Klasifikasi Tipe dan Format Data dalam Konteks Big Data dan Data Science

Fadhlan Taufiqurrohman[1], Siti Nurul Zakiyyah[2], M. Naufal Rona[3], Muhamad Gilang Rigan Agasi[4], Muhamad Nuril Hikmana Kosim[5], Muhammad Encep[6]

[1]Universitas Djuanda, ftblues30@gmail.com
[2]Universitas Djuanda, nurulzakiyyah35@gmail.com
[3]Universitas Djuanda, muhamadnaufal1213@gmail.com
[4]Universitas Djuanda, m.gilank123@gmail.com
[5]Universitas Djuanda, nurilhikmana@gmail.com
[6]Universitas Djuanda, ahmadpoac@unida.ac.id

ABSTRAK

Perkembangan teknologi informasi yang pesat telah menyebabkan ledakan data dalam berbagai bentuk dan ukuran, yang dikenal sebagai fenomena Big Data. Dalam konteks ini, Data Science menjadi pendekatan utama untuk mengolah dan menganalisis data, namun keberagaman tipe dan format data masih menjadi tantangan utama. Meskipun telah banyak penelitian yang membahas pengolahan data secara teknis, kajian komprehensif yang mengklasifikasikan tipe dan format data dalam perspektif Big Data dan Data Science masih terbatas. Oleh karena itu, penelitian ini bertujuan untuk mengidentifikasi dan mengklasifikasikan tipe serta format data yang sering digunakan dalam bidang tersebut. Penelitian ini menggunakan pendekatan studi literatur (library research) terhadap berbagai jurnal ilmiah dan referensi akademik lainnya. Hasil penelitian menunjukkan bahwa data dalam konteks Big Data dapat diklasifikasikan ke dalam tiga tipe utama, yaitu data terstruktur, semi-terstruktur, dan tidak terstruktur. Selain itu, ditemukan berbagai format data yang digunakan sesuai kebutuhan penyimpanan dan pemrosesan, seperti CSV, JSON, XML, Avro, Parquet, dan lainnya. Pemilihan format ini berpengaruh langsung terhadap efisiensi proses analitik, interoperabilitas sistem, dan pengelolaan data skala besar. Penelitian ini menyadari adanya batasan berupa keterbatasan cakupan literatur serta belum adanya validasi empiris dalam praktik industri. Namun demikian, hasil kajian ini memberikan kontribusi awal bagi pengembangan kerangka kerja klasifikasi data serta menjadi referensi konseptual bagi peneliti dan praktisi. Implikasi penelitian ini juga membuka peluang

eksplorasi lebih lanjut terhadap pemanfaatan dan optimalisasi format data dalam berbagai sektor berbasis data.

Kata Kunci: Big Data, Data Science, Tipe Data, Format Data, Klasifikasi

PENDAHULUAN

Big Data adalah istilah yang diberikan pada kumpulan data yang berukuran sangat besar dan kompleks, sehingga tidak memungkinkan untuk diproses menggunakan perangkat pengelola database konvensional ataupun aplikasi pemroses data lainnya (Fajriyah et al., 2022). Revolusi digital yang ditandai dengan peningkatan kapasitas penyimpanan, kecepatan pemrosesan, dan konektivitas jaringan telah memunculkan era Big Data, di mana volume, kecepatan, dan variasi data meningkat secara eksponensial. Pemanfaatan big data di berbagai industri telah membawa tantangan baru dalam analisis, pengelolaan, dan pemrosesan data yang besar dan kompleks (Juroihan et al., 2024). Seiring dengan perkembangan ini, Data Science hadir sebagai pendekatan interdisipliner yang menggabungkan statistik, ilmu komputer, dan domain ilmu pengetahuan lainnya untuk mengekstraksi pengetahuan dari data. Menurut Syamsu & Widodo (2021), Data Science berfungsi untuk mengolah data yang sangat besar, baik terstruktur, semi-terstruktur maupun tidak terstruktur, menjadi informasi yang bernilai. Big Data memiliki karakteristik volume, variety, dan velocity yang mengharuskan data dikelola dengan cara yang berbeda tergantung pada bentuk dan tipenya, baik itu terstruktur maupun tidak terstruktur (Maryanto, 2017). Namun, salah satu tantangan fundamental dalam penerapan Big Data dan Data Science adalah keragaman tipe dan format data yang harus dikelola dan dianalisis secara efektif. Tipe data yang berbeda—baik terstruktur, semi-terstruktur, maupun tidak terstruktur—memiliki karakteristik teknis dan kebutuhan penanganan yang berbeda. Di samping itu, berbagai format data seperti CSV, JSON, XML, hingga format kolumnar seperti Parquet dan Avro, digunakan untuk menyimpan dan

mentransfer data dalam berbagai sistem dan platform, yang memerlukan pemahaman mendalam terhadap fungsinya.

Meskipun telah banyak studi yang membahas alat dan teknik dalam pengolahan data, masih terdapat kekosongan dalam literatur mengenai klasifikasi sistematis terhadap tipe dan format data dalam konteks Big Data dan Data Science. Padahal, pemahaman mengenai klasifikasi ini sangat penting sebagai landasan dalam pengambilan keputusan teknis maupun strategis dalam proyek-proyek data. Penelitian ini bertujuan untuk menyusun klasifikasi komprehensif mengenai tipe dan format data yang umum digunakan dalam lingkungan Big Data dan Data Science. Dengan menggunakan pendekatan studi pustaka terhadap berbagai sumber ilmiah, penelitian ini berupaya memberikan pemetaan konseptual yang dapat digunakan oleh peneliti dan praktisi sebagai referensi awal dalam memahami dan memilih tipe serta format data yang tepat sesuai kebutuhan analitik dan pengelolaan data skala besar.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan kualitatif deskriptif dengan metode studi literatur (library research) yang bertujuan untuk mengkaji berbagai sumber akademik terkait tipe dan format data dalam konteks Big Data dan Data Science. Pendekatan ini dipilih untuk menggali dan menyusun pemahaman teoritis mengenai klasifikasi data yang telah dikembangkan dalam berbagai penelitian sebelumnya. Pengumpulan data dilakukan dengan cara menelaah pustaka dari jurnal ilmiah, buku referensi, dan artikel akademik yang memiliki relevansi dengan topik penelitian. Untuk menghasilkan artikel dengan relevansi dan kemutakhiran yang tinggi, referensi yang digunakan merupakan artikel dengan maksimal penerbitan 5 tahun terakhir (Encep et al., 2024).

Pemilihan literatur didasarkan pada kesesuaian tema, kejelasan konsep, dan keterbaruan informasi dalam sumber-sumber tersebut. Data yang diperoleh dari literatur kemudian dikelompokkan berdasarkan topik pembahasan seperti tipe data, format data, serta hubungannya dengan kebutuhan penyimpanan dan analisis. Selanjutnya, informasi tersebut dianalisis dan disusun secara sistematis untuk membentuk kerangka klasifikasi yang komprehensif. Dengan metode ini, penelitian diharapkan dapat memberikan kontribusi teoritis yang kuat dan menjadi dasar untuk eksplorasi lebih lanjut dalam pengembangan sistem data berbasis Big Data dan Data Science.

HASIL DAN PEMBAHASAN

Hasil dari analisis literatur menunjukkan bahwa data dalam konteks Big Data dan Data Science dapat diklasifikasikan secara umum ke dalam tiga tipe utama, yaitu data terstruktur, data semi-terstruktur, dan data tidak terstruktur. Data terstruktur merupakan kelompok data yang memiliki tipe data, format dan struktur yang telah terdefinisi. Sumber datanya dapat berupa data transaksional, OLAP data, tradisional RDBMS, file CSV dan spreadsheets (Saputra et al., 2022). Data semi-terstruktur, seperti XML dan JSON, memiliki elemen-elemen yang memungkinkan organisasi data tetapi tidak sepenuhnya terikat pada skema yang kaku. Tipe ini sering digunakan dalam aplikasi web, pertukaran data antar sistem, dan representasi data yang fleksibel. Sementara itu, data tidak terstruktur mencakup data dalam bentuk teks bebas, gambar, audio, video, dan sensor, yang tidak memiliki format standar dan sering kali memerlukan pendekatan analitik khusus seperti pemrosesan bahasa alami (NLP), computer vision, atau analisis sinyal.

Selain klasifikasi berdasarkan tipe, penelitian ini juga mengidentifikasi berbagai format data yang umum digunakan dalam proses penyimpanan dan pemrosesan data skala besar. Format seperti CSV (Comma Separated Values) banyak

digunakan untuk data tabular sederhana dan sangat kompatibel dengan berbagai alat analisis data. JSON (JavaScript Object Notation) dan XML (eXtensible Markup Language) populer dalam pertukaran data antar sistem karena mendukung representasi data kompleks secara hierarkis. Dalam konteks Big Data yang menuntut efisiensi dalam penyimpanan dan pemrosesan, format seperti Avro, Parquet, dan ORC banyak digunakan karena mendukung kompresi data, penyimpanan kolumnar, dan performa baca/tulis yang tinggi. Format-format ini sangat relevan dalam platform big data seperti Apache Hadoop dan Apache Spark. Temuan ini menunjukkan bahwa pemilihan format data memiliki pengaruh langsung terhadap kecepatan pemrosesan, konsumsi memori, skalabilitas sistem, serta interoperabilitas antar platform analitik.

Dalam pembahasan lebih lanjut, klasifikasi tipe dan format data ini menunjukkan adanya keterkaitan erat dengan kebutuhan spesifik dari masing-masing proyek data. Misalnya, analisis real-time seperti pada data IoT atau media sosial memerlukan format yang ringan dan mendukung stream processing seperti JSON atau format biner ringan seperti Avro. Sebaliknya, untuk kebutuhan analisis historis pada volume besar, format seperti Parquet lebih efisien digunakan karena mendukung pemrosesan batch yang masif. Selain itu, format penyimpanan juga memengaruhi kompleksitas pipeline data, termasuk dalam tahap ekstraksi, transformasi, dan loading (ETL). Oleh karena itu, pemahaman terhadap klasifikasi ini tidak hanya bersifat teknis, tetapi juga strategis dalam perencanaan arsitektur data dan pengembangan sistem berbasis big data.

KESIMPULAN

Penelitian ini mengkaji dan mengklasifikasikan tipe serta format data yang umum digunakan dalam konteks Big Data dan Data Science melalui pendekatan studi literatur. Hasil kajian menunjukkan bahwa data dalam lingkungan Big Data dapat dikelompokkan menjadi tiga tipe utama, yaitu data terstruktur, semi-terstruktur, dan

tidak terstruktur, masing-masing dengan karakteristik teknis dan implikasi analitik yang berbeda. Selain itu, ditemukan bahwa berbagai format data seperti CSV, JSON, XML, Avro, Parquet, dan ORC memiliki peran penting dalam menyimpan dan mengelola data sesuai dengan kebutuhan spesifik, baik untuk pemrosesan real-time maupun batch, serta dalam konteks efisiensi penyimpanan dan performa analitik.

Klasifikasi yang disusun dalam penelitian ini memberikan kontribusi konseptual yang dapat dijadikan acuan awal dalam proses pemilihan dan perancangan sistem data, terutama dalam proyek-proyek yang memerlukan penanganan data dalam skala besar dan kompleks. Meskipun demikian, penelitian ini memiliki beberapa batasan, antara lain tidak dilakukannya validasi langsung terhadap praktik industri dan fokus hanya pada studi literatur tanpa pengumpulan data lapangan. Oleh karena itu, disarankan agar penelitian selanjutnya memperluas cakupan dengan pendekatan empiris, seperti studi kasus atau survei terhadap praktisi data, guna menguji kepraktisan dan relevansi klasifikasi yang diusulkan.

REFERENSI

Saputra, A., Firdaus, M. I., Wahyudi, R., Mohdo, L., Gunawan, M. E., Encep, M., & Khaira, M. (2022). Big data. *Karimah Tauhid*, *1*(6), 880-889.

Juroihan, M., Fikri, W. K., Mohdo, L., Fikri, M., Romadhon, R. N., & Encep, M. (2024). Integrasi Cloud Computing untuk Analisis Big Data. *Karimah Tauhid*, *3*(4), 4387-4399.

Encep, M., Rianto, M. R., Faris, B. A., Mutahari, M. I., & Rahman, R. A. (2024). Manfaat Implementasi Big Data pada Berbagai Sektor. *Karimah Tauhid*, *3*(8), 8957-8968.

Fajriyah, N., Setiawan, W., Dewi, E., & Duha, T. (2022). Implementasi Teknologi Big Data di Era Digital. *Jurnal Informatika*, *1*(1), 1-7.

Maryanto, B. (2017). Big Data dan Pemanfaatannya dalam Berbagai Sektor. *Media Informatika*, *16*(2), 14-19.

Syamsu, M., & Widodo, W. (2021). Peran Data Science dan Data Scientist Untuk Mentransformasi Data Dalam Industri 4.0. *Jurnal Teknologi Informasi (JUTECH)*, *2*(1), 27-36.