and Parallel Databases*. https://doi.org/10.1007/s10619-018-7248-y

Kaushik, M., Sharma, R., Peious, S. A., Shahin, M., Yahia, S. Ben, & Draheim, D. (2021). A Systematic Assessment of Numerical Association Rule Mining Methods. *SN Computer Science*, *2*(5). https://doi.org/10.1007/s42979-021-00725-2

Khoirunnisa, A., & Ramadhan, N. G. (2023). Improving malaria prediction with ensemble learning and robust scaler: An integrated approach for enhanced accuracy. *JURNAL INFOTEL*, *15*(4), 326–334. https://doi.org/10.20895/infotel.v15i4.1056

Kumar, D., Bezdek, J. C., Palaniswami, M., Rajasegarar, S., Leckie, C., & Havens, T. C. (2016). A Hybrid Approach to Clustering in Big Data. *IEEE Transactions on Cybernetics*, *46*(10), 2372–2385. https://doi.org/10.1109/TCYB.2015.2477416

Li, D., Zhang, H., Li, T., Bouras, A., Yu, X., & Wang, T. (2022). Hybrid Missing Value Imputation Algorithms Using Fuzzy C-Means and Vaguely Quantified Rough Set.

IEEE Transactions on Fuzzy Systems, *30*(5), 1396–1408. https://doi.org/10.1109/TFUZZ.2021.3058643

Li, J. H., Guo, S. X., Ma, R. L., He, J., Zhang, X. H., Rui, D. S., Ding, Y. S., Li, Y., Jian, L. Y., Cheng, J., & Guo, H. (2024). Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets. *BMC Medical Research Methodology*, *24*(1). https://doi.org/10.1186/s12874-024-02173-x

Liao, Z., Lu, X., Yang, T., & Wang, H. (2009). Missing data imputation: A fuzzy k-means clustering algorithm over sliding window. *6th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2009*, *3*, 133–137. https://doi.org/10.1109/FSKD.2009.407

Liu, X., & Thomsen, C. (2014). *CloudETL : Scalable Dimensional ETL for Hive Categories and Subject Descriptors*. 195–206.

Liu, X., Thomsen, C., & Pedersen, T. B. (2013). ETLMR: A highly scalable dimensional ETL framework based on MapReduce. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7790 LNCS*, 1–31. https://doi.org/10.1007/978-3-642-37574-3_1

Liu, X., Thomsen, C., & Pedersen, T. B. (2014). CloudETL: Scalable dimensional ETL for hive. *ACM International Conference Proceeding Series*, 195–206. https://doi.org/10.1145/2628194.2628249

Maliakel, P. J., Ilager, S., & Brandic, I. (2024). FLIGAN: Enhancing Federated Learning with Incomplete Data using GAN. *EdgeSys 2024 - Proceedings of the 7th International Workshop on Edge Systems, Analytics and Networking, Part of: EuroSys 2024*, 1–6. https://doi.org/10.1145/3642968.3654813

Manickam, V., & Rajasekaran Indra, M. (2023). Dynamic multi-variant relational scheme-based intelligent ETL framework for healthcare management. *Soft Computing*, *27*(1), 605–614. https://doi.org/10.1007/s00500-022-06938-8

Mixed Data Imputation using Generative Adversarial Networks. (n.d.).

Neepa Biswas, Samiran Chattopadhyay, Gautam Mahapatra, Santanu Chatterjee, K. C. M. (2017). *SysML Based Conceptual ETL Process Modeling*. *776*, 72–83. https://doi.org/10.1007/978-981-10-6430-2

Nikfalazar, S., Yeh, C. H., Bedingfield, S., & Khorshidi, H. A. (2020). Missing data imputation using decision trees and fuzzy clustering with iterative learning. *Knowledge and Information Systems*, *62*(6), 2419–2437. https://doi.org/10.1007/s10115-019-01427-1

Nisha.C.M, & N. Thangarasu. (2023). Deep learning algorithms and their relevance: A review. *International Journal of Data Informatics and Intelligent Computing*, *2*(4), 1–10. https://doi.org/10.59461/ijdiic.v2i4.78

Ou, H., Yao, Y., & He, Y. (2024). Missing Data Imputation Method Combining Random Forest and Generative Adversarial Imputation Network. *Sensors*, *24*(4). https://doi.org/10.3390/s24041112

Pandey, K. K., Shukla, D., & Milan, R. (2020). A Comprehensive Study of Clustering Algorithms for Big Data Mining with MapReduce Capability. In *Lecture Notes in Networks and Systems* (Vol. 100). Springer Singapore. https://doi.org/10.1007/978-981-15-2071-6_34

Pavan Kumar, C. S., & Dhinesh Babu, L. D. (2019). Review on big data and its impact on business intelligence. *Advances in Intelligent Systems and Computing*, *862*, 93–109. https://doi.org/10.1007/978-981-13-3329-3_10

Ponniah, P. (2010). *Data Warehousing Fundamentals for IT Professionals, Second Edition*.

Putra, I. M. S., & Adhitya Putra, D. K. T. (2019). Rancang Bangun Engine ETL Data Warehouse dengan Menggunakan Bahasa Python. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *3*(2), 113–123. https://doi.org/10.29207/resti.v3i2.872

Ram, J., Zhang, C., & Koronios, A. (2016a). The Implications of Big Data Analytics on Business Intelligence: A Qualitative Study in China. *Procedia Computer Science*, *87*, 221–226. https://doi.org/10.1016/j.procs.2016.05.152

Ram, J., Zhang, C., & Koronios, A. (2016b). The Implications of Big Data Analytics on Business Intelligence: A Qualitative Study in China. *Procedia Computer Science*, *87*, 221–226. https://doi.org/10.1016/j.procs.2016.05.152

Ramzan, F., Sartori, C., Consoli, S., & Reforgiato Recupero, D. (2024). Generative Adversarial Networks for Synthetic Data Generation in Finance: Evaluating Statistical Similarities and Quality Assessment. *AI (Switzerland)*, *5*(2), 667–685. https://doi.org/10.3390/ai5020035

Razavi-Far, R., Cheng, B., Saif, M., & Ahmadi, M. (2020). Similarity-learning information-fusion schemes for missing data imputation. *Knowledge-Based Systems*, *187*. https://doi.org/10.1016/j.knosys.2019.06.013

Richards, G., Yeoh, W., Chong, A. Y. L., & Popovič, A. (2019). Business Intelligence Effectiveness and Corporate Performance Management: An Empirical Analysis. *Journal of Computer Information Systems*, *59*(2), 188–196. https://doi.org/10.1080/08874417.2017.1334244

Rodzi, N. A. H. M., Othman, M. S., & Yusuf, L. M. (2016). Significance of data integration and ETL in business intelligence framework for higher education. *Proceedings - 2015 International Conference on Science in Information Technology: Big Data Spectrum for Future Information Economy, ICSITech 2015*, 181–186. https://doi.org/10.1109/ICSITech.2015.7407800

Rosado-Galindo, H., & Dávila-Padilla, S. (2020). Tree-Based Missing Value Imputation Using Feature Selection. *Journal of Data Science*, *18*(4), 606–631. https://doi.org/10.6339/JDS.202010_18(4).0002

Runtuwene, J. P. A., Tangkawarow, I. R. H. T., Manoppo, C. T. M., & Salaki, R. J. (2018). A Comparative Analysis of Extract, Transformation and Loading (ETL) Process. *IOP Conference Series: Materials Science and Engineering*, *306*(1). https://doi.org/10.1088/1757-899X/306/1/012066

R.Vishwanath, P., Rajyalakshmi, R., & Reddy, S. (2015). An Association Rule Mining for Materialized View Selection and View Maintanance. *International Journal of Computer Applications*, *109*(5), 15–20. https://doi.org/10.5120/19184-0670

Sarker, I. H. (2021). Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. In *SN Computer Science* (Vol. 2, Issue 6). Springer. https://doi.org/10.1007/s42979-021-00815-1

Sefidian, A. M., & Daneshpour, N. (2019). Missing value imputation using a novel grey based fuzzy c-means, mutual information based feature selection, and regression model. *Expert Systems with Applications*, *115*, 68–94. https://doi.org/10.1016/j.eswa.2018.07.057

Setiawan, I., Gernowo, R., & Warsito, B. (2023). A Systematic Literature Review on Missing Values: Research Trends, Datasets, Methods and Frameworks. *E3S Web of Conferences*, *448*. https://doi.org/10.1051/e3sconf/202344802020

Shahbazian, R., & Greco, S. (2023). Generative Adversarial Networks Assist Missing Data Imputation: A Comprehensive Survey and Evaluation. *IEEE Access*, *11*, 88908–88928. https://doi.org/10.1109/ACCESS.2023.3306721

Shahbazian, R., & Trubitsyna, I. (2023a). *DEGAIN as tool for Missing Data Imputation*. http://ceur-ws.org

Shahbazian, R., & Trubitsyna, I. (2023b). *DEGAIN as tool for Missing Data Imputation*. http://ceur-ws.org

She, X., & Zhang, L. (2016). Apriori parallel improved algorithm based on MapReduce distributed architecture. *Proceedings - 2016 6th International Conference on Instrumentation and Measurement, Computer, Communication and Control, IMCCC 2016*, 517–521. https://doi.org/10.1109/IMCCC.2016.59

Sherman, R. (2014). *Business Intelligence Guidebook: From Data Integration to Analytics 1st Edition, Kindle Edition*.

Sherman, R. (2015). *Business Intelligence Guidebook: From Data Integration to Analytics - Books24x7*.

Simitsis, A. (2005). Mapping conceptual to logical models for ETL processes. *DOLAP 2005 - Proceedings of the 8th ACM International Workshop on Data Warehousing and OLAP, Co-Located with CIKM 2005*, 67–76. https://doi.org/10.1145/1097002.1097014

Simitsis, A., & Vassiliadis, P. (2008). A method for the mapping of conceptual designs to logical blueprints for ETL processes. *Decision Support Systems*, *45*(1), 22–40. https://doi.org/10.1016/j.dss.2006.12.002

Simitsis, A., Vassiliadis, P., Terrovitis, M., & Skiadopoulos, S. (2005). Graph-based modeling of ETL activities with multi-level transformations and updates. *Lecture Notes in Computer Science*, *3589*, 43–52. https://doi.org/10.1007/11546849_5

Sinha, A., & Jana, P. K. (2018). A hybrid MapReduce-based k-means clustering using genetic algorithm for distributed datasets. *Journal of Supercomputing*, *74*(4), 1562–1579. https://doi.org/10.1007/s11227-017-2182-8

Sirin, E., & Karacan, H. (2017). *A Review on Business Intelligence and Big Data*. https://doi.org/10.1039/b0000

Skoutas, D., & Simitsis, A. (2006). Designing ETL processes using semantic web technologies. *DOLAP: Proceedings of the ACM International Workshop on Data Warehousing and OLAP*, 67–74. https://doi.org/10.1145/1183512.1183526

Souibgui, M., Atigui, F., Zammali, S., Cherfi, S., & Yahia, S. Ben. (2019a). Data quality in ETL process: A preliminary study. *Procedia Computer Science*, *159*, 676–687. https://doi.org/10.1016/j.procs.2019.09.223

Souibgui, M., Atigui, F., Zammali, S., Cherfi, S., & Yahia, S. Ben. (2019b). Data quality in ETL process: A preliminary study. *Procedia Computer Science*, *159*, 676–687. https://doi.org/10.1016/j.procs.2019.09.223

Souibgui, M., Atigui, F., Zammali, S., Cherfi, S., & Yahia, S. Ben. (2019c). Data quality in ETL process: A preliminary study. *Procedia Computer Science*, *159*, 676–687. https://doi.org/10.1016/j.procs.2019.09.223

Sreemathy, J., Infant Joseph, V., Nisha, S., Chaaru Prabha, I., & Gokula Priya, R. M. (2020). Data Integration in ETL Using TALEND. *2020 6th International Conference on Advanced Computing and Communication Systems, ICACCS 2020*, 1444–1448. https://doi.org/10.1109/ICACCS48705.2020.9074186

Sudrajat, W., & Cholid, I. (2023). K-NEAREST NEIGHBOR (K-NN) UNTUK PENANGANAN MISSING VALUE PADA DATA UMKM. In *Jurnal Rekayasa Sistem Informasi dan Teknologi* (Vol. 1, Issue 2).

Sudrajat, W., Cholid, I., & Petrus, J. (n.d.). *Wahyu Sudrajat et al, Penerapan Algoritma K-Means Untuk …….*

Sun, L. (2017). *A Novel Processing Model For Scds In ETL*. *62*(Jimec), 133–136.

Sureddy, M. R., & Yallamula, P. (2020). Data Quality Architecture for Data Warehouses. *International Journal of Research Culture Society*, *4*(6), 95–100. https://doi.org/10.2017/IJRCS.2456.6683/202006017

Syarif, A., Desti Riana, O., Asiah Shofiana, D., & Junaidi, A. (n.d.-a). A Comprehensive Comparative Study of Machine Learning Methods for Chronic Kidney Disease Classification: Decision Tree, Support Vector Machine, and Naive Bayes. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 14, Issue 10). www.ijacsa.thesai.org

Syarif, A., Desti Riana, O., Asiah Shofiana, D., & Junaidi, A. (n.d.-b). A Comprehensive Comparative Study of Machine Learning Methods for Chronic Kidney Disease Classification: Decision Tree, Support Vector Machine, and Naive Bayes. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 14, Issue 10). www.ijacsa.thesai.org

Ta'a, A., & Abdullah, M. S. (2011). Goal-ontology approach for modeling and designing ETL processes. *Procedia Computer Science*, *3*, 942–948. https://doi.org/10.1016/j.procs.2010.12.154

Thomas, T., & Rajabi, E. (2021). A systematic review of machine learning-based missing value imputation techniques. *Data Technologies and Applications*, *55*(4), 558–585. https://doi.org/10.1108/DTA-12-2020-0298

Trujillo, J., & Luján-Mora, S. (2003). A UML based approach for modeling ETL processes in data warehouses. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *2813*, 307–320. https://doi.org/10.1007/978-3-540-39648-2_25

UU No. 9 Tahun 1995 Tentang Usaha Kecil, 11 296 (1995).

Vaisman, A. (2014). *Data Warehouse Systems*.

Vassiliadis, P. (n.d.). *Extraction, transformation, and loading*.

Vassiliadis, P., Simitsis, A., Georgantas, P., & Terrovitis, M. (2003). A framework for the design of ETL scenarios. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *2681*, 520–535. https://doi.org/10.1007/3-540-45017-3_35

Vassiliadis, P., Simitsis, A., Georgantas, P., Terrovitis, M., & Skiadopoulos, S. (2005). A generic and customizable framework for the design of ETL scenarios. *Information Systems*, *30*(7), 492–525. https://doi.org/10.1016/j.is.2004.11.002

Vassiliadis, P., Simitsis, A., & Skiadopoulos, S. (2002). Conceptual modeling for ETL processes. *ACM International Workshop on Data Warehousing and OLAP (DOLAP)*, *December 2002*, 14–21. https://doi.org/10.1145/583890.583893

Vercellis, C. (2009). Business Intelligence: Data Mining and Optimization for Decision Making. In *Business Intelligence: Data Mining and Optimization for Decision Making*. https://doi.org/10.1002/9780470753866

Vincenzo Deufemia*,†, Massimiliano Giordano, G. P. and G. T. (2009). A visual language-based system for extraction–transformation–loading development Vincenzo. *Software - Practice and Experience*, *39*(7), 701–736. https://doi.org/10.1002/spe

Visalakshi, K. N., Shanthi, S., & Lakshmi, K. (2021). MapReduce-Based Crow Search-Adopted Partitional Clustering Algorithms for Handling Large-Scale Data. *International Journal of Cognitive Informatics and Natural Intelligence*, *15*(4), 1–23. https://doi.org/10.4018/IJCINI.20211001.oa32

Wang, H. Bin, & Gao, Y. J. (2021). Research on parallelization of Apriori algorithm in association rule mining. *Procedia Computer Science*, *183*, 641–647. https://doi.org/10.1016/j.procs.2021.02.109

Wang, Y., Li, D., Li, X., & Yang, M. (2020). *PC-GAIN: Pseudo-label Conditional Generative Adversarial Imputation Networks for Incomplete Data*. http://arxiv.org/abs/2011.07770

Wijaya, R., & Pudjoatmodjo, B. (2015). An overview and implementation of extraction-transformation-loading (ETL) process in data warehouse (Case study: Department of agriculture). *2015 3rd International Conference on Information and Communication Technology, ICoICT 2015*, 70–74. https://doi.org/10.1109/ICoICT.2015.7231399

Williams, S. (2016). Business Intelligence Strategy and Big Data Analytics: A General Management Perspective. In *Business Intelligence Strategy and Big Data Analytics: a General Management Perspective*.

Williams, S., & Williams, N. (2007). *The Profit Impact of Business Intelligence*.

Wulandari, G. F. (2014). Segmantasi Pelanggan Menggunakan Algoritma K-Means Untuk Customer Relationship Management (CRM) Pada Hijab Miulan. *Industrial Marketing Management*, *I*(segmentasi pelanggan), 7.

Yan, D., Zhao, X., Lin, R., & Bai, D. (2018). PPQAR: Parallel PSO for Quantitative Association Rule Mining. *Proceedings - 2018 IEEE International Conference on Big Data and Smart Computing, BigComp 2018*, 163–169. https://doi.org/10.1109/BigComp.2018.00032

Yoon, J., Jordon, J., & van der Schaar, M. (2018a). *GAIN: Missing Data Imputation using Generative Adversarial Nets*. http://arxiv.org/abs/1806.02920

Yoon, J., Jordon, J., & van der Schaar, M. (2018b). *GAIN: Missing Data Imputation using Generative Adversarial Nets*. http://arxiv.org/abs/1806.02920

Yoon, J., Jordon, J., & van der Schaar, M. (2018c). *GAIN: Missing Data Imputation using Generative Adversarial Nets*. http://arxiv.org/abs/1806.02920

Yoon, S., & Sull, S. (2020). Gamin: Generative adversarial multiple imputation network for highly missing data. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 8453–8461. https://doi.org/10.1109/CVPR42600.2020.00848

Yulian Pamuji, F., Ahmad Rofiqul Muslikh, Rizza Muhammad Arief, & Delviana Muti. (2024). Komparasi Metode Mean dan KNN Imputation dalam Mengatasi Missing Value pada Dataset Kecil. *Jurnal Informatika Polinema*, *10*(2), 257–264. https://doi.org/10.33795/jip.v10i2.5031

Zada, I., Ali, S., Khan, I., Hadjouni, M., Elmannai, H., Zeeshan, M., Serat, A. M., & Jameel, A. (2022). Performance Evaluation of Simple K -Mean and Parallel K -Mean Clustering Algorithms: Big Data Business Process Management Concept. *Mobile Information Systems*, *2022*. https://doi.org/10.1155/2022/1277765