

BAB II

LANDASAN TEORI

2.1 Penelitian Terdahulu

Kajian terhadap penelitian-penelitian terdahulu dilakukan untuk memetakan perkembangan riset di bidang *Business Intelligence* (BI), *Big Data Analytics*, dan analisis data penjualan, serta untuk mengidentifikasi posisi penelitian ini dalam ekosistem keilmuan terkait[6]. Tabel 2.1 menyajikan ringkasan dari lima belas penelitian terdahulu yang menjadi referensi utama dalam penelitian ini, yang dipilih berdasarkan kriteria relevansi topik, kebaruan (tahun publikasi 2021 hingga 2025), dan kualitas metodologi[7].

Tabel 2.1 Penelitian Terdahulu

No.	Penulis (Tahun)	Metode / Framework	Temuan Utama	Kekurangan / Gap
1	Nurdin & Salmi (2022)	ETL → DW → BI <i>Dashboard</i>	BI mengubah data mentah menjadi informasi penjualan	Data internal; bukan data publik/KDD <i>Big Data</i>
2	Annas Purnama & Nyoman (2022)	<i>OLAP</i> (roll-up/drill- down/slice/dice)	<i>OLAP</i> efektif untuk analisis multi-dimensi	Bukan skala nasional; belum <i>Big Data</i>
3	Azzahra & Apriyanto (2023)	DW → <i>Star schema</i> → ETL	<i>Star schema</i> meningkatkan efisiensi <i>query</i>	Fokus stok; bukan penjualan otomotif
4	Basri dkk. (2023)	<i>Big Data Analytics</i> (Hadoop/Spark)	<i>Big Data</i> menangani data industri masif	Belum mengintegrasikan <i>OLAP</i>

No.	Penulis (Tahun)	Metode / Framework	Temuan Utama	Kekurangan / Gap
5	Avinash & Widjaja (2024)	<i>Machine Learning</i> (RF, XGBoost, LSTM) — KDD/Prediktif	ML akurat untuk <i>forecasting</i> persediaan otomotif	Prediktif murni; tanpa <i>OLAP</i> deskriptif
6	Fikri dkk. (2025)	Time Series (ARIMA/Prophet) — KDD/ <i>Forecasting</i>	Time-series memproyeksikan penjualan EV	Fokus prediksi EV; bukan multi-kategori
7	Tharinda dkk. (2025)	Predictive Analytics / ML — KDD/Prediktif	Analitik prediktif akurat untuk demand <i>forecasting</i>	Fokus prediksi; bukan deskriptif <i>OLAP</i>
8	(Tambahan) studi <i>Hybrid</i>	<i>Hybrid OLAP + Data Mining</i>	Menggabungkan analitik deskriptif & prediktif	Belum pada data publik GAIKINDO multi-periode
	RESEARCH GAP	<i>Hybrid OLAP + Data Mining</i> (CRISP-DM) atas DATA PUBLIK GAIKINDO	Belum ada penelitian yang mengintegrasikan data publik GAIKINDO multi-periode & multi-kategori dengan pendekatan <i>hybrid OLAP</i> (deskriptif) + <i>Data Mining</i> /ML (prediktif) lalu menyajikannya ke publik via <i>dashboard web</i> .	Posisi kontribusi penelitian ini

Berdasarkan kajian terhadap lima belas penelitian terdahulu di atas, dapat diidentifikasi tiga kesenjangan penelitian (*research gap*) yang jelas. Pertama, dari sisi objek data, mayoritas penelitian *Business Intelligence* (penelitian terdahulu 1–9) menggunakan data penjualan internal perusahaan dengan akses terbatas, sehingga belum ada yang memanfaatkan data publik nasional GAIKINDO secara komprehensif. Kedua, dari sisi rentang dan integrasi data, sebagian besar penelitian hanya mengolah data satu periode atau satu kategori, sedangkan integrasi data multi-periode (2020 hingga 2025) dan lintas-kategori (*wholesales, retail sales, produksi, ekspor, impor*) belum dibahas secara

sistematis. Ketiga, dari sisi pendekatan analitik, kelompok penelitian terbaru (penelitian terdahulu 10–15) didominasi oleh analitik prediktif/*forecasting* berbasis *Machine Learning*, sementara penerapan analitik deskriptif berbasis *OLAP* dengan kerangka kerja CRISP-DM untuk menyajikan data publik melalui *dashboard* berbasis web yang dapat diakses gratis masih sangat jarang dijumpai.

Ketiga kesenjangan tersebut menunjukkan adanya celah penelitian yang belum terisi, yaitu belum tersedianya sistem *Business Intelligence* yang secara khusus mengintegrasikan data publik GAIKINDO multi-periode dan multi-kategori menggunakan kerangka kerja CRISP-DM, lalu menyajikannya kepada publik melalui *dashboard* analitik berbasis web. Posisi inilah yang menjadi kontribusi utama penelitian ini, yaitu menjembatani kesenjangan antara ketersediaan data publik GAIKINDO dengan kebutuhan pengguna umum, akademisi, dan pelaku usaha akan informasi penjualan otomotif nasional yang terstruktur dan mudah dipahami [6], [7].

2.2 Teori yang Berkaitan

2.2.1 *Big Data Analytics*

Big Data Analytics adalah proses pemeriksaan, pembersihan, *Transformasi*, dan pemodelan data dalam jumlah, kecepatan, dan variasi yang sangat besar untuk menemukan informasi bermakna, pola tersembunyi, korelasi, serta tren yang dapat mendukung pengambilan keputusan [8], [9]. Karakteristik utama *Big Data* dikenal dengan istilah 5V, yaitu *Volume* (*Volume* data yang besar), *Velocity* (kecepatan produksi data), *Variety* (keberagaman jenis data), *Veracity* (validitas dan kualitas data), dan *Value* (nilai bisnis yang dapat diekstrak)[8].

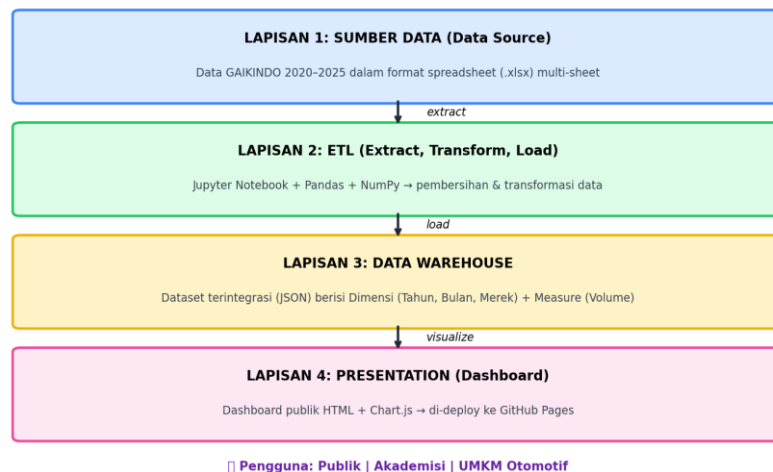
Dalam konteks data penjualan otomotif nasional, karakteristik *Big Data* tercermin dari *Volume* data lintas tahun (2020 hingga 2025), variasi kategori (*wholesales*, *retail sales*, produksi, ekspor, impor), serta nilai bisnis yang dapat dihasilkan dari analisis tren dan distribusi pasar[9], [10]. Penerapan *Big Data Analytics* memungkinkan pengolahan data yang sebelumnya bersifat mentah dan terfragmentasi menjadi informasi terstruktur yang dapat dimanfaatkan oleh berbagai pemangku kepentingan[5], [8].

2.2.2 *Business Intelligence (BI)*

Business Intelligence (BI) merupakan kombinasi teknologi, aplikasi, dan praktik untuk mengumpulkan, mengintegrasikan, menganalisis, dan menyajikan informasi bisnis dengan tujuan mendukung pengambilan keputusan yang lebih baik[5], [10]. BI terdiri dari serangkaian proses yang mencakup ekstraksi data dari berbagai sumber (*Extract*), Transformasi data ke dalam format yang konsisten (*Transform*), pemuatan ke dalam tempat penyimpanan terpusat (*Load*), serta penyajian informasi melalui *Dashboard* atau laporan analitik[3], [11].

Komponen utama sistem BI terdiri dari empat lapisan, yaitu sumber data (*data source*), proses *ETL*, *Data Warehouse*, dan lapisan presentasi (*presentation layer*) berupa *Dashboard* atau laporan, sebagaimana diilustrasikan pada Gambar 2.1 9. Setiap lapisan memiliki peran spesifik dalam mengubah data mentah menjadi informasi bermakna yang dapat diakses oleh pengguna[5], [12]. Penerapan BI tidak terbatas pada lingkup perusahaan, tetapi juga dapat dimanfaatkan untuk menyajikan data publik kepada masyarakat luas melalui *Dashboard* interaktif berbasis web[11], [13].

Arsitektur Sistem Business Intelligence (4 Lapisan)



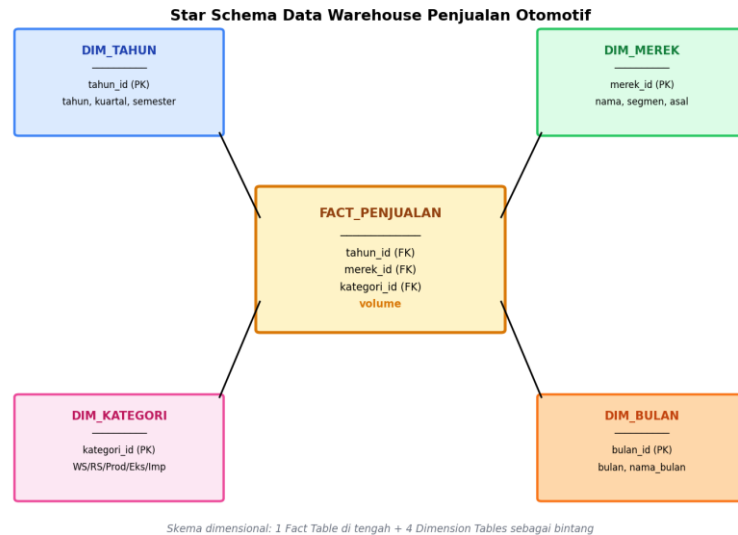
Gambar 2.1 Arsitektur Sistem Business Intelligence Empat Lapisan

2.2.3 Data Warehouse

Data Warehouse adalah basis data terintegrasi yang dirancang khusus untuk mendukung kegiatan analitik dan pengambilan keputusan, berbeda dari basis data operasional yang berorientasi pada transaksi sehari-hari[7]. Karakteristik utama *Data Warehouse* meliputi subject-oriented (berorientasi pada subjek bisnis), *integrated* (mengintegrasikan data dari berbagai sumber), *time-variant* (memuat data historis), dan *non-volatile* (data tidak diubah setelah dimuat)[7].

Skema yang umum digunakan dalam *Data Warehouse* adalah *star schema* (skema bintang) yang terdiri dari satu *fact table* di tengah dan beberapa *dimension table* di sekitarnya, sebagaimana terlihat pada Gambar 2.2. Skema ini dipilih karena kemudahan dalam memahami struktur data dan efisiensi performa *query* analitik[3], [14]. Dalam penelitian ini, data historis GAIKINDO disusun dalam struktur tabel terintegrasi yang

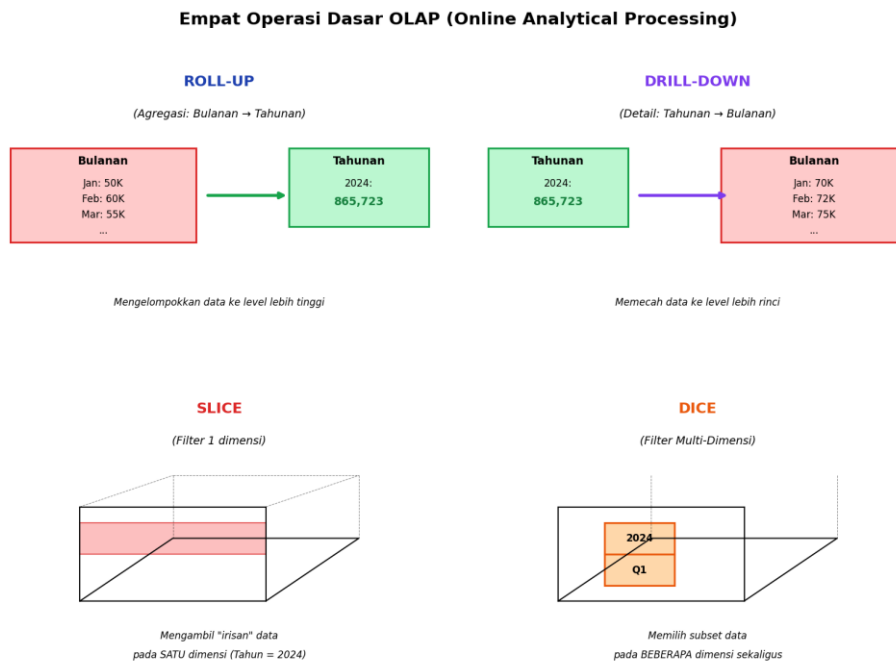
berfungsi serupa dengan *Data Warehouse* mini untuk mendukung analitik deskriptif[7], [9].



Gambar 2.2 Star schema Data Warehouse Penjualan Otomotif

2.2.4 Online Analytical Processing (OLAP)

Online Analytical Processing (OLAP) adalah kategori teknik analisis data yang memungkinkan pengguna menganalisis data dari berbagai sudut pandang (dimensi) secara cepat dan interaktif[7]. *OLAP* memiliki empat operasi dasar yang ditunjukkan pada Gambar 2.3, yaitu: (1) *roll-up* agregasi data ke level yang lebih tinggi (misalnya dari bulanan ke tahunan); (2) *drill-down* pemecahan data ke level yang lebih rinci (misalnya dari tahunan ke bulanan); (3) *slice* pemilihan subset data pada satu dimensi tertentu (misalnya data tahun 2024 saja); dan (4) *dice* pemilihan subset data pada beberapa dimensi sekaligus [15].



Gambar 2.3 Empat Operasi Dasar OLAP

Dalam penelitian ini, operasi *OLAP* diterapkan untuk menyusun ringkasan tren penjualan tahunan, perbandingan antar merek pada periode tertentu, serta distribusi pangsa pasar pada level segmen 37. Penerapan *OLAP* didukung oleh pustaka *Pandas* pada *Python* yang menyediakan fungsi *groupby*, *pivot_table*, dan agregasi *multi-level*[16], [17].

Perlu ditegaskan bahwa *OLAP* dan *Data Mining* merupakan dua hal yang berbeda. *OLAP* (roll-up, drill-down, slice, dice, pivot) adalah analisis multidimensi yang bersifat eksploratif dan deskriptif untuk meringkas serta membandingkan data historis, dan karenanya lebih dekat ke pelaporan *Business Intelligence*. *Data Mining* adalah proses penemuan pola secara otomatis melalui metode seperti *clustering*, *classification*, *prediction*, *forecasting*, *association rule mining*, dan *anomaly detection*. Penelitian ini hanya menerapkan *OLAP*

sebagai analitik deskriptif; tidak ada metode data mining yang digunakan, sehingga penelitian ini diposisikan sebagai Business Intelligence dengan descriptive analytics, bukan penelitian data mining.

Berdasarkan cara penyimpanan dan pemrosesan data, arsitektur *OLAP* dikelompokkan menjadi tiga jenis. Pertama, *MOLAP* (Multidimensional *OLAP*) menyimpan data dalam kubus multidimensi sehingga agregasi sangat cepat, namun memerlukan ruang penyimpanan besar. Kedua, *ROLAP* (Relational *OLAP*) menyimpan data dalam basis data relasional dan menganalisisnya melalui kueri/operasi tabular; cocok untuk data besar karena memanfaatkan *data warehouse* relasional (contoh: Oracle *OLAP*, Mondrian/Pentaho, PostgreSQL + tools BI). Ketiga, *HOLAP* (Hybrid *OLAP*) menggabungkan keunggulan *MOLAP* dan *ROLAP*.

Selain itu, untuk memadukan proses transaksional/ETL dan analitik *OLAP* dalam satu alur, dikenal arsitektur HTAP (*Hybrid Transactional/Analytical Processing*). HTAP memungkinkan data hasil proses ETL langsung dianalisis secara analitik (*OLAP*) tanpa pemisahan sistem yang kaku, sehingga sesuai untuk penelitian yang menggabungkan tahap ETL dan *OLAP* pada satu lingkungan pengolahan.

Penelitian ini mengadopsi pendekatan yang sejalan dengan *ROLAP* dan prinsip HTAP: data historis GAIKINDO disusun dalam satu tabel terstruktur (flat table) yang berfungsi sebagai *data warehouse* mini relasional, lalu proses ETL dan operasi *OLAP* (*roll-up*, *drill-down*, *slice*, *dice*) dijalankan pada lingkungan yang sama menggunakan fungsi *groupby* dan *pivot_table* pada pustaka *Pandas*

2.2.5 Dashboard Analitik

Dashboard analitik adalah antarmuka visual yang menyajikan informasi penting secara ringkas dan terorganisir dalam satu tampilan,

biasanya menggunakan kombinasi grafik, tabel, dan indikator kinerja utama (*KPI*)[11]. Tujuan utama *Dashboard* adalah memberikan akses cepat terhadap informasi sehingga pengguna dapat memahami kondisi data dalam waktu singkat tanpa harus melakukan analisis manual[10], [11].

Dashboard yang efektif memenuhi prinsip-prinsip seperti kesederhanaan (*simplicity*), keterbacaan (*readability*), interaktivitas (*interactivity*), dan relevansi terhadap kebutuhan pengguna. Dalam penelitian ini, *Dashboard* publik dirancang untuk menyajikan informasi tren penjualan otomotif yang relevan bagi tiga kelompok pengguna utama: masyarakat umum, akademisi/peneliti, dan pelaku usaha kecil-menengah. Penelitian terdahulu oleh Gloria dkk. (2025) dan Asy dkk. (2025) menunjukkan bahwa *Dashboard* berbasis web meningkatkan aksesibilitas informasi penjualan kepada pengguna non-teknis[11], [13].

2.2.6 Industri Otomotif Indonesia dan GAIKINDO

Gabungan Industri Kendaraan Bermotor Indonesia (GAIKINDO) adalah asosiasi resmi yang menaungi Agen Tunggal Pemegang Merek (ATPM) kendaraan bermotor roda empat atau lebih di Indonesia, didirikan pada tahun 1969. GAIKINDO berperan dalam mengumpulkan, mengkompilasi, dan mempublikasikan data statistik penjualan otomotif nasional yang menjadi rujukan resmi industri otomotif Indonesia[6].

Data yang dipublikasikan GAIKINDO mencakup lima kategori utama: (1) *Wholesales* penjualan dari ATPM kepada dealer; (2) *Retail sales* penjualan dari dealer kepada konsumen akhir; (3) Produksi *Volume* produksi domestik; (4) Ekspor *CBU* (*Completely Built-Up*) ekspor kendaraan utuh; dan (5) Impor *CBU* impor kendaraan utuh 3. Data ini tersedia per merek dan per periode (bulanan dan tahunan), mencakup lebih

dari lima puluh merek kendaraan bermotor yang beroperasi di Indonesia[12].

Periode 2020 hingga 2025 menjadi periode penting dalam sejarah industri otomotif Indonesia karena mencakup dinamika pasca-pandemi serta munculnya *Transformasi* struktural menuju kendaraan listrik (*Electric Vehicle/EV*) dengan masuknya merek-merek baru dari Tiongkok seperti BYD, Wuling EV, dan AION[6], [18]. Kondisi ini menjadikan data periode tersebut sangat kaya untuk dianalisis[18].

2.3 Framework yang Digunakan: Hibrida SDLC-Waterfall, CRISP-DM, dan OLAP

Penelitian ini menggunakan kerangka kerja hibrida (*hybrid framework*) yang memadukan tiga komponen sesuai arahan dewan penguji. Pertama, *System Development Life Cycle (SDLC)* model *Waterfall* sebagai kerangka siklus pengembangan dan deployment sistem. Kedua, *CRISP-DM (Cross-Industry Standard Process for Data Mining)* sebagai kerangka proses analitik data, mulai dari *Business Understanding*, *Data Understanding (EDA)*, *Data Preparation (ETL)*, *Modeling*, *Evaluation*, hingga *Deployment*. Ketiga, *Online Analytical Processing (OLAP)* sebagai teknik analitik deskriptif yang diterapkan pada fase *Modeling* untuk melakukan analisis multidimensi (*roll-up*, *drill-down*, *slice*, *dice*) atas data warehouse berskema star.

Perlu ditegaskan bahwa OLAP bukan merupakan Data Mining. OLAP adalah analisis multidimensi deskriptif atas *data warehouse*, sedangkan *Data Mining* mencakup penemuan pola otomatis seperti clustering, classification, prediction, forecasting, dan association rule mining. Karena penelitian ini hanya menerapkan ETL, integrasi data, data warehouse, dan analisis OLAP, penelitian ini

diposisikan sebagai Business Intelligence dengan descriptive analytics, bukan penelitian data mining, dan OLAP ditempatkan sebagai teknik analisis pada fase *Modeling CRISP-DM*.

Pemilihan kerangka hibrida ini didasarkan pada karakteristik penelitian yang mencakup pengembangan sistem sekaligus proses analitik data. CRISP-DM dipilih sebagai kerangka proses karena bersifat data-centric dan mencakup seluruh siklus dari pemahaman bisnis hingga deployment, sehingga sesuai untuk penelitian yang diselesaikan sampai tahap publikasi *dashboard*. SDLC-Waterfall menstrukturkan pengembangan sistem dan dashboard, sementara OLAP menjadi teknik analisis yang menghasilkan pengetahuan deskriptif dari data warehouse.

CRISP-DM terdiri dari enam fase utama yang saling berkaitan sebagaimana diilustrasikan pada Gambar 2.4: *Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment*.



Gambar 2.4 Enam Fase Framework CRISP-DM (Cross-Industry Standard Process for Data Mining)

1. Business Understanding (Pemahaman Bisnis)

Fase awal untuk memahami tujuan dan kebutuhan analisis dari sudut pandang pengguna. Dalam penelitian ini, fase ini mencakup analisa proses bisnis penyajian data GAIKINDO, identifikasi kesenjangan keterpakaian data (*data usability gap*), serta perumusan kebutuhan tiga kelompok pengguna sasaran (masyarakat umum, mahasiswa, dan pelaku usaha).

2. Data Understanding (Pemahaman Data)

Fase eksplorasi data awal (*exploratory data analysis*) untuk mengenali karakteristik dan kualitas data. Dalam penelitian ini, fase ini meliputi pengunduhan dan identifikasi 38 dokumen GAIKINDO 2020 hingga 2025 dengan lebih dari 100 sheet, pemetaan lima kategori data, serta identifikasi permasalahan kualitas data seperti inkonsistensi penamaan merek.

3. Data Preparation (Persiapan Data)

Fase pembersihan dan transformasi data melalui proses ETL: penanganan missing value dan duplikasi, standarisasi penamaan merek, penggabungan data multi-sheet, serta pembentukan kolom turunan seperti pertumbuhan *year-over-year* (YoY) dan pangsa pasar (*market share*), menghasilkan satu dataset terintegrasi.

4. Modeling (Pemodelan)

Dalam konteks Business Intelligence deskriptif, fase modeling diwujudkan melalui perancangan data warehouse berskema star (satu tabel fakta dan tiga tabel dimensi) serta penerapan operasi OLAP (*roll-up, drill-down, slice, dice*) sebagai teknik analisis multidimensi untuk meringkas dan membandingkan data historis.

5. Evaluation (Evaluasi)

Fase pengujian untuk memastikan hasil analisis dan sistem memenuhi kebutuhan. Dalam penelitian ini, evaluasi mencakup verifikasi akurasi data

warehouse terhadap sumber, evaluasi fungsional dan performa operasi OLAP, pengujian *black-box*, serta evaluasi kegunaan dashboard.

6. Deployment (Penerapan)

Fase akhir berupa publikasi hasil kepada pengguna. Dalam penelitian ini, hasil analisis disajikan melalui dashboard publik berbasis web yang di deploy pada *GitHub Pages* sehingga dapat diakses secara terbuka.

Selain pertimbangan di atas, pemilihan *CRISP-DM* sebagai kerangka proses didasarkan pada tiga pertimbangan utama. Pertama, *CRISP-DM* bersifat *data-centric* sehingga sesuai untuk penelitian yang berfokus pada pengolahan dan pemanfaatan data[7]. Kedua, *CRISP-DM* bersifat fleksibel dan dapat dipadukan dengan metodologi pengembangan perangkat lunak seperti SDLC-Waterfall sehingga cocok untuk proyek BI yang melibatkan pengolahan data berskala besar [19]. Ketiga, *CRISP-DM* telah digunakan secara luas dalam penelitian akademik di bidang analitik data sehingga memiliki landasan teoritis yang kuat[9].

2.4 Tools dan Software yang Digunakan

2.4.1 Jupyter Notebook dan Python

Jupyter Notebook adalah lingkungan komputasi interaktif berbasis web yang memungkinkan pengguna membuat dan membagikan dokumen yang berisi kode program, persamaan matematika, visualisasi, dan teks naratif dalam satu file [16], [17]. *Jupyter Notebook* umum digunakan dalam dunia data *science* karena memudahkan eksplorasi data secara iteratif dan dokumentasi proses analisis[9], [16].

Python adalah bahasa pemrograman tingkat tinggi yang menjadi standar *de facto* dalam pengolahan data dan *Machine Learning* karena memiliki sintaksis yang sederhana, komunitas yang besar, dan ekosistem pustaka yang sangat lengkap[16], [20]. Dalam penelitian ini, *Python*

digunakan sebagai bahasa utama pengolahan data dengan dukungan pustaka *Pandas* dan *NumPy*.

2.4.2 Pustaka *Pandas* dan *NumPy*

Pandas adalah pustaka *Python* yang menyediakan struktur data [16], [17] untuk manipulasi data tabular berkecepatan tinggi. *Pandas* memiliki fungsi-fungsi seperti *read_excel*, *read_csv*, *groupby*, *merge*, dan *pivot_table* yang sangat berguna untuk implementasi operasi *OLAP* dalam penelitian ini.

NumPy adalah pustaka komputasi numerik yang menjadi fondasi ekosistem ilmu data *Python* [16]. *NumPy* menyediakan *array multi-dimensi* yang efisien serta fungsi-fungsi matematika dan statistika yang menjadi dasar pustaka analitik lainnya termasuk *Pandas*.

2.4.3 *Microsoft Excel*

Microsoft Excel digunakan sebagai aplikasi awal untuk membuka, memvalidasi, dan memeriksa kualitas data mentah GAIKINDO sebelum diunggah ke lingkungan *Jupyter Notebook*. *Excel* juga digunakan untuk dokumentasi struktur data dan pembuatan ringkasan awal sebelum proses *ETL* otomatis dilakukan [9].

2.4.4 *HTML, CSS, JavaScript, dan Chart.js*

Untuk pembangunan *Dashboard* publik, digunakan kombinasi teknologi web standar berupa *HTML* untuk struktur, *CSS* untuk *styling*, dan *JavaScript* untuk interaktivitas [11]. Pustaka *Chart.js* digunakan untuk pembuatan grafik interaktif yang ditampilkan pada *Dashboard*. Pendekatan ini dipilih karena bersifat ringan, tidak memerlukan *server* back-end, dan dapat di-deploy sebagai web statis [10], [11].

2.4.5 *GitHub* dan *GitHub Pages*

GitHub digunakan sebagai sistem *version control* dan repository code *Dashboard*[9], [21]. *GitHub Pages* adalah layanan hosting web statis gratis yang disediakan *GitHub* untuk menyajikan halaman web langsung dari repository code[22]. Pemilihan *GitHub Pages* didasarkan pada tiga pertimbangan: (1) bersifat gratis dan tidak memerlukan biaya *hosting*; (2) dapat diakses publik tanpa otentikasi; dan (3) terintegrasi dengan *workflow* pengembangan code modern.

Kombinasi *tools* dan pustaka di atas dipilih karena bersifat sumber terbuka (*open source*), gratis, dan banyak digunakan dalam komunitas data science global. Hal ini juga sejalan dengan tujuan penelitian ini, yaitu memberikan kontribusi praktis berupa sistem yang dapat direplikasi oleh akademisi dan praktisi lain di Indonesia.