Pertanika J. Sci. & Technol. 33 (3): 1231 - 1256 (2025)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Review Article

Thematic Trends on Data Quality Studies in Big Data Analytics: A Review

Nazliah Chikon, Shuzlina Abdul-Rahman* and Syaripah Ruzaini Syed Aris

College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, 40000 Shah Alam, Selangor, Malaysia

ABSTRACT

Data quality has become a critical issue in research and practice in the era of exponential data generation and increasing reliance on big data analytics (BDA) across industries. This study conducts a thematic analysis of literature published between 2020 and 2024 to examine the prevailing trends, challenges, and advancements in data quality studies within the domain of BDA. Guided by the systematic thematic review methodology, the research analysed 34 peer-reviewed studies identified from SCOPUS and Web of Science (WoS) databases, using qualitative data analysis tools such as ATLAS.ti. The findings reveal five major themes: Ontology and Data Quality Frameworks, Big Data Analytics in Various Industries, Machine Learning and AI Integration, Governance and Data Stewardship, and Tools and Techniques for Data Analysis. These themes highlight a shift towards interdisciplinary approaches, integrating advanced technologies like Artificial Intelligence (AI) and the Internet of Things (IoT) to address data quality issues. Limitations include potential selection bias from database restrictions and the exclusion of subscription-based journals, which may limit the generalisability of the findings. The study contributes to the theory by providing a comprehensive synthesis of data quality trends and their implications across various sectors. Methodologically, it demonstrates the utility of thematic analysis for consolidating diverse research. Practically, the insights inform data practitioners and policymakers on governance and technological strategies for ensuring data integrity. This review is original in its systematic exploration of thematic trends in data quality, offering a valuable roadmap for future research and addressing the critical intersection of data quality and BDA.

Keywords: Artificial intelligence, big data analytics, data analytics, data quality, governance

ARTICLE INFO

Article history:
Received: 12 August 2024
Accepted: 07 February 2025
Published: 26 March 2025

DOI: https://doi.org/10.47836/pjst.33.3.07

E-mail addresses:
nazliah.uitm@gmail.com (Nazliah Chikon)
shuzlina@uitm.edu.my (Shuzlina Abdul Rahman)
ruzaini@uitm.edu.my (Syaripah Ruzaini Syed Aris)
*Corresponding author

e-ISSN: 2231-8526

INTRODUCTION

In an era characterised by an exponential increase in data generation and computational capabilities, the role of data quality within Big Data Analytics (BDA) has become paramount. As various sectors, including shipping, facilities management, and healthcare, increasingly rely on big data to drive decision-making and operational efficiency, the integrity and reliability of this factor cannot be overstated. Barba-González et al. (2024) emphasised that data quality should be the cornerstone of Artificial Intelligence (AI) initiatives from the onset, where measurement and evaluation of data quality are crucial to determining its usability for specific tasks. This perspective is crucial as industries adopt digital transformations, venturing into domains like Shipping 4.0, where BDA serves as a disruptive force enhancing operational energy efficiency (Bui & Perera, 2021).

Moreover, integrating BDA with technologies such as the Internet of Things (IoT) is recognised as a strategic investment by firms aiming to differentiate themselves in competitive markets (Côrte-Real et al., 2020). This integration also highlights the importance of data quality from sensor input in leveraging business value from BDA investments. The quality of sensor data is crucial in BDA, as errors like missing values, outliers, and drift can lead to incorrect decisions. Teh et al. (2020) provide a comprehensive review of sensor data quality issues and solutions, highlighting the importance of addressing these errors to ensure accurate insights.

The multifaceted impact of data quality is also evident in its role in enhancing machine learning (ML) readiness for large-scale datasets (Hart et al., 2022) and in driving data-driven decision-making in marketing (Johnson et al., 2021). Medeiros et al. (2021) identify the increasing necessity for organisations to establish robust data quality policies and processes as regulatory demands grow and the scope of database analyses broadens across various business areas. As industries evolve, expectations from big data also shift, demanding not only the collection and storage of vast amounts of data but also ensuring that this data is of high quality to generate accurate and actionable insights. Therefore, it will be interesting to see the trends and patterns that emerge in research integrating data quality in BDA in the literature and the direction of future research. The purpose of this article is to conduct a thematic analysis of discussions on data quality in the domain of BDA that have been published between 2020 and 2024.

METHODS

This section describes the methodology used in this thematic review. The term "thematic review" and the use of ATLAS.ti as a tool to conduct thematic review were introduced by Zairul (2020; 2021), Zairul et al. (2023), and Zairul and Zaremohzzabieh (2023). Clarke and Braun (2013) defined thematic analysis as a systematic process of identifying patterns and developing themes from a meticulous reading of the subject matter's contents (Zairul,

2021). The study employs Thematic Review FlowZ (TreZ), which is protected by copyright under the registration number CRLY2023W02032 (Zairul, 2023). The method is employed because the study's methodology adheres to the thematic analysis procedure for conducting literature reviews. Figure 1 illustrates the thematic review flow in TreZ.

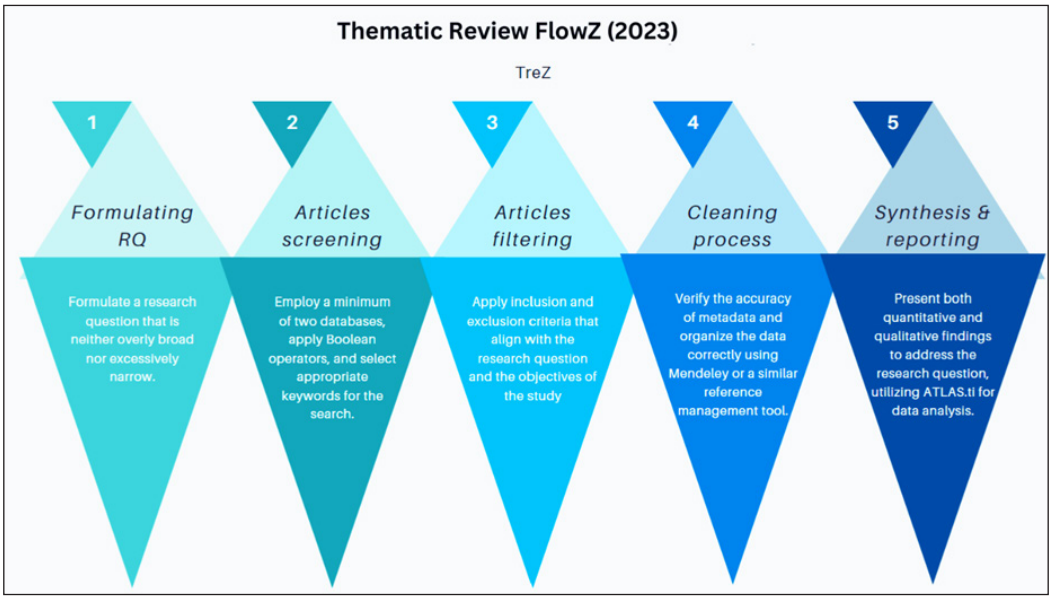

Figure 1. Thematic Review FlowZ (TreZ) (Zairul, 2023)

The TreZ outlines a systematic approach for conducting thematic literature reviews. The process begins with formulating a research question, defining the study's scope and focus, and ensuring the research is manageable and the conclusions are meaningful. The process followed with article screening using a minimum of two databases, Boolean operators and appropriate keywords for the search. Subsequently, articles are filtered using inclusion and exclusion criteria that are aligned with the research question and objective of the study. The fourth step involves a thorough double-checking of the metadata of the article and organising the data using reference management software like Mendeley. The final step is data extraction, where a thematic analysis is conducted using tools like ATLAS.ti to analyse the data and develop themes based on discussions of the subject matter in the selected articles (Figure 2). This methodology provides a structured framework for executing an exhaustive and effective literature review.

The next step involves identifying patterns and constructing themes to comprehend the trends related to data quality studies in BDA, as discussed in the literature from 2020 to 2024. To ensure a comprehensive and relevant analysis, the selection of literature for this review was guided by specific criteria: (1) the publication date range from 2020

to 2024, ensuring the research is current and significant, (2) focusing on open-access articles, ensuring unrestricted access to the full text of studies, (3) including only articles in their final stage of publication, ensuring the use of completed and validated research, and (4) inclusion of key terms such as "data quality" and "big data analytics", ensuring the focus on studies that are directly relevant to the main themes of this research. This methodical selection process helps capture a broad spectrum of contemporary insights and developments in this field.

Figure 2. Word cloud generated from all 34 articles

In the context of a thematic review (TreZ) utilising specific search strings, this research meticulously outlined and executed a selection process to identify pertinent literature across two major academic databases, SCOPUS and Web of Science (WoS). Here, we detail the methodological steps undertaken, ensuring a robust selection of studies that enhance the validity and reliability of this review's findings.

The search began with carefully formulating queries tailored to this study's objectives. These queries were deployed in the SCOPUS and WoS databases. The two databases are chosen for their comprehensive coverage of peer-reviewed journals relevant to data quality studies within the BDA domain. In SCOPUS, the search was defined with the keywords "data quality" and "big data analytics" in the title, abstract, and keywords (**TITLE-ABS-KEY**) targeting publications from 2020 to 2024, and restricted to academic articles in English that were open-access and in the final stage of publication (**LIMIT-TO (DOCTYPE, "ar")**, **LIMIT-TO (LANGUAGE, "English")**, **LIMIT-TO (OA, "all"),**

(LIMIT-TO (PUBSTAGE, "final")). This search strategy yielded 28 articles, indicating a substantial body of recent literature. Conversely, the search in WoS was smaller, using the same keywords across all fields without specific field restrictions, and focused only on open-access articles in English. This approach yielded 27 results. The difference in the number of articles retrieved from each database is primarily due to their varying coverage and focus and distinct indexing policies. Table 1 shows the search strings used, resulting in the initial search, which guarantees a thorough selection of studies aiming to encompass a broad spectrum of relevant literature.

Table 1
Search strings from Scopus and WoS

Source	Keywords	Results
SCOPUS	TITLE-ABS-KEY ("data quality" AND "big data analytics") AND PUBYEAR > 2019 AND PUBYEAR < 2025 AND (LIMIT-TO (DOCTYPE, "ar")) AND (LIMIT-TO (SRCTYPE, "j")) AND (LIMIT-TO (LANGUAGE, "English")) AND (LIMIT-TO (PUBSTAGE, "final")) AND (LIMIT-TO (OA, "all")).	28 results
Web of Science (WoS)	"data quality" AND "big data analytics" (Topic) and Open Access and 2024 or 2023 or 2022 or 2021 or 2020 (Publication Years) and Article (Document Types) and English (Languages).	27 results

Upon merging the search results from both databases, we proceeded to identify and remove duplicate entries to maintain the uniqueness of each record in subsequent analyses. A total of 15 duplicates were identified and excluded from the dataset. The consolidated list of records then underwent rigorous screening based on predefined inclusion and exclusion criteria. These criteria were meticulously developed to align closely with the research question and objective of this study: (1) the record must be related to the objective of this study, and (2) the record must provide empirical findings. Six records were excluded during this phase as they did not meet the necessary criteria, ensuring that only the most pertinent studies were retained.

After this thorough screening process, 34 studies were selected for inclusion in this thematic review (TreZ). These studies collectively met all specified eligibility requirements and are expected to provide substantial insights pertinent to the research question. The systematic approach to selecting relevant papers highlights the diligence required to conduct a thorough review. This selection process ensures the inclusion of relevant data and minimises biases, contributing significantly to the reliability of the review's conclusions. This report serves as a foundational component of this review paper, providing clarity and transparency about the methods used in study selection, which is critical for replicability and trust in the findings presented. This process is illustrated in Figure 3.

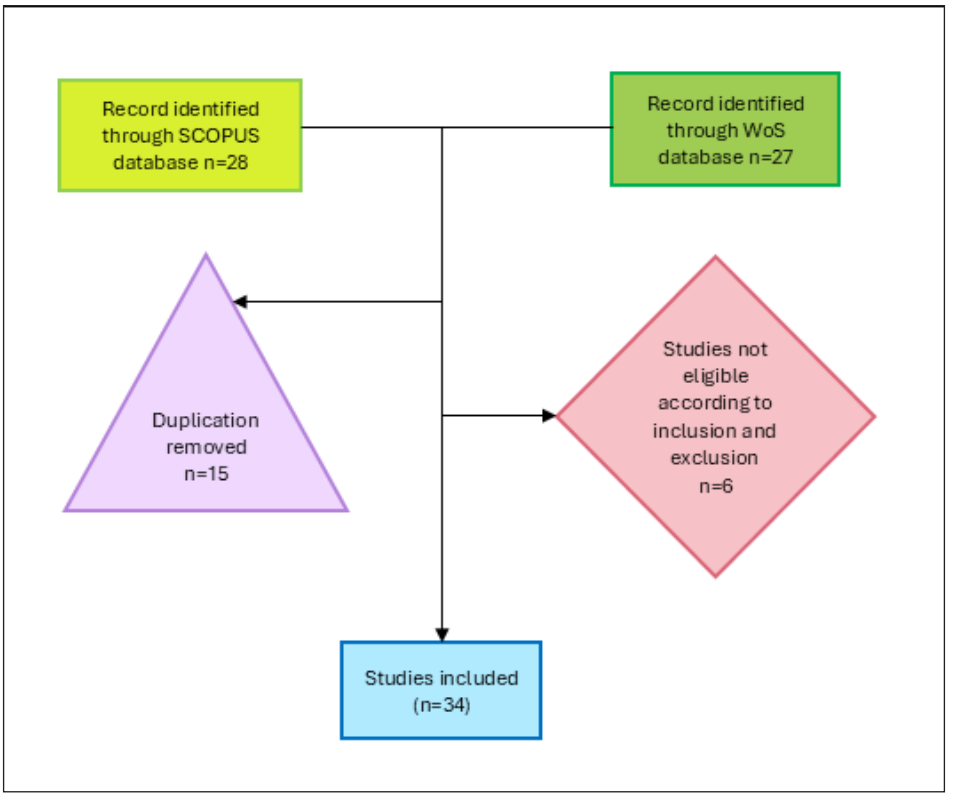

Figure 3. Selection process of studies in TreZ (Zairul, 2023)

RESULTS AND DISCUSSIONS

Following the meticulous methodology described in the previous section, 34 records were collected and analysed to identify the main themes and emerging patterns, which will help answer the research question. This analysis used ATLAS.ti, a qualitative data analysis tool that helps organise and interpret data systematically. Important results of the thematic review are summarised in this section. Both the quantitative and qualitative findings will be discussed next.

Quantitative Finding

Year of publications, industrial background, research location, and focal concept were used to analyse the study trends, which somewhat mirror the patterns of the data quality studies in the BDA domain to some extent. The number of relevant articles published gradually increased from 2020 to 2021 and reached a significant peak in 2022, as shown in Figure 4, but fell in 2023 before having a slight increase in 2024, indicating a continued interest in the study. Additionally, the trend of publications is illustrated in Figure 5.

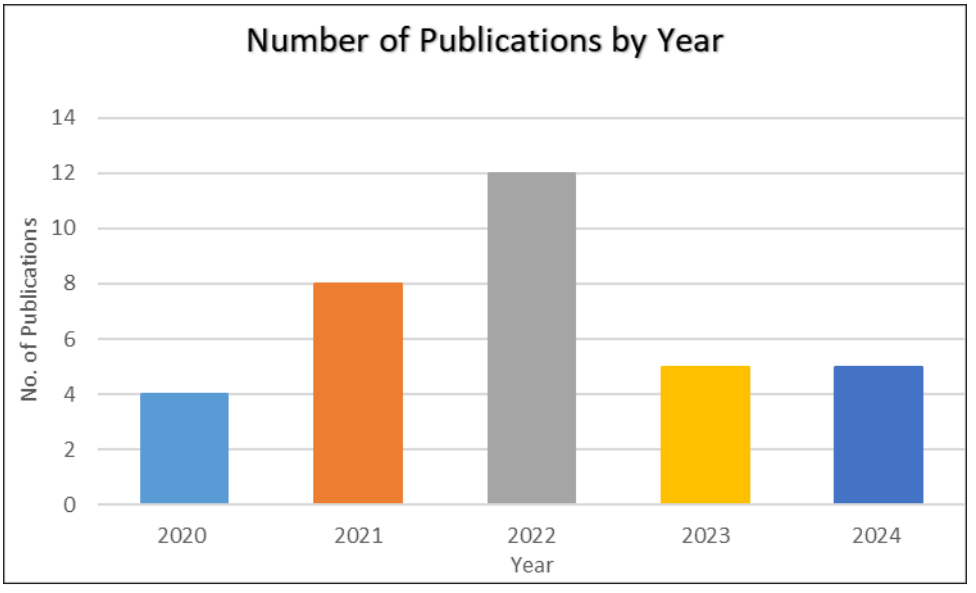

Figure 4. Paper breakdown by year of publication

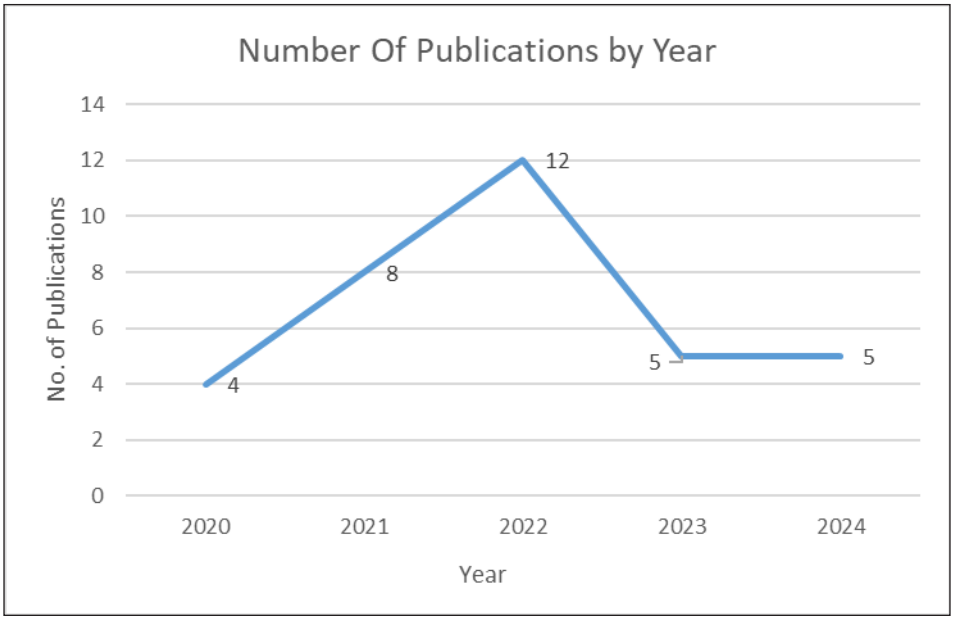

Figure 5. Publication trends from 2020 until 2024

The geographical dispersion reflects the global interest in BDA and disparities in research output that may correlate with different countries' economic, educational, and technological capacities. Figure 6 depicts the geographical distribution of articles published

across various countries from 2020 to 2024. In terms of publications, the topic of data quality in BDA is highly popular in developed countries, particularly in the United States of America (US) and the United Kingdom (UK), highlighting its significant role as a key player in the global research arena, especially within technology-driven sectors.

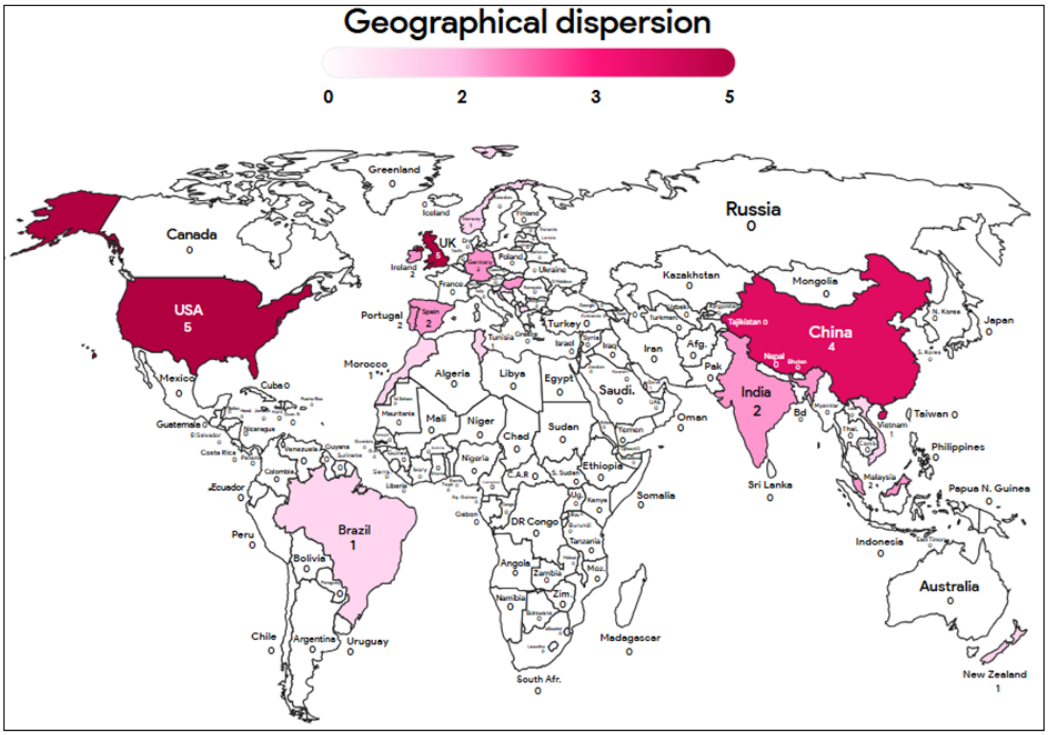

Figure 6. Geographical dispersion of articles published

The US and UK emerge as the frontrunners with the highest number of five publications. Most of the discussion is on data analytics and machine learning. (Hart et al., 2022), marketing insights for decision-making (Johnson et al., 2021), quality of Volunteer-contributed Geographic Data (VGI) (Zhang, 2022), use of BDA in purchasing and supply management (Patrucco et al., 2023), and factors of BDA success in various industries (Chen et al., 2022). The UK discussion, on the other hand, centres on the challenges of BDA implementation in the Facilities Management industry (Konanahalli et al., 2022), governance considerations of BDA in food consumer science (Timotijevic et al., 2022), AI in supply-chain decision-making (Hao & Demir, 2024), role-based access control using AI in Agriculture 4.0 (Spanaki et al., 2021), and aligning BDA capabilities (BDAC) and business models (BM) in small and medium-sized enterprises (SMEs) (Song et al., 2022).

China was ranked third in terms of article numbers and had many technical developments in BDA, which include integration of Morris design with AI and BDA to enhance network

performance and user experience (F. Song, 2024), detection of environmental violators using a big data approach (Chang et al., 2021), data pricing and utility evaluation in big data trading market (Chen et al., 2023), and determinants of BDA adoption in decision-making, specifically in New Zealand, China, and Vietnam (Yu et al., 2022). This latest study is consistent with the former perspective on data quality, organisational support, and technology readiness (Chen et al., 2022; Konanahalli et al., 2022; Timotijevic et al., 2022). The rest of the nations having publications on the topic are Germany (2), Hungary (2), India (2), Ireland (2), Malaysia (2), Portugal (2), Spain (2), Brazil (1), Croatia (1), Macedonia (1), Morocco (1), New Zealand (1), Norway (1), Qatar (1), Tunisia (1) and Vietnam (1).

Next, the study identified the themes and concerns of shortlisted articles. Five themes, namely T1 (Ontology and Data Quality Frameworks), T2 (Big Data Analytics in Various Industries), T3 (Machine Learning and AI Integration), T4 (Governance and Data Steward), and T5 (Tools and Techniques for Data Analysis) were specifically identified from the articles examined following the study's focus topic, as shown in Table 2. Figure 7 illustrates themes discussed in the literature. Each theme plays a distinct role in propelling the field of BDA forward, showcasing an equitable emphasis on a wide array of aspects.

The consistent importance attributed to "Tools and Techniques for Data Analysis" and "Machine Learning and AI Integration" highlights their foundational significance. At the same time, the enduring relevance of "Ontology and Data Quality Frameworks", "Big Data Analytics in Various Industries", and "Governance and Data Stewardship" emphasises their crucial contributions to the discipline. This equilibrium approach guarantees a thorough and resilient advancement in BDA, tackling diverse challenges and seizing opportunities to foster innovative and efficient solutions.

Figure 2 depicts the word cloud derived from collecting all 34 articles focusing on data quality within the domain of BDA from 2020 to 2024. It displays significant themes and core concepts relevant to scholarly investigations during this timeframe. The word cloud is generated using ATLAS.ti, utilising frequency-based weighting to highlight terms that appear frequently across numerous articles, resulting in their larger size in the word cloud. The prevalence of terms like "data", "quality", "analytics", "big", and "management" emphasise their fundamental significance within the discipline. These terms highlight the importance of upholding high standards and dependability of data, emphasising managerial procedures to sustain and enhance data quality. Keywords such as "information", "business", "decision", and "performance" indicate the considerable influence of data quality on organisational decision-making and overall performance. At the same time, terms like "research", "systems", "processing", and "technologies" suggest a combination of theoretical underpinnings and practical methodologies in the field.

In addition to that, operational terms like "implementation", "analysis", "models", "ai", and "tools" suggest the utilisation of analytical techniques and technological tools

to ensure the integrity of data. The terms "value", "adoption", and "impact" emphasise the broader consequences and advantages of implementing data quality protocols within organisations and industries. Furthermore, terms like "governance", "capability", and "innovation" propose a strategic viewpoint, highlighting the significance of governance structures and innovative strategies. Consequently, the word cloud portrays the intricate nature of data quality research in the domain of BDA, emphasising their strategic relevance and immediate influence on business and organisational outcomes.

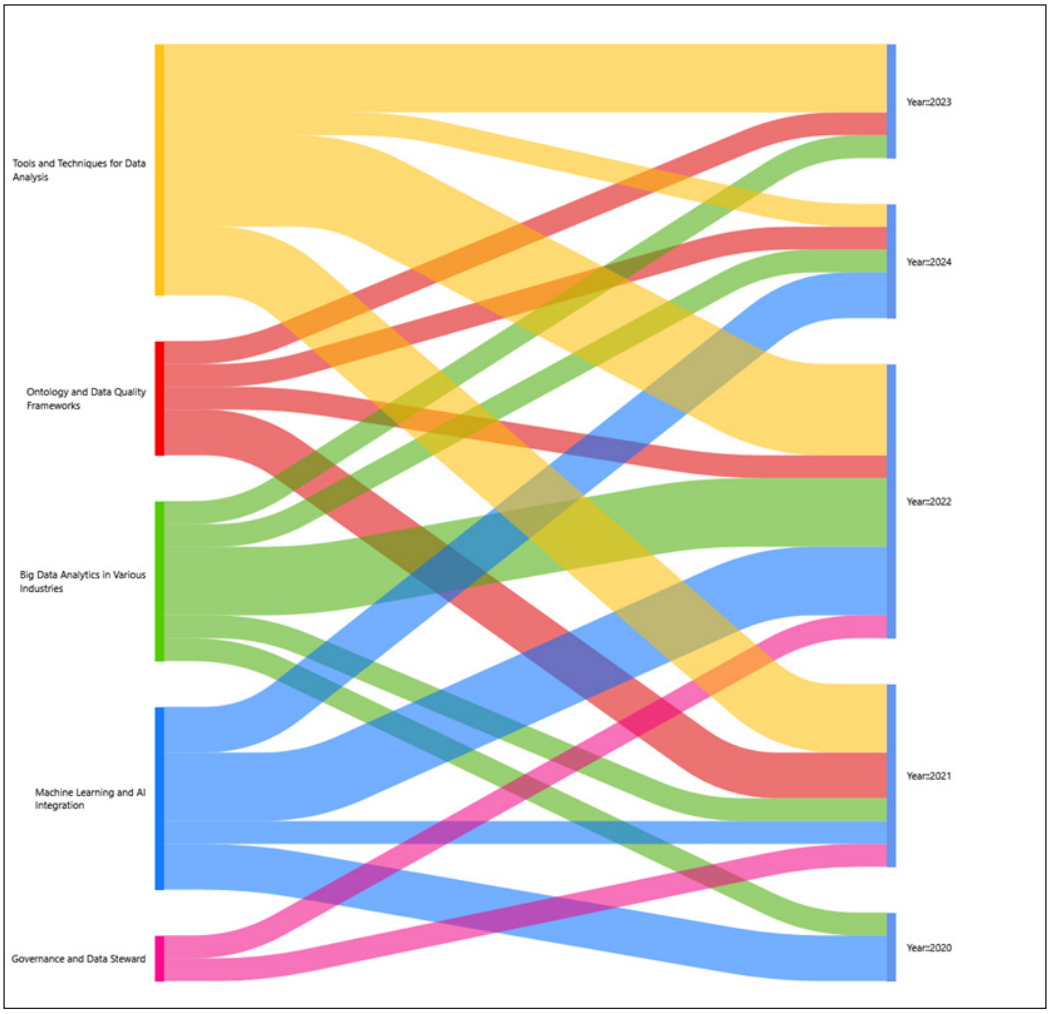

Figure 7. Themes discussed in the literature

Table 2
Tabulation of authors vs theme

	Theme 1: Ontology and Data Quality Frameworks	**Theme 2: Big Data Analytics in Various Industries**	**Theme 3: Machine Learning and AI Integration**	**Theme 4: Governance and Data Steward**	**Theme 5: Tools and Techniques for Data Analysis**
Barba-González et al. (2024)	/	-	-	-	-
Bui & Perera (2021)	-	-	-	-	/
Côrte-Real et al. (2020)	-	-	-	-	/
Hart et al. (2022)	/	-	-	-	-
Johnson et al. (2021)	-	/	-	-	-
Konanahalli et al. (2022)	-	/	-	-	-
Medeiros et al. (2021)	-	-	-	/	-
Radhakrishnan et al. (2022)	-	-	/	-	-
Song (2024)	-	-	/	-	-
Šprem et al. (2024)	-	-	-	-	/
Szukits & Móricz (2023)	-	-	-	-	/
Timotijevic et al. (2022)	-	-	-	/	-
Widad et al. (2023)	/	-	-	-	-
Wurster et al. (2024)	-	/	-	-	-
Yahia et al. (2021)	-	-	-	-	/
Yu et al. (2022)	-	-	-	-	/
Chen et al. (2023)	-	-	-	-	/
Al-Madhrahi et al. (2022)	-	-	-	-	/
Stach et al. (2022)	-	-	-	-	/
Lavalle et al. (2020)	-	-	/	-	-
Shidaganti & Prakash (2021)	/	-	-	-	-
(Shahi, 2023)	-	-	-	-	/
Zhang (2022)	-	-	-	-	/
Wook et al. (2021)	-	-	/	-	-
Phan & Tran (2022)	/	-	-	-	-
Jha et al. (2020)	-	-	/	-	-
Chang et al. (2021)	-	-	-	-	/
Savoska & Ristevski (2020)	-	/	-	-	-
Hao & Demir (2024)	-	-	-	/	-
Patrucco et al. (2023)	-	/	-	-	-
Spanaki et al. (2021)	/	-	-	-	-
Chen et al. (2022)	-	/	-	-	-
Rana et al. (2022)	-	-	/	-	-
Song et al. (2022)	-	/	-	-	-

Qualitative Finding

This thematic review paper studied publications and coded the data quality patterns in BDA. However, it did not address the future direction of BDA implementation. The initial codes were recorded, combined, and categorised in several rounds. Codes that were occasionally used and could not be categorised into any topic were removed since this study concerned aspects extensively discussed and investigated by researchers. Results from quantitative investigations that were not statistically significant were also removed. Furthermore, generic sociodemographic characteristics were not considered because they may not be universally applicable to all circumstances.

The first round of coding produced 12 initial codes, and the first stage was to become acquainted with the data. The first stage was to complete a thorough review of all the gathered articles and find relevant codes to develop the possible themes. BDA issues, BDA capabilities, data quality, data management, techniques and tools were among the phrases that prompted the creation of the initial subject. The next phase was generating, reviewing, and defining the final themes. A few rounds of discussions were held to refine these themes, resolving differences through consensus. This step ensured that the themes were consistent and accurately represented the data. Next, the themes were compared with findings from previous studies to confirm their relevance and alignment with existing research in the field. To ensure reliability, multiple researchers independently reviewed the themes. Finally, five major themes were identified, focusing on different aspects of data quality in BDA. The themes highlight diverse approaches and dimensions researchers cover, from data quality management, theoretical frameworks, and practical applications to technical advancements. Each theme contributes uniquely to advancing the understanding of data quality in BDA to ensure a balanced and comprehensive review of the study by addressing theoretical, applied, and technical dimensions.

Figure 8 systematically categorises significant themes that emerged from the analysis of 34 articles in the research area from 2020 to 2024. Theme 1, "Ontology and Data Quality Frameworks," focuses on developing standardised data definitions and the assurance of data integrity across various applications, underlining the need for robust frameworks to support data reliability. Theme 2, "Big Data Analytics in Various Industries," explores the widespread application of BDA across different sectors, indicating customised adaptations to meet specific industry challenges and foster innovation. A central query about the trends in data quality within this period intersects with Theme 3, "Machine Learning and AI Integration," which suggests incorporating advanced algorithms to enhance data analytics processes. Themes 4 and 5, "Governance and Data Steward" and "Tools and Techniques for Data Analysis," address the regulatory frameworks and practical tools necessary for effective data management and analysis. This diagram provides a structured framework for understanding the current and emerging trends in BDA, particularly focusing on enhancing data quality and integrating machine learning and AI.

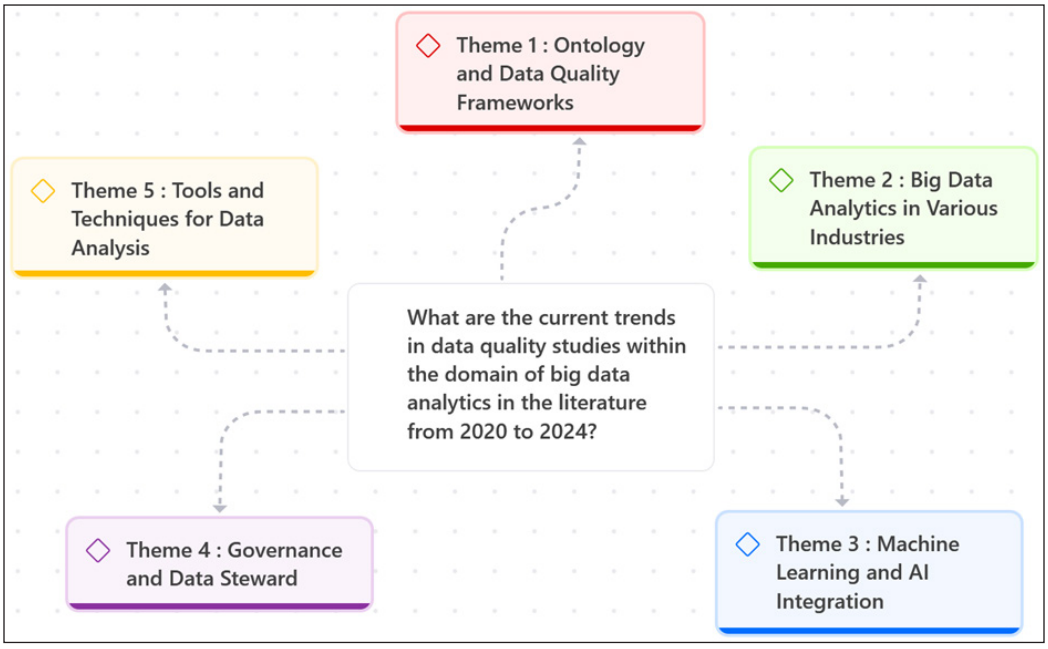

Figure 8. Overall network to answer research question

Theme 1: Ontology and Data Quality Frameworks

Figure 9 for "Theme 1: Ontology and Data Quality Frameworks" presents five articles, each contributing uniquely to data quality frameworks through ontology-based approaches. These articles collectively offer insights and methodologies applicable across various industries, emphasising the critical role of data quality in BDA. Spanaki et al. (2021) address the agricultural sector by proposing a framework for role-based data access control, emphasising the need for secure and efficient data management in environments where data sharing is crucial. In contrast, Phan and Tran (2022) focus on the banking sector, arguing that the complexities and regulatory demands of banking require a tailored approach to data quality management through BDA to improve operational efficiency and decision-

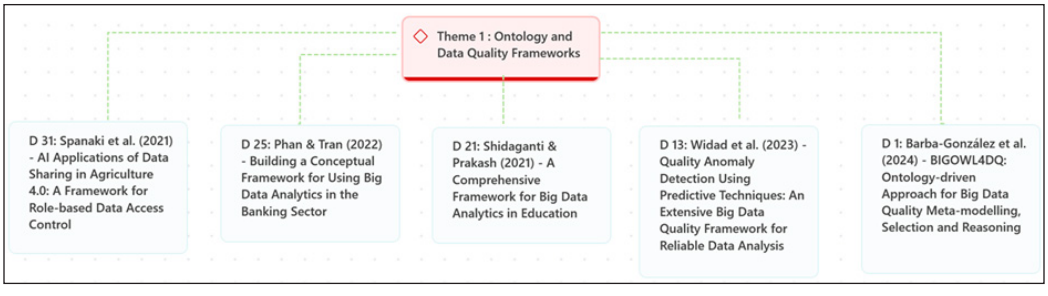

Figure 9. Theme 1: Ontology and Data Quality Frameworks

making. Shidaganti and Prakash (2021) take a different sectoral approach, presenting a framework for the educational sector that highlights the necessity of high data quality for effective educational analytics and the unique challenges in educational data management.

Widad et al. (2023) offer a broader perspective by introducing a framework for anomaly detection using predictive techniques applicable across various industries to maintain high data quality standards through advanced analytics. Barba-González et al. (2024) further broaden the scope with an ontology-driven approach, BIGOWL4DQ, for big data quality meta-modelling, selection, and reasoning; these studies argue that sophisticated ontology-based frameworks are essential for enhancing data integration and interoperability across diverse data environments. Together, these studies highlight the varied approaches and sector-specific needs in developing robust data quality frameworks, underscoring the importance of tailored and advanced methodologies to ensure high data quality standards in BDA.

Theme 2: Big Data Analytics in Various Industries

Figure 10 for "Theme 2: Big Data Analytics in Various Industries" presents articles exploring the diverse applications and implications of BDA across different sectors. Patrucco et al. (2023) emphasise the importance of absorptive capacity in strategic purchasing and supply chain management, arguing that the ability to absorb and utilise big data can significantly enhance decision-making processes. In contrast, using an exploratory factor analysis approach, Konanahalli et al. (2022) identify the drivers and challenges of implementing big data within the UK facilities management sector. This study suggests that while big data holds potential, its implementation faces numerous hurdles, including organisational and attitudinal barriers. Chen et al. (2022) delve deeper into these barriers, highlighting data, attitudinal, and organisational determinants that affect adopting BDA systems. The findings of Chen et al. (2022) align with those of Konanahalli et al. (2022), reinforcing that successful implementation requires addressing these fundamental issues. Meanwhile, Wurster et al. (2024) examine the impact of big data on the healthcare sector, focusing on implementing electronic medical records in a German hospital. Wurster argues that big data can improve documentation completeness and healthcare delivery, a perspective that contrasts with the more cautious views of Chen et al. (2022) and Konanahalli et al. (2022), who emphasise the challenges over the benefits.

Savoska and Ristevski (2020) discuss the pharmaceutical industry, advocating for adopting big data concepts to improve operational efficiencies and innovation. This perspective adds to Patrucco et al.'s (2023) arguments for strategic advantages in supply chain management. Both studies demonstrate how big data can transform business operations. Song et al. (2022) further support this view by demonstrating how SMEs leveraged BDA capabilities to maintain competitive performance during COVID-19. Song

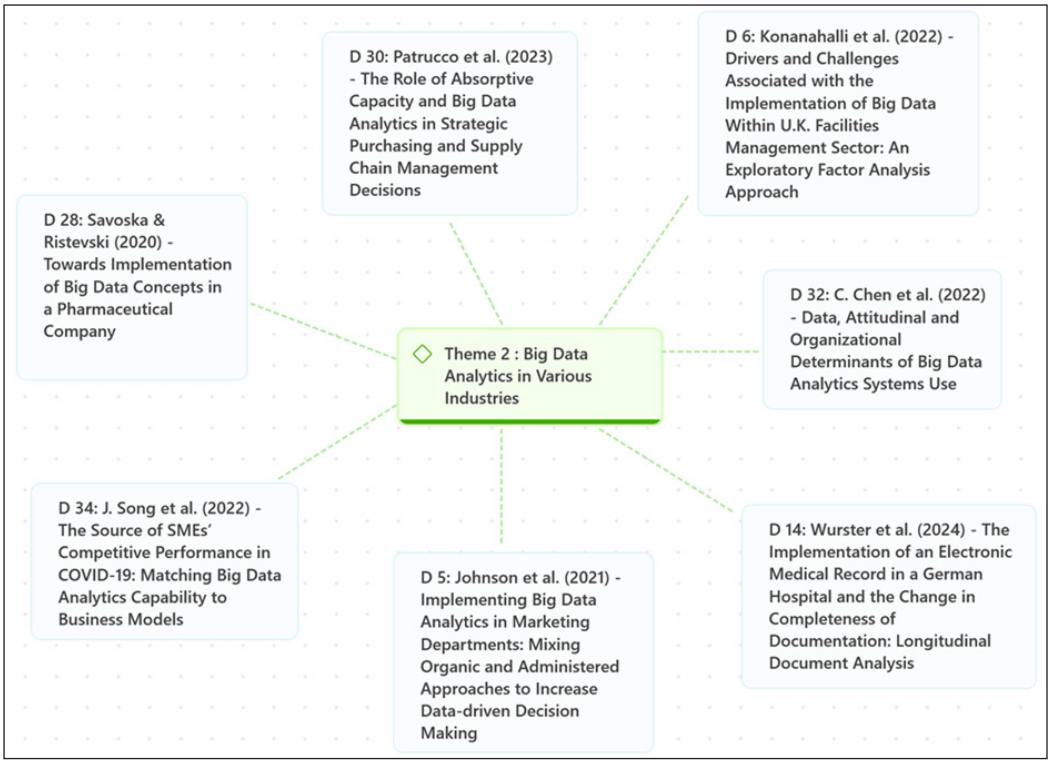

Figure 10. Theme 2: Big Data Analytics in Various Industries

et al.'s (2022) study highlight the adaptability and resilience that big data can provide to businesses, which resonates with the optimistic outlook of Savoska and Ristevski (2020). Johnson et al. (2021) explore the integration of BDA in the marketing sector, proposing a mix of organic and administered approaches to enhance data-driven decision-making. This approach highlights the need for flexibility and innovation in implementing big data solutions, echoing the broader themes of adaptability and strategic advantage found in Patrucco et al.'s (2023) and Song et al.'s (2022) studies. Although BDA offers significant benefits across various industries, its successful implementation requires overcoming substantial organisational, attitudinal, and data-related challenges. The contrasting perspectives and sector-specific insights emphasise the need for tailored strategies to harness the full potential of BDA.

Theme 3: Machine Learning and AI Integration

Several authors explore the integration of machine learning (ML) and AI across different sectors, highlighting the benefits and challenges of such implementations (Figure 11). Jha et al. (2020) emphasise the importance of developing BDA capabilities within supply chains, arguing that ML and AI can enhance efficiency and responsiveness. This perspective aligns

with Hart et al. (2022), who apply ML to large-scale synchrophasor datasets, demonstrating the need to evaluate the "machine learning-readiness" of data to ensure effective power grid management. Contrasting these optimistic views, Rana et al. (2022) explore the dark side of AI in business analytics, highlighting potential operational inefficiencies and competitiveness issues that can arise from AI integration. Rana et al.'s (2022) findings serve as a cautionary tale, suggesting that while AI has substantial potential, it also carries risks that need careful management. This argument is supported by Hao and Demir (2024), who propose an environmental, social, and governance (ESG) framework for AI in supply chain decision-making, emphasising the need for responsible AI integration that considers broader societal impacts.

Wook et al. (2021) take a more technical approach, exploring big data's traits and quality dimensions necessary for successful ML and AI applications. By using partial least squares structural equation modelling, they highlight the critical role of high-quality data in achieving effective AI outcomes. This technical focus is echoed by Lavalle et al. (2020), who advocate using visualisation techniques in smart cities to enhance sustainability through big data captured by IoT devices. Their work illustrates how ML and AI can be leveraged to improve urban efficiency and sustainability. Song (2024) offers another dimension by incorporating Morris' design thoughts for AI and big data to optimise wireless communication networks in China. This study highlights the technological advancements

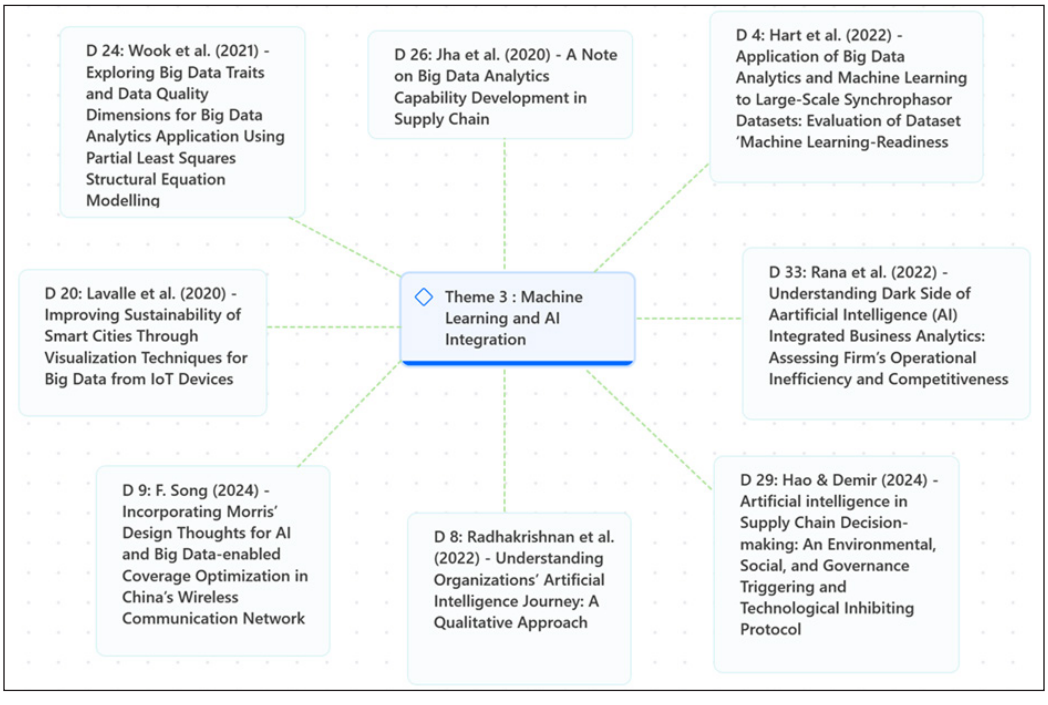

Figure 11. Theme 3: Machine Learning and AI Integration

that ML and AI can drive, showcasing their potential to optimise complex systems. Meanwhile, Radhakrishnan et al. (2022) provide a qualitative approach to understanding the AI journey of organisations, offering insights into the practical challenges and strategies for AI adoption. Radhakrishnan et al.'s (2022) work complements the technical and sector-specific studies by focusing on AI integration's human and organisational aspects. ML and AI integration has potential benefits and significant challenges and can drive efficiency, sustainability, and technological optimisation. However, successful implementation requires careful consideration of data quality, potential risks, and responsible governance. This balanced perspective emphasises the need for tailored, context-specific strategies to fully harness the potential of ML and AI in BDA.

Theme 4: Governance and Data Steward

The theme of "Governance and Data Steward" within BDA is explored through two scholarly works that illustrate the critical impact of governance in different contexts: sector-specific and corporate (Figure 12). Timotijevic et al. (2022) focus on implementing governance within the food and nutrition sector, highlighting how specialised governance frameworks are necessary to ensure data integrity and reliability, particularly when data accuracy directly impacts public health. This approach highlights the protective and regulatory dimensions of governance tailored to meet stringent sector-specific challenges. In contrast, Medeiros et al. (2021) discuss the strategic role of data stewardship in corporate performance management, advocating that effective governance is pivotal for managing data and harnessing it to drive corporate strategies and enhance operational efficiencies. This perspective positions data stewardship as a transformative tool for achieving competitive advantages in the business field. These contrasting viewpoints emphasise the versatility of governance roles: On one hand, Timotijevic et al. (2022) present a protective stance focused on compliance and safety; on the other hand, Medeiros et al. (2021) highlight the

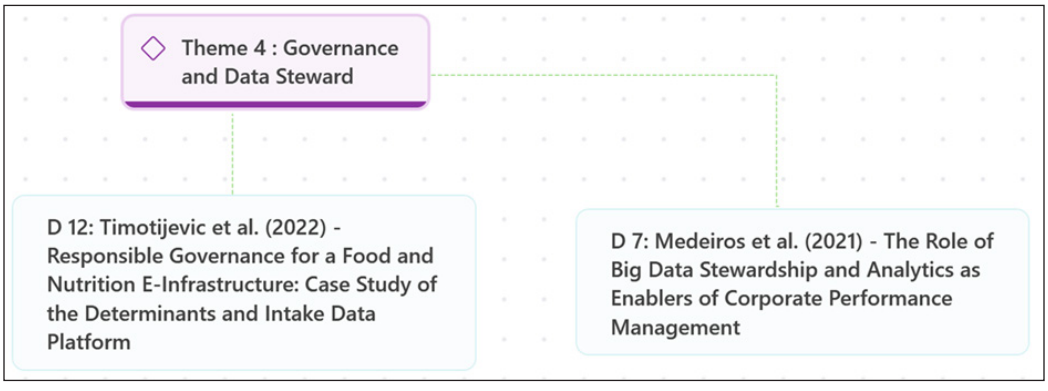

Figure 12. Theme 4: Governance and Data Steward

potential of governance to catalyse business insights and success. Despite an adaptable governance strategy that aligns with specific organisational goals and sector requirements, data integrity and ethical management principles have varying applications and impacts depending on the context.

Theme 5: Tools and Techniques for Data Analysis

The "Tools and Techniques for Data Analysis" theme reveals diverse approaches and insights across various studies (Figure 13). The works of Bui & Perera (2021) and Corte-Real et al. (2020) highlight the integration of IoT with BDA to address domain-specific challenges. While Bui & Perera (2021) focus on monitoring ship performance under unique operational conditions, Corte-Real et al. (2020) examine IoT and big data use within European and American firms. These studies suggest a preference for combining IoT-generated data with advanced analytics, as this synergy provides actionable insights tailored to the operational context. Similarly, Šprem et al. (2024) and Yahia et al. (2021) delve into developing tools and frameworks for handling large-scale data processes. Šprem et al. (2024) focus on creating web applications to streamline data ingestion and processing,

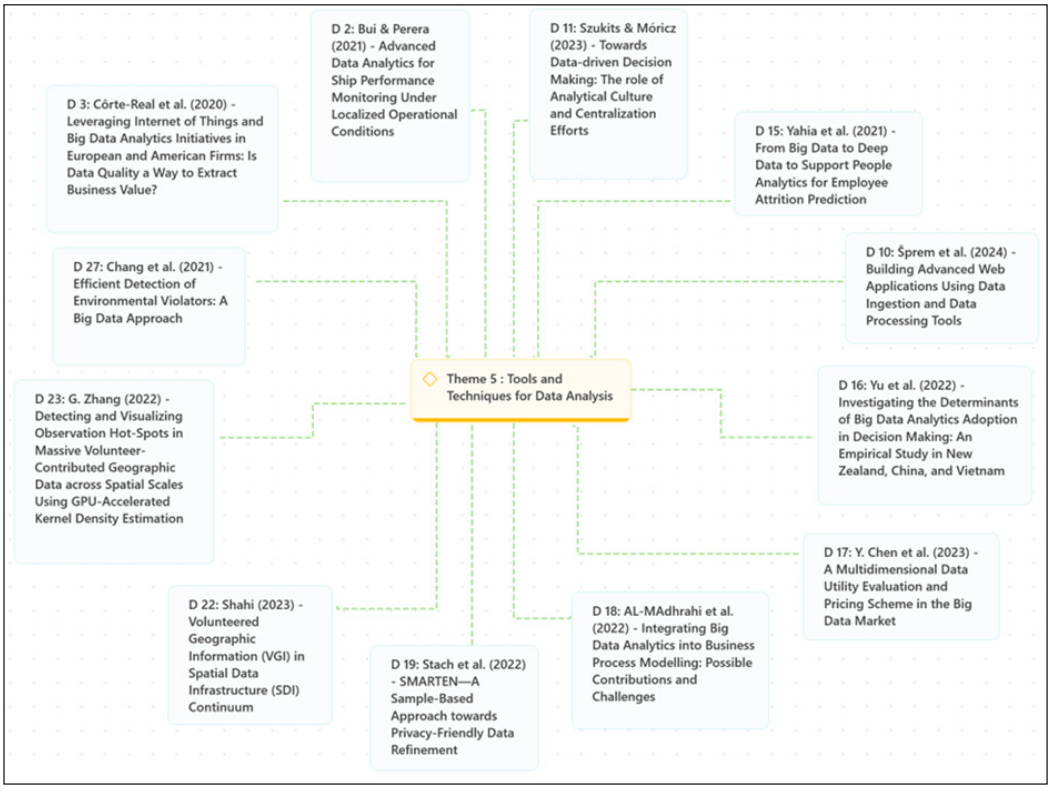

Figure 13. Theme 5: Tools and Techniques for Data Analysis

while Yahia et al. (2021) leverage advanced big data techniques, such as ML, to predict employee attrition. Both studies demonstrate a trend toward creating specialised tools that simplify data handling and enhance business decision-making capabilities.

Chen et al. (2023) and Yu et al. (2022) extend this focus by analysing economic and decision-making implications. Chen et al. (2023) evaluate pricing schemes in the big data marketplace, while Yu et al. (2022) investigate the factors influencing big data adoption in decision-making within organisations across New Zealand, China, and Vietnam. Their findings reveal an emerging emphasis on understanding the economic impact of data analysis tools and their effectiveness in supporting organisational strategies. Environmental and spatial data applications are explored in studies like Chang et al. (2021), Zhang (2022), and Shahi (2023). Chang et al. (2021) address detecting environmental violations using big data, while Zhang (2022) presents GPU-accelerated methods to visualise geographic data at scale. Shahi (2023) investigates volunteered geographic information (VGI) in spatial data infrastructures. These studies illustrate a preference for leveraging domain-specific tools, such as geographic visualisation and environmental monitoring platforms, to address unique analytical challenges.

Stach et al. (2022) and Al-Madhrahi et al. (2022) tackle privacy and ethical considerations. Stach et al. (2022) introduce the SMARTE approach for privacy-focused data refinement, and AL-MAdhrahi et al. (2022) explore the integration of analytics into business process modelling. These contributions highlight the growing concern for ethical practices and the necessity of embedding privacy-preserving measures in analytics frameworks. Finally, Szukits and Móricz (2023) examine the organisational culture around data-driven decision-making. The study emphasises the role of analytical centralisation in fostering effective data usage, reflecting a broader pattern where organisations are increasingly investing in cultural shifts to embrace data-centric approaches.

Across these studies, certain patterns emerge, such as the reliance on domain-specific tools and a growing emphasis on integrating advanced techniques like machine learning and GPU acceleration. There is also a noticeable focus on the ethical and economic aspects of data analysis tools, highlighting evolving priorities in the field. By considering these patterns, businesses and researchers can align their strategies with proven practices and adopt tools that meet their specific analytical needs.

LIMITATION OF THE STUDY

This study offers valuable insights into trends in data quality within the BDA domain. However, several limitations exist that may affect the generalizability and interpretation of the findings. These limitations are particularly related to database selection bias, dependence on open-access literature, and geographical dispersion of research.

First, the search strategy employed in this study focused on two databases, SCOPUS and WoS, both known for their coverage of high-quality, peer-reviewed research. However, selection bias arises from excluding other databases, such as IEEE Xplore or Google Scholar, which may include additional relevant studies. By limiting the scope to these databases, the study may have missed significant contributions from disciplines or regions underrepresented in SCOPUS and WoS. For instance, emerging economies advancing in BDA may publish relevant works in local or regional journals not indexed in these databases. Additionally, reliance on open-access publications may introduce potential publication bias. Open-access studies do not always represent the broader scientific discourse, as they often exclude works published in subscription-based journals, which may offer valuable and diverse perspectives. This limitation may restrict the comprehensiveness of the findings and the generalizability of the conclusions drawn.

Second, the geographical distribution of research articles highlights disparities in research productivity, with significant contributions from the US, UK and China. Conversely, regions in developed countries such as Canada, Australia, and Russia are underrepresented in the findings despite their potential contributions to BDA. This uneven representation may reflect systemic biases in publication practices or limitations in the search methodology. Future studies could explore alternative sources or collaborate with researchers from underrepresented regions to achieve a more balanced geographical representation.

Finally, while the study identifies critical themes, such as ontology frameworks, machine learning integration, and governance in BDA, the focus remains on theoretical and methodological discussions. The absence of extensive real-world applications limits the contextual relevance of the findings. Addressing practical challenges, particularly from the perspective of BDA implementation in diverse sectors, could provide a more comprehensive understanding of how data quality issues affect decision-making and outcomes.

DISCUSSION AND RECOMMENDATION

Artificial intelligence (AI) has emerged as a pivotal tool in addressing the long-standing data quality issue in BDA. The literature consistently highlights the criticality of data quality in ensuring accurate, reliable, and actionable insights from large datasets. High-quality data enhances the value of BDA, while poor-quality data compromises decision-making and operational efficiency. AI offers a range of capabilities that can directly address data quality issues. Machine learning algorithms can identify patterns and anomalies in data, flagging inconsistencies and potential errors. For example, predictive models have been employed to detect quality anomalies, as highlighted by Widad et al. (2023). Similarly, ontology-driven approaches, such as BIGOWL4DQ proposed by Barba-González et al. (2024), facilitate data integration and interoperability, addressing structural and semantic

inconsistencies. These approaches highlight the potential of AI to enhance data accuracy and reliability across various sectors.

AI's ability to learn and adapt also allows for dynamic quality monitoring. AI-powered tools can continuously audit datasets to ensure compliance with quality standards. For instance, role-based access frameworks discussed by Spanaki et al. (2021) integrate AI to manage data securely, ensuring that only high-quality and relevant data is accessed and utilised. Moreover, AI-driven natural language processing (NLP) and automation reduce the manual effort required for data cleaning and preparation, making the process more efficient and less error-prone.

Despite its potential, AI-integrated BDA faces several challenges. Data governance frameworks, essential for managing and maintaining data quality, often lag behind technological advancements. Many frameworks fail to adequately address the complexities of real-time data processing and multi-source data integration. Moreover, ethical concerns such as data privacy, algorithmic bias, and implications of poor data quality on decision-making remain significant barriers. Data privacy is a critical concern in BDA, especially with the increasing use of personal and sensitive information. AI models often require large volumes of data for training, which may lead to privacy violations if data is inadequately anonymised or secured. Studies like Timotijevic et al. (2022) highlight the need for stringent governance mechanisms to protect sensitive data, particularly in sectors like healthcare and food safety. Bias in AI models is another pressing issue. If training datasets are skewed or incomplete, AI algorithms may perpetuate existing biases, leading to unfair or inaccurate outcomes. Rana et al. (2022) emphasise how biases in business analytics can exacerbate operational inefficiencies and competitiveness issues. Furthermore, poor data quality, manifesting as missing, inaccurate, or outdated data, can severely compromise decision-making. For instance, decision-support systems relying on flawed data may produce misleading insights, ultimately affecting organisational performance and trust in AI systems.

Integrating governance frameworks with AI and BDA offers a pathway to mitigate these challenges. Effective governance ensures data is managed responsibly, ethically, and aligned with regulatory standards. AI can be harnessed within these frameworks to automate compliance monitoring, flag governance breaches, and ensure adherence to ethical norms. For example, Hao and Demir (2024) propose an ESG framework for AI-driven supply chain decision-making, which balances technological efficiency with environmental and social considerations. Such frameworks can also address the dual challenges of privacy and security by embedding robust encryption and access control mechanisms. By incorporating AI, governance frameworks can become adaptive and responsive, dynamically evolving with emerging challenges in the data ecosystem. Governance frameworks can also play a crucial role in mitigating bias. Combined with explainable AI models, transparent governance structures can ensure accountability and fairness. For instance, Wook et al.

(2021) highlight the importance of data quality dimensions in achieving unbiased AI outcomes. By integrating governance mechanisms with AI, organisations can foster ethical decision-making processes while leveraging the full potential of BDA.

The future of data quality in BDA is intertwined with advancements in emerging technologies such as blockchain, edge computing, and quantum computing. These technologies offer transformative potential to address current limitations and shape the future trajectory of the field. Blockchain technology provides a decentralised and immutable ledger for data storage and sharing, ensuring transparency and traceability. Blockchain can enhance data quality by preventing tampering and maintaining an auditable record of data modifications. This technology is particularly valuable in sectors where data integrity is paramount, such as healthcare and finance. Future studies could explore the integration of blockchain with AI to create secure and transparent data ecosystems. Edge computing shifts data processing closer to the source, reducing latency and improving the quality of real-time analytics. This approach is especially relevant for IoT applications, where timely insights are critical. By enabling localised data processing, edge computing minimises the risk of data degradation during transmission. Researchers could investigate how edge computing can be combined with AI to enhance the accuracy and timeliness of data analytics in distributed environments. Quantum computing, though still in its nascent stages, holds promise for solving complex optimisation problems and processing vast datasets with unparalleled speed. Future research could examine how quantum algorithms might be applied to data quality challenges, particularly in handling large-scale unstructured data.

In addition to these technologies, there is a growing need for frameworks prioritising ethical considerations in AI and BDA. The development of explainable AI (XAI) models, which provide transparent reasoning for algorithmic decisions, can build trust and accountability. Studies could also focus on developing international data quality and governance standards, ensuring consistency across borders and industries.

CONTRIBUTIONS AND BENEFITS OF STUDY

The contributions of this study are manifold, offering a comprehensive analysis of present trends in data quality studies within the BDA field from 2020 to 2024. This review emphasises the multidimensional nature of data quality issues and their broad applicability across different sectors. The study reveals significant trends, such as the maximum number of publications recorded in 2022, emphasising how important data quality is for using big data for various applications. The geographical dispersion of research output, particularly the prominence of the United States and the United Kingdom, reflects the global interest and its correlation with the economic, educational, and technological capacities of different countries.

The study's thematic review classifies publications into five themes: Ontology and Data Quality Frameworks, Big Data Analytics in Various Industries, Machine Learning and AI Integration, Governance and Data Stewardship, and Tools and Techniques for Data Analysis. This method efficiently synthesises much information, which makes it easier to identify gaps, new trends, and areas that require more research. In addition to highlighting the fundamental importance of tools and methods for data analysis and the fusion of AI and machine learning, this organised framework also highlights the vital function that strong data quality frameworks and governance structures play. The study shows how big data technology may be flexible and transformational in various fields by looking at industry-specific applications, such as marketing, supply chain management, and healthcare. This comprehensive overview not only guides future research but also informs best practices in BDA implementation, capturing the complexity and dynamic nature of this field.

ACKNOWLEDGEMENTS

The authors express their sincere gratitude to the College of Computing, Informatics, and Mathematics, the Institute of Postgraduate Studies (IPSis), and the Research Management Centre, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia, for providing the infrastructure and facilities for this research and supporting the publication of this paper.

REFERENCES

Al-Madhrahi, Z., Singh, D., & Yadegaridehkordi, E. (2022). Integrating big data analytics into business process modelling: Possible contributions and challenges. *International Journal of Advanced Computer Science and Applications*, *13*(6), 461–468. https://doi.org/10.14569/IJACSA.2022.0130657

Barba-González, C., Caballero, I., Varela-Vaca, Á. J., Cruz-Lemus, J. A., Gómez-López, M. T., & Navas-Delgado, I. (2024). BIGOWL4DQ: Ontology-driven approach for big data quality meta-modelling, selection and reasoning. *Information and Software Technology*, *167*, Article 107378. https://doi.org/10.1016/j.infsof.2023.107378

Bui, K. Q., & Perera, L. P. (2021). Advanced data analytics for ship performance monitoring under localized operational conditions. *Ocean Engineering*, *235*, Article 109392. https://doi.org/10.1016/j.oceaneng.2021.109392

Chang, X., Huang, Y., Li, M., Bo, X., & Kumar, S. (2021). Efficient detection of environmental violators: A big data approach. *Production and Operations Management*, *30*(5), 1246–1270. https://doi.org/10.1111/poms.13272

Chen, C., Choi, H. S., & Ractham, P. (2022). Data, attitudinal and organizational determinants of big data analytics systems use. *Cogent Business & Management*, *9*(1), Article 2043535. https://doi.org/10.1080/23311975.2022.2043535

Chen, Y., Bai, R., Wu, Y., Li, T., & Zhou, H. (2023). A multidimensional data utility evaluation and pricing scheme in the big data market. *Wireless Communications and Mobile Computing, 2023*(1), Article 6217495. https://doi.org/10.1155/2023/6217495

Clarke, V., & Braun, V. (2013). Teaching thematic analysis: Overcoming challenges and developing strategies for effective learning. *The Psychologist*, *26*(2), 120–123.

Côrte-Real, N., Ruivo, P., & Oliveira, T. (2020). Leveraging internet of things and big data analytics initiatives in European and American firms: Is data quality a way to extract business value? *Information & Management*, *57*, Article 103141. https://doi.org/10.1016/j.im.2019.01.003

Hao, X., & Demir, E. (2024). Artificial intelligence in supply chain decision-making: an environmental, social, and governance triggering and technological inhibiting protocol. *Journal of Modelling in Management*, *19*(2), 605–629. https://doi.org/10.1108/JM2-01-2023-0009

Hart, P., He, L., Wang, T., Kumar, V. S., Aggour, K., Subramanian, A., & Yan, W. (2022). Application of big data analytics and machine learning to large-scale synchrophasor datasets: Evaluation of dataset 'Machine Learning-Readiness'. *IEEE Open Access Journal of Power and Energy*, *9*, 386–397. https://doi.org/10.1109/OAJPE.2022.3197553

Jha, A. K., Agi, M. A. N. N., & Ngai, E. W. T. T. (2020). A note on big data analytics capability development in supply chain. *Decision Support Systems*, *138*(2020), Article 113382. https://doi.org/10.1016/j.dss.2020.113382

Johnson, D. S., Sihi, D., & Muzellec, L. (2021). Implementing big data analytics in marketing departments: Mixing organic and administered approaches to increase data-driven decision making. *Informatics, 8*(4), Article 66. https://doi.org/10.3390/informatics8040066

Konanahalli, A., Marinelli, M., & Oyedele, L. (2022). Drivers and challenges associated with the implementation of big data within U.K. facilities management sector: An exploratory factor analysis approach. *IEEE Transactions on Engineering Management*, *69*(4), 916–929. https://doi.org/10.1109/TEM.2019.2959914

Lavalle, A., Teruel, M. A., Maté, A., & Trujillo, J. (2020). Improving sustainability of smart cities through visualization techniques for big data from IoT devices. *Sustainability*, *12*(14), Article 5595. https://doi.org/10.3390/su12145595

Medeiros, M. M., MaçAda, A. C. G., & Hoppen, N. (2021). The role of big data stewardship and analytics as enablers of corporate performance management. *Revista de Administracao Mackenzie*, *22*(6), Article eRAMD210063. https://doi.org/10.1590/1678-6971/eRAMD210063

Patrucco, A. S., Marzi, G., & Trabucchi, D. (2023). The role of absorptive capacity and big data analytics in strategic purchasing and supply chain management decisions. *Technovation*, *126*(2023), Article 102814. https://doi.org/10.1016/j.technovation.2023.102814

Phan, D. T., & Tran, L. Q. T. (2022). Building a conceptual framework for using big data analytics in the banking sector. *Intellectual Economics*, *16*(1), 5–23. https://doi.org/10.13165/IE-22-16-1-01

Radhakrishnan, J., Gupta, S., & Prashar, S. (2022). Understanding organizations' artificial intelligence journey: A qualitative approach. *Pacific Asia Journal of the Association for Information Systems*, *14*(6), 43–77. https://doi.org/10.17705/1pais.14602

Rana, N. P., Chatterjee, S., Dwivedi, Y. K., & Akter, S. (2022). Understanding dark side of artificial intelligence (AI) integrated business analytics: assessing firm's operational inefficiency and competitiveness. *European Journal of Information Systems*, *31*(3), 364–387. https://doi.org/10.1080/0960085X.2021.1955628

Savoska, S., & Ristevski, B. (2020). Towards implementation of big data concepts in a pharmaceutical company. *Open Computer Science*, *10*(1), 343–356. https://doi.org/10.1515/comp-2020-0201

Shahi, K. (2023). Volunteered Geographic Information (VGI) in Spatial Data Infrastructure (SDI) continuum. *EAI Endorsed Transactions on Internet of Things*, *9*(1), Article e3. https://doi.org/10.4108/eetiot.v9i1.2979

Shidaganti, G., & Prakash, S. (2021). A comprehensive framework for big data analytics in education. *International Journal of Advanced Computer Science and Applications*, *12*(9), 218–227. https://doi.org/10.14569/IJACSA.2021.0120926

Song, F. (2024). Incorporating Morris' design thoughts for AI and big data-enabled coverage optimization in China's wireless communication network. *Journal of Information Systems Engineering and Management*, *9*(1), Article 23622. https://doi.org/10.55267/iadt.07.14076

Song, J., Xia, S., Vrontis, D., Sukumar, A., Liao, B., Li, Q., Tian, K., & Yao, N. (2022). The source of SMEs' competitive performance in COVID-19: Matching big data analytics capability to business models. *Information Systems Frontiers*, *24*, 1167–1187. https://doi.org/10.1007/s10796-022-10287-0

Spanaki, K., Karafili, E., & Despoudi, S. (2021). AI applications of data sharing in agriculture 4.0: A framework for role-based data access control. *International Journal of Information Management*, *59*, Article 102350. https://doi.org/10.1016/j.ijinfomgt.2021.102350

Šprem, Š., Tomažin, N., Matečić, J., & Horvat, M. (2024). Building advanced web applications using data ingestion and data processing tools. *Electronics*, *13*(4), Article 0709. https://doi.org/10.3390/electronics13040709

Stach, C., Behringer, M., Bräcker, J., Gritti, C., & Mitschang, B. (2022). SMARTEN - A sample-based approach towards privacy-friendly data refinement. *Journal of Cybersecurity and Privacy*, *2*(3), 606–628. https://doi.org/10.3390/jcp2030031

Szukits, Á., & Móricz, P. (2023). Towards data-driven decision making: The role of analytical culture and centralization efforts. *Review of Managerial Science, 18*(10), 2849-2887. https://doi.org/10.1007/s11846-023-00694-1

Teh, H. Y., Kempa-Liehr, A. W., & Wang, K. I. K. (2020). Sensor data quality: A systematic review. *Journal of Big Data*, *7*(1), 1–49. https://doi.org/10.1186/s40537-020-0285-1

Timotijevic, L., Carr, I., De La Cueva, J., Eftimov, T., Hodgkins, C. E., Seljak, B. K., Mikkelsen, B. E., Selnes, T., Van't Veer, P., & Zimmermann, K. (2022). Responsible governance for a food and nutrition e-infrastructure: Case study of the determinants and intake data platform. *Frontiers in Nutrition*, *8*, Article 795802. https://doi.org/10.3389/fnut.2021.795802

Widad, E., Saida, E., & Gahi, Y. (2023). Quality anomaly detection using predictive techniques: An extensive big data quality framework for reliable data analysis. *IEEE Access*, *11*, 103306–103318. https://doi.org/10.1109/ACCESS.2023.3317354

Wook, M., Hasbullah, N. A., Zainudin, N. M., Jabar, Z. Z. A., Ramli, S., Razali, N. A. M., & Yusop, N. M. M. (2021). Exploring big data traits and data quality dimensions for big data analytics application using partial least squares structural equation modelling. *Journal of Big Data*, *8*(1), 1-15. https://doi.org/10.1186/s40537-021-00439-5

Wurster, F., Beckmann, M., Cecon-Stabel, N., Dittmer, K., Jes Hansen, T., Jaschke, J., Köberlein-Neu, J., Okumu, M. R., Rusniok, C., Pfaff, H., & Karbach, U. (2024). The implementation of an electronic medical record in a German Hospital and the change in completeness of documentation: Longitudinal document analysis. *JMIR Medical Informatics*, *12*(1), Article e47761. https://doi.org/10.2196/47761

Yahia, N. B., Hlel, J., & Colomo-Palacios, R. (2021). From big data to deep data to support people analytics for employee attrition prediction. *IEEE Access*, *9*, 60447–60458. https://doi.org/10.1109/ACCESS.2021.3074559

Yu, J., Taskin, N., Nguyen, C. P., Li, J., & Pauleen, D. J. (2022). Investigating the determinants of big data analytics adoption in decision making: An empirical study in New Zealand, China, and Vietnam. *Pacific Asia Journal of the Association for Information Systems*, *14*(4), 62–99. https://doi.org/10.17705/1pais.14403

Zairul, M. (2020). A thematic review on student-centered learning in the studio education. *Journal of Critical Reviews*, *7*(2), 504–511. https://doi.org/10.31838/jcr.07.02.95

Zairul, M. (2021). A thematic review on Industrialised Building System (IBS) publications from 2015-2019: Analysis of patterns and trends for future studies of IBS in Malaysia. *Pertanika Journal of Social Sciences and Humanities*, *29*(1), 635–652. https://doi.org/10.47836/PJSSH.29.1.35

Zairul, M. (2023). *Thematic Review template (Patent No. CRLY2023W02032).* Controller of Copyright.

Zairul, M., Azli, M., & Azlan, A. (2023). Defying tradition or maintaining the status quo? Moving towards a new hybrid architecture studio education to support blended learning post-COVID-19. *Archnet-IJAR: International Journal of Architectural Research*, *17*(3), 554–573. https://doi.org/10.1108/ARCH-11-2022-0251

Zairul, M., & Zaremohzzabieh, Z. (2023). Thematic trends in Industry 4.0 Revolution potential towards sustainability in the construction industry. *Sustainability*, *15*, Article 7720. https://doi.org/10.3390/su15097720

Zhang, G. (2022). Detecting and visualizing observation hot-spots in massive volunteer-contributed geographic data across spatial scales using GPU-accelerated kernel density estimation. *ISPRS International Journal of Geo-Information*, *11*(1), Article 55. https://doi.org/10.3390/ijgi11010055