A SYSTEMATIC LITERATURE REVIEW OF FUNDAMENTAL DATA SCIENCE CONCEPTS

[1]Akila Rianti, [2]Danu Alpin Gustio, [3]M Rizqy Fadlan, [4]Nadia Nadawina

[1,2,3,4]*Sekolah Tinggi Keguruan Dan Ilmu Pendidikan Amal Bakti*
Email: [1]Akilarianti788@email.com, [2]danualpingustio@gmsil.com, [3]riskyfadlan564@gmail.com, [4] nadianadawina273@gmail.com

ABSTRACT

Data science is an interdisciplinary field that integrates multiple domains of knowledge and plays a critical role in transforming raw data into meaningful information to support evidence-based decision making. The rapid advancement of digital technology has led to a significant increase in both the volume and complexity of data, thereby necessitating systematic and high-quality data management approaches. This study aims to examine the fundamental concepts of data science and its key stages, including data collection, data cleaning, data exploration and analysis, data modeling, evaluation, and data visualization. This research adopts a literature review method by analyzing 20 scientific articles relevant to the foundational principles of data science. The findings indicate that the stages of the data science process are highly interconnected, and data science should be viewed not merely as a technical tool but as a robust foundation for innovation and advancement across various aspects of life. High-quality data that undergo rigorous analytical processes can generate deeper insights and support more effective and efficient decision making.

Keywords : *Data Science, Data Collection, Data Cleaning, Data Analysis, Data Visualization*

INTRODUCTION

Data science is a discipline that applies scientific approaches to collect data from diverse sources and subsequently process and clean it to eliminate errors. This process aims to enhance understanding of real-world phenomena, generate predictions about future trends, and support more accurate, evidence-based decision making across various sectors, including business, healthcare, and the environment (Athoillah & Fitriani, 2024).

The data science process represents a skill-oriented framework that emphasizes learning through practical and systematic stages. This process encourages individuals to independently explore and deepen their understanding of facts, concepts, and values essential to research activities. Through this approach, researchers are not merely passive recipients of information but are guided to develop creative and analytical thinking patterns that broaden scientific insight. Consequently, data science functions not only as a tool for comprehension but also as a means of fostering innovative and forward-looking mindsets (Setyo Retno, 2022).

From multiple perspectives, science is regarded as a methodological process for acquiring new knowledge. It involves systematic observation and data analysis through experimentation to construct an understanding of natural and social phenomena. This approach provides a reliable foundation for producing factual evidence and promotes the advancement of science in an ethical and responsible manner (Izzuddin Ahmad, 2024).

Data reflect various aspects of human life and enable a comprehensive understanding of behavioral and social patterns. In this context, data are not merely numerical values or isolated information but serve as a robust foundation for interpreting how individuals and societies function. When managed and analyzed appropriately, data facilitate more intelligent decision making and ultimately become a powerful instrument for representing real-world conditions (Nasution, 2021).

In the current era of rapid digital technological advancement, data are increasingly recognized as a valuable resource that plays a strategic role in enhancing efficiency and personalization across systems and services (Agustiana et al., 2023). Fundamentally, data science begins with data collection, which refers to the process of obtaining information from various sources, including both primary and secondary data. Primary data are collected directly from original sources through methods such as interviews, surveys, or experiments, whereas secondary data are derived from existing materials, including previously published articles, books, and academic journals (Sulung & Muspawi, 2024). This stage is critical, as high-quality data form the basis for subsequent analytical processes in data science.

One of the primary challenges in the field of data science is poor data quality, which may manifest in the form of incomplete records, inaccuracies, or duplicate entries. Such issues can compromise analytical outcomes and reduce the accuracy of data-driven models. As data availability continues to expand alongside rapid technological progress, the role of data science becomes increasingly vital in addressing these challenges and generating innovative solutions to complex problems (Sulianta, 2024).

Given the wide range of challenges present in data science, the main objective of this study is to enhance data management and analytical practices to ensure that the data used are accurate, reliable, and fit for purpose. Through improved data processing, decision-making processes can become more informed and rational. This study is expected to contribute meaningful insights by

reinforcing the view of data science not merely as a technical instrument, but as a strong foundational framework for innovation and advancement across various fields of life.

METHOD

This study employed a library research approach. This method involved an in-depth exploration of various scholarly sources, including academic books, peer-reviewed journals, and research articles authored by previous scholars that specifically address the fundamental concepts of data science. The research process consisted of collecting and analyzing 20 scientific articles that focus on core data science principles.

The use of a library research methodology ensured that the processes of source identification, selection, and analysis were conducted in a structured and systematic manner. This approach provided a strong methodological foundation for the study, allowing for a comprehensive synthesis of existing knowledge. Consequently, the findings are expected to present well-organized and coherent information aligned with the research objectives, while facilitating an effective analytical process to optimally support the achievement of the study's aims..

RESULTS AND DISCUSSIONS

Data Collection

Data collection techniques refer to the methods employed by researchers to obtain the information required for a study. Appropriate data collection plays a crucial role in establishing a solid research foundation. When suitable techniques are carefully selected, the collected data tend to be accurate and relevant, thereby strengthening the reliability of subsequent analyses. Common techniques aligned with research objectives include surveys, interviews, and observations. Researchers are also required to ensure that data collection is conducted ethically, such as obtaining informed consent from respondents and protecting personal privacy to avoid legal and ethical violations (Nafisatur, 2024).

Data collection may involve multiple strategies, including direct observation to examine phenomena in real-world settings and interviews to gain in-depth insights through structured or semi-structured interactions with participants (Haki et al., 2024). These techniques are essential components of the research methodology. However, survey-based data are often susceptible to inaccuracies, such as recall bias, where respondents may misremember past events, potentially affecting data validity (Apriliawati, 2020).

Data Cleaning

Data cleaning is a critical step in enhancing data quality to ensure its readiness for analysis. This process involves identifying and correcting various data issues, such as duplicate records, inconsistencies, and incomplete entries, in order to improve data usability. Through effective data cleaning, the risk of errors in analytical reports can be significantly reduced, allowing datasets to be processed without major obstacles (Wijanarko & Santoso, 2024).

Furthermore, data cleaning ensures that raw data can be utilized optimally and in accordance with analytical objectives. Without this process, raw datasets often contain inaccuracies and missing information that may compromise analytical outcomes. By implementing systematic data cleaning procedures, each data element becomes valid, consistent, and analysis-ready, thereby supporting the successful achievement of research objectives (Hidayat et al., 2023).

Data Exploration and Analysis

Data exploration refers to the process of examining and presenting data in order to uncover meaningful patterns and insights embedded within the dataset. Although closely related to data cleaning, data exploration places greater emphasis on developing a deeper understanding of data characteristics. This stage serves as a fundamental basis for conducting accurate and efficient analyses without altering the core structure of the data (Muhammad Radhi, 2022).

Data analysis is conducted after all relevant data have been comprehensively collected. This stage is essential to prevent premature conclusions and to ensure analytical accuracy. Through systematic analysis, potential errors arising from incomplete or inconsistent data can be minimized, enabling the findings to support more informed and evidence-based decision-making processes (Millah et al., 2023).

Data Modeling

Data modeling represents a strategic approach to identifying and replicating specific patterns, such as structural relationships, semantic meanings, and interdependencies among data elements. The primary objective is to enable these patterns to be practically implemented and transferred into computational systems. As a result, data modeling facilitates more effective, accurate, and innovative analytical outcomes. This process involves iterative data collection and repeated evaluations to ensure model reliability and robustness (Sujacka Retno & Rozzi Kesuma Dinata, 2020).

Moreover, data modeling serves as a foundational mechanism for addressing various analytical challenges. By establishing appropriate data structures, this approach enables the identification of patterns, outcome prediction, and informed decision-making. In technological applications, data modeling simplifies informational complexity, thereby supporting the development of innovative, efficient, and effective solutions. Ultimately, data modeling provides a strong framework for overcoming practical barriers in both data analysis and research contexts (Muhammad Riza Alifi et al., 2022).

Evaluation

Evaluation aims to collect essential data, process it comprehensively, and present the results in an interpretable format. The analyzed outcomes are subsequently examined and compared against predefined evaluation criteria. The final results serve as a basis for making informed decisions regarding the evaluated object or system. This process ensures that decisions are grounded in objective and verifiable evidence (Gusmaningsih et al., 2023).

In addition, evaluation provides concrete evidence regarding whether the initial objectives have been successfully achieved and whether the implemented actions were appropriate and effective. Through systematic evaluation, decision-makers are able to refine strategies and make necessary adjustments to improve future outcomes, ensuring that efforts remain purposeful and impactful (Evita Sari Dalimunthe & Muhammad Syahbudi, 2023).

Data Visualization

Data visualization can be broadly defined as the process of making data easier to understand by transforming raw information into graphical representations such as charts, graphs, and diagrams that highlight key insights. The primary objective is to ensure that data are not perceived merely as static numbers, but rather as meaningful information that conveys clear and actionable insights to diverse audiences (Ariandi & Rahma Puteri, 2022).

More specifically, data visualization involves converting complex datasets into visual formats that facilitate the identification of patterns, trends, and critical insights. Without

visualization, data often remain difficult to interpret due to their abstract and numerical nature. In the contemporary digital era, data visualization plays an increasingly vital role by transforming ambiguous data into a powerful medium that supports understanding, inspires action, and drives informed change (Hartono et al., 2025).

Table 1. Evaluation of Reference Articles (Literature)

No	Author(s)	Year	Article Title	Key Conclusions
1	Athoillah & Fitriani	2024	Scientific Writing Training in the Field of Data Science for Senior High School (MAN 2 Nganjuk) Teachers	Structured training activities, including observation, material delivery, and mentoring, produced positive outcomes, with high participant satisfaction due to the relevance of materials and instructional methods.
2	Agustiana et al.	2024	Data Science Learning in Elementary Schools and Its Impact on Students' Mathematical Literacy	Data science-based learning enhanced elementary students' conceptual understanding, critical thinking, and problem-solving skills through child-friendly media and teacher training aligned with modern education.
3	Sulung & Muspawi	2024	Understanding Research Data Sources: Primary, Secondary, and Tertiary Data	Understanding distinctions among primary, secondary, and tertiary data is essential, as data source selection directly affects research accuracy and validity; integrating these sources improves research quality and relevance.
4	Sulianta	2024	Fundamentals and Concepts of Data Science	The book serves as a comprehensive and accessible guide for understanding core concepts and foundational principles of data science.
5	Nafisatur	2024	Research Data Collection Methods	The study emphasizes that data collection methods must be aligned with research objectives to obtain accurate and meaningful information for analysis and decision-making.
6	Apriliawati	2020	Diary Study as a Data Collection Method in Quantitative Research: A Literature Review	Diary studies are shown to be an effective quantitative data collection method that addresses limitations commonly found in surveys and experimental approaches.
7	Wijanarko & Santoso	2024	Application of MID and FIND Functions for Address Data Cleaning	The use of MID and FIND functions in Microsoft Excel effectively improves data cleanliness, organization, and

No	Author(s)	Year	Article Title	Key Conclusions
				readability, facilitating easier data retrieval and analysis.
8	Hidayat et al.	2023	Exploratory Data Analysis of Key Performance Indicator Achievements in the Faculty of Engineering	The study recommends focusing on lecturer development, particularly for those with limited external engagement, as a strategic foundation for improving faculty performance indicators.
9	Muhammad Radhi	2022	Big Data Analysis Using Exploratory Data Analysis (EDA) and Visualization with Jupyter Notebook	The findings demonstrate that EDA combined with visualization using Jupyter Notebook is effective in extracting meaningful insights from sales data.
10	Millah et al.	2023	Data Analysis in Classroom Action Research	Continuous and interconnected data analysis is essential to ensure accurate conclusions and to support evidence-based improvements in learning quality.
11	Sujacka Retno & Rozzi Kesuma Dinata	2020	Development of an Information Chatbot System	The study successfully developed and implemented a chatbot (Lhokbot), demonstrating its potential as an effective information source for both tourists and local communities.
12	Muhammad Riza Alifi et al.	2022	Relational Data Modeling in Document-Oriented NoSQL Databases	Relational data modeling can be effectively applied in document-oriented NoSQL databases using Chen's ERD notation at the conceptual level and CDR at logical and physical levels.
13	Gusmaningsih et al.	2023	Reflection and Evaluation Strategies in Classroom Action Research	Reflection is used to identify strengths and weaknesses in implemented learning actions, while evaluation measures success by comparing outcomes with predefined criteria.
14	Evita Sari Dalimunthe & Muhammad Syahbudi	2023	Data-Based Evaluation in Educational Policy	Evaluation outcomes serve as a primary foundation for data-driven decision-making, supporting fair educational policies and adaptive education systems.
15	Ariandi & Rahma Puteri	2022	Data Visualization Analysis of Kertapati District Using Tableau Public	Data visualization significantly accelerates analysis processes, improving accuracy and time efficiency compared to conventional analytical methods.

No	Author(s)	Year	Article Title	Key Conclusions
16	Hartono et al.	2025	Data Visualization Using Tableau: A Literature Review	Tableau is identified as an effective and efficient visualization platform suitable for both beginners and professionals, producing more communicative and insightful visual outputs.
17	Setyo Retno	2022	Analysis of Project-Based Learning Using Video Content in Teaching Basic Science Concepts	Project-based learning encourages active student engagement in real-world problem-solving while fostering critical thinking, collaboration, communication, and innovation skills.
18	Izzuddin Ahmad	2024	Improving Science Concept Understanding of Pre-Service Elementary Teachers through a Scientific Approach	The scientific approach actively engages students in observation, logical reasoning, and idea expression, resulting in improved comprehension of science concepts.
19	Nasution	2021	Understanding Data: An Introduction	Data cannot be fully understood through a single disciplinary perspective; interdisciplinary approaches are essential for comprehensive interpretation.
20	Haki et al.	2024	Data Collection and Analysis Strategies in Qualitative Educational Research	Techniques such as observation and interviews are effective in addressing complex issues in educational and social research contexts.

CONCLUSIONS

Based on the literature review, it can be concluded that data science plays a crucial role in transforming raw data into meaningful insights and serving as a foundation for evidence-based decision-making. Each stage of the data science process ranging from data collection, data cleaning, data exploration and analysis, data modeling, evaluation, to data visualization is interconnected and collectively determines the quality of the final outcomes. This study emphasizes that a solid understanding of fundamental data science concepts is essential to ensure that data can be optimally utilized as a basis for innovation and development across various fields. Consequently, data science should be viewed not merely as a technical tool, but as a strategic framework that supports sustainable progress and informed decision-making.

ACKNOWLEDGMENTS

The author would like to express sincere gratitude to all individuals and institutions that provided support and assistance throughout the completion of this study. Special appreciation is

extended to colleagues and mentors for their valuable insights, guidance, and constructive feedback. The author also acknowledges the educational institutions and scholarly literature that served as the intellectual foundation for this research. All forms of support, both direct and indirect, were instrumental in ensuring the smooth execution and successful completion of this study.

REFERENCES

Agustiana, A., Arisetyawan, A., & Anusaen, K. (2023). Pembelajaran Data Science di Sekolah Dasar Terhadap Kemampuan Literasi Matematika Siswa. *Jurnal Didaktika Pendidikan Dasar*, *7*(3), 891–910. https://doi.org/10.26811/didaktika.v7i3.1387

Apriliawati, D. (2020). Diary Study sebagai Metode Pengumpulan Data pada Riset Kuantitatif: Sebuah Literature Review. *Journal of Psychological Perspective*, *2*(2), 79–89. https://doi.org/10.47679/jopp.022.12200007

Ariandi, M., & Rahma Puteri, S. (2022). Analisis Visualisasi Data Kecamatan Kertapati menggunakan Tableau Public. *Jurnal Jupiter*, *14*(2), 366–373. http://repository.binadarma.ac.id/id/eprint/7199

Athoillah, M., & Fitriani, F. (2024). Pelatihan Karya Tulis Ilmiah Bidang Sains Data Bagi Guru Madrasah Aliyah Negeri (Man) 2 Nganjuk. *Kanigara*, *4*(1), 34–44. https://doi.org/10.36456/kanigara.v4i1.8535

Evita Sari Dalimunthe, K., & Muhammad Syahbudi. (2023). Jurnal mudabbir. *Jurnal Research and Education Studies*, *3*(1), 11–20.

Gusmaningsih, I. O., Azizah, N. L., Suciani, R. N., & Fajrin, R. A. (2023). Strategi Refleksi dan Evaluasi Penelitian Tindakan Kelas. *Jurnal Kreativitas Mahasiswa*, *1*(2), 114–123.

Haki, U., Prahastiwi, E. D., & Selatan, U. T. (2024). Strategi Pengumpulan dan Analisis Data dalam Penelitian Kualitatif Pendidikan. *Jurnal Inovasi Dan Teknologi Pendidikan*, *3*(1), 1–19. https://doi.org/10.46306/jurinotep.v3i1.67

Hartono, F., Somad, A., & Ezamora, I. (2025). *Visualisasi Data dengan Tableau : Sebuah Literature Review Presentasi , penelitian , dan pengamatan merupakan kegiatan yang sudah sangat umum di lakukan di kehidupan sehari hari . Mulai dari hal kecil seperti seorang anak yang mengamati kejadian di sekita*. *2025*, 1–6. https://doi.org/10.65344/practic.v1i1.125

Hidayat, I., Tolago, A. I., Dako, R. D. R., & Ilham, J. (2023). Analisis Data Eksploratif Capaian Indikator Kinerja Utama 3 Fakultas Teknik. *Jambura Journal of Electrical and Electronics Engineering*, *5*(2), 185–191. https://doi.org/10.37905/jjeee.v5i2.18397

Izzuddin Ahmad. (2024). Peningkatan Konsep Sains Mahasiswa Calon Guru SD/MI Pada Mata Kuliah Konsep Dasar IPA Menggunakan Pendekatan Saintifik. *NUSANTARA: Jurnal Pendidikan Dan Ilmu Sosial*, *6*, 34–46.

Millah, A. S., Apriyani, Arobiah, D., Febriani, E. S., & Ramdhani, E. (2023). Analisis Data dalam Penelitian Tindakan Kelas. *Jurnal Kreativitas Mahasiswa*, *1*(2), 140–153.

Muhammad Radhi, A. R. H. S. H. S. I. (2022). *ANALISIS BIG DATA DENGAN METODE EXPLORATORY DATA ANALYSIS (EDA) DAN METODE VISUALISASI MENGGUNAKAN JUPYTER NOTEBOOK | Jurnal Sistem Informasi dan Ilmu Komputer*. *4*(2), 2–6. https://jurnal.unprimdn.ac.id/index.php/JUSIKOM/article/view/2475

Muhammad Riza Alifi, Transmissia Semiawan, Djoko C.U. Lieharyani, & Hashri Hayati. (2022). Pemodelan Data Relasional pada NoSQL Berorientasi Dokumen. *Jurnal Nasional Teknik*

Elektro Dan Teknologi Informasi, *11*(3), 183–191. https://doi.org/10.22146/jnteti.v11i3.3704

Nafisatur, M. (2024). Metode Pengumpulan Data Penelitian. *Metode Pengumpulan Data Penelitian*, *3*(5), 5423–5443.

Nasution, M. K. M. (2021). Memahami Data: Suatu Pengantar. *Sains Data*, *2*(July), 3. https://www.researchgate.net/profile/Mahyuddin-Nasution/publication/352991330_Memahami_Data_Suatu_Pengantar/links/61079ed21e95fe241aa18822/Memahami-Data-Suatu-Pengantar.pdf

Setyo Retno, R. (2022). Analisis Penerapan Model Pembelajaran Project Based Learning Berbasis Content Video Pada Pembelajaran Konsep Dasar Sains Mahasiswa. *Jurnal Pemikiran Dan Pengembangan Sekolah Dasar (JP2SD)*, *10*(1), 1–11. https://doi.org/10.22219/jp2sd.v10i1.19850

Sujacka Retno, Rozzi Kesuma Dinata, N. H. (2020). Jurnal Computer Science and Information Technology (CoSciTech) manusia. *Jurnal Sains Dan Teknologi*, *3*(2), 1–11.

Sulianta, F. (2024). *Buku Dasar & Konsep Data Science*.

Sulung, U., & Muspawi, M. (2024). MEMAHAMI SUMBER DATA PENELITIAN : PRIMER, SEKUNDER, DAN TERSIER. *Jurnal Edu Reseach*, *5*(3), 110–116.

Wijanarko, S., & Santoso, A. (2024). *VOL. X NO. 1 FEBRUARI 2024 JURNAL TEKNIK INFORMATIKA STMIK ANTAR BANGSA Penerapan Fungsi Mid Dan Find pada Pembersihan Data Alamat*. *X*(1), 14–18.