Applied Mathematics and Computational Intelligence
Volume 15, No. 3, 2026 [108-122]

Smart Transportation System: A Literature and Bibliometric Analysis

Mohd Azri Rosli[1], Nor Azuana Ramli[2*], Lilik Jamilatul Awalin[3]

[1,2]Centre for Mathematical Science, Universiti Malaysia Pahang Al-Sultan Abdullah, Lebuh Persiaran Tun Khalil Yaakob, 26300 Kuantan, Pahang, Malaysia.
[3]Department of Engineering, Faculty of Advanced Technology and Multidicipline, Universitas Airlangga, Jl. Mulyorejo, Surabaya 60115, Indonesia.

[*] Corresponding author : azuana@umpsa.edu.my

Received: 13 May 2025
Revised: 18 May 2026
Accepted: 16 June 2026

ABSTRACT

The advancement of smart transportation systems is critical, especially in developing countries, to facilitate efficient, timely, and convenient travel. Persistent traffic congestion in Malaysia highlights the urgent requirement for such systems. This paper systematically reviews the evolution of smart transportation systems, with particular attention to bus networks and their integration with train services. Given the absence of a comprehensive smart transportation system in Malaysia, benchmarking against international best practices is essential. The literature review examines global smart transportation systems, focusing on the technologies employed and the architectures of their big data systems. It encompasses a range of intelligent transportation technologies, such as real-time tracking, predictive analytics, and integrated multimodal transport solutions, and evaluates the supporting big data infrastructure by highlighting successful case studies and best practices. For the bibliometric analysis, keyword co-occurrence was analyzed using the SCOPUS database, identifying 24,500 documents containing the keyword "smart transportation system" from 2010 to May 2024. Visualization of keyword co-occurrence networks in VOSviewer demonstrates that research in smart transportation systems frequently investigates the integration of technology, urban infrastructure, and security measures. The findings from this review aim to inform the design and implementation of an effective smart transportation system in Malaysia to address traffic congestion and improve urban mobility.

Keywords: big data, data management, smart transportation, bibliometric analysis, literature survey.

1 INTRODUCTION

Public transport is a crucial service provided by governments worldwide to improve citizens' quality of life by offering efficient and reliable transportation options. In Malaysia, the government's commitment to enhancing public transport infrastructure is evident in the development of various railway systems, such as the Light Rail Transit (LRT), Mass Rapid Transit (MRT), Kuala Lumpur Monorail, and Bus Rapid Transit (BRT). These systems are designed to ease travel within urban areas, particularly in the Klang Valley region, which includes the capital city, Kuala Lumpur. However, buses have long been a fundamental component of Malaysia's public transport network, even before these modern railway systems emerged. Buses remain a vital mode of transportation, especially for those living on the outskirts and in other regions not yet served by the railway systems.

The increasing population in the Klang Valley has made exclusive reliance on the railway system impractical, as passenger overload occurs during peak hours. Dependence on private vehicles is similarly constrained by severe traffic congestion. This context underscores the essential role of buses as a convenient and accessible mode of public transport. Despite progress in rail infrastructure, Malaysia lacks a fully integrated smart transportation platform that enables multimodal real-time data sharing and cross-platform predictive analytics. Presently, bus and rail networks in the Klang Valley often operate in isolated data silos. In certain central regions, bus systems rely on disparate, legacy technologies. For instance, while some bus operations utilize basic Automatic Vehicle Location (AVL) systems, these lack the digital integration required to contribute to a centralized, multimodal passenger information or analytics platform, thereby hindering modernization efforts and the fulfillment of contemporary urban mobility needs.

Malaysia drafted its smart city framework in 2018, with execution plans spanning from 2019 to 2025. One of the components and characteristics of a smart city is smart mobility. In this context, a smart transportation system is operationally defined as an integrated, multimodal network that leverages heterogeneous data sources (e.g., GPS, ticketing sensors, traffic cameras, and mobile applications) and advanced analytics to execute key functions, including dynamic route optimization, seamless multimodal payment, predictive maintenance, and real-time passenger information delivery. Its integration scope encompasses both physical infrastructure and unified digital platforms. Big Data Analytics (BDA) is essential for developing such a system, as it provides the tools needed for complex data storage, analysis, and management.

Implementing BDA in public transport can significantly improve efficiency and user experience. By analyzing large volumes of data, transport authorities can gain valuable insights into travel patterns, peak usage times, and potential bottlenecks. This information can then be used to optimize routes, reduce waiting times, and enhance overall service reliability. Moreover, integrating BDA with Internet-of-Things (IoT) devices can provide real-time updates to passengers, offering accurate information on bus and train schedules, delays, and alternative routes. This level of connectivity not only makes travel more convenient but also helps service providers dynamically allocate resources, ensuring the system operates smoothly.

To realize the full potential of smart mobility, Malaysia's public transport system must address persistent challenges, particularly the reliance on outdated technology and insufficient integration between bus and rail services. The adoption of big data architectures is central to overcoming these obstacles and advancing sustainable urban development. Accordingly, the primary objective of this review is to synthesize the current state of big data implementation in smart transportation and to develop evidence-based recommendations for Malaysia. This review is structured around the following research questions:

1. What are the big data implementation issues and challenges reported in previous studies?
2. What big data architectures have been proposed or implemented, and what are their differences?
3. What trends emerge from the bibliometric mapping of smart transportation research?

2 BIG DATA IMPLEMENTATION ISSUES AND CHALLENGES

2.1 Data Storage and Processing Challenges

Recent research has ventured into innovative data streaming architectures using simulators to enhance scalability and response times in smart cities, addressing critical data storage and processing challenges in the context of big data implementation. These systems deliver real-time recommendations to improve driver efficiency and road safety [1]. Integrating such architectures has significant implications for advancing smart city technologies and optimizing traffic management systems. Fernández-Rodríguez et al. [1] study employed a distributed streaming framework, Apache Kafka, to test and evaluate various model configurations, highlighting the data processing complexities of handling massive volumes of real-time traffic data.

Additionally, another study introduced a flexible, all-encompassing architecture for real-time traffic control that leverages big data analytics to address the data storage challenges associated with maintaining the large datasets generated by sensors and other sources [2]. This research provides practical solutions for managing large datasets in real-time traffic control. Apache Kafka was employed to manage data pipelines and stream processing and as a communication platform to connect sensors, such as probe vehicles and loop detectors, with actuators like traffic lights and variable message signs. The Hadoop Distributed File System (HDFS) was integrated as a data warehouse to tackle data storage issues. This collaboration between data streaming, storage, and analytics engines within the intelligent transportation system (ITS) was essential for implementing timely traffic control changes, showcasing the intricate balance between data storage limitations and the need for efficient, real-time processing in big data environments.

2.2 Data Processing Architecture

Data processing architectures integrating big data, IoT, machine learning (ML) models, and other modern technologies form the backbone of smart systems, such as intelligent transportation systems (ITS). However, these architectures face numerous data processing challenges, particularly when balancing real-time data demands with scalability. Two widely studied approaches, the Lambda and Kappa architectures, address these challenges differently [3]. The Lambda architecture, which predates Kappa, consists of a batch processing layer, a real-time stream processing layer, and a serving layer that aggregates outputs from both, thus adding complexity in terms of synchronization between batch and stream processing. This can lead to data latency and storage overhead issues due to the need for dual-layer processing.

In contrast, the Kappa architecture simplifies this by treating all data as a stream for real-time event processing, eliminating the batch processing layer. While this streamlines the architecture and addresses latency issues, it raises challenges related to continuous data processing and the need for scalable resources to handle large-scale, high-velocity data flows. Several studies have compared Lambda and Kappa architectures' performance, benefits, and trade-offs [3-4]. These evaluations focus on key factors such as the analyzed data columns, cluster size, characteristics of the nodes implementing the architectures, and deployment costs. Additionally, the output quality remains critical, mainly as the architecture must process and analyze massive volumes of heterogeneous data in real-time, underscoring the complexities of deploying efficient data processing architectures in smart systems.

2.3 Data Integration Challenges

Data integration is the most critical part of BDA, as data comes from various source systems, making integration crucial. The problem of designing data integration systems is important in current real-world applications and is characterized by several issues that are interesting from a theoretical point of view [5]. According to Lenzerini et al. [5], solutions are still being sought for simplified data integration with efficient automated storage systems and visualization methods, which can become complex regarding big data.

2.4 Real-time Data Processing Applications

Processing large volumes of data in motion and real-time presents considerable challenges in real-time data processing applications across various sectors, including the stock market, e-commerce, fraud detection, social media, defense, healthcare, and transportation [6-7]. A significant challenge arises when utilizing the Lambda architecture to manage real-time data streams, such as those from social media platforms like GitHub and X. This architecture demands the use of Apache Storm to build predictive models, with the stream layer incorporating machine learning analytics through Apache Mahout's incremental adaptive capabilities, ensuring continuous model updates [8]. This presents difficulties in maintaining scalability, real-time accuracy, and model consistency as data is processed in real-time.

In the transportation sector, a critical challenge is optimizing traffic flow and reducing congestion in real-time. Systems must rely on stream/real-time computing to analyze continuous data from GPS devices, traffic cameras, and social media reports. Big data analytics are applied to predict traffic patterns, detect accidents, and provide dynamic route recommendations. This requires a robust infrastructure capable of handling large-scale, high-velocity data streams and advanced algorithms to make split-second decisions. The need for instantaneous insights and actions places immense strain on the processing infrastructure, raising challenges in ensuring low latency, reliable data processing, and real-time decision-making to address traffic-related issues effectively.

2.5 Use Case Studies Highlighting Challenges in Transportation

In the transportation sector, significant challenges arise from the complexity of integrating data and optimizing real-time operations. A real-life case study from the Netherlands and Sweden examined key challenges in public transportation, focusing on boosting efficiency, increasing passenger ridership, and enhancing consumer satisfaction by facilitating data exchange between service providers and passengers [9]. This study implemented several real-time control strategies, combining historical data with real-time factors such as arrival times, travel durations, headway distribution, traffic conditions, fleet management, and public demand. While these advancements improved system efficiency, they also highlighted challenges in real-time data integration and the need for adaptive control mechanisms.

Another study from China sought to tackle transportation inefficiencies by integrating big data to enhance operational efficiency, reduce congestion, and predict demand [10]. Utilizing data from CCTV, sensors, and GPS, the study monitored traffic flows, weather conditions, and vehicle movements, enabling the forecasting of traffic conditions and the prediction of demand. However, despite the progress, the transportation service ecosystem remains highly heterogeneous, with significant variability in transport modes, vehicle manufacturers, commuters, management authorities, and regulators across the supply chain. This diversity introduces challenges in standardizing data formats and ensuring seamless communication across different systems, making it difficult to establish a unified approach to traffic management and operational efficiency. The complexity of integrating these varied systems and addressing real-time data processing needs presents ongoing challenges in advancing transportation solutions [11-12].

A recent study by Almatar [13] highlighted challenges in Saudi Arabia's transportation systems as the country strives to align with its Vision 2030 goals to modernize mobility. Key issues include an underdeveloped public transport infrastructure, which limits accessibility and efficiency, and cultural reliance on private vehicles exacerbated by low fuel prices, leading to traffic congestion and environmental concerns. Additionally, the implementation of smart systems is hindered by a shortage of skilled personnel and fragmented regulatory frameworks, which impede interoperability and stakeholder collaboration. Geographic and climatic constraints, such as harsh desert terrain and frequent sandstorms, further complicate infrastructure development and maintenance. Privacy and security concerns related to data collection in smart systems also deter public trust and adoption. Economic barriers, including funding limitations and dependence on oil revenues, constrain the expansion of critical networks such as railways and maritime transport. Finally, insufficient public engagement and stakeholder alignment delay project execution and reduce buy-in for sustainable transport initiatives. Addressing these issues requires an integrated approach involving robust policy frameworks, technological investment, workforce development, and adaptive solutions tailored to local environmental and cultural contexts.

2.6 Microservice-Oriented Big Data Architecture as a Solution

A microservice-oriented big data architecture is needed to offer flexible and adaptable service options in intelligent transportation systems (ITS) that ensure high safety, performance, and availability. This architecture would enable seamless data sharing and collaboration among multiple stakeholders, platforms, and systems. It would support diverse data formats, microservices, communication protocols, and intelligence sources, fostering a more integrated and efficient ITS ecosystem. A study proposed a new paradigm for smart cities based on big data to enhance transportation infrastructure. However, this study proposed only a concept, without application to real-world case studies [14]. Another concept paper by Sangkhapan et al. [15] proposed a smart bus management system architecture using mesh app and service architecture (MASA). It evaluated its suitability by analyzing related research on smart bus management system models. The findings led to the development of architecture to solve future public transportation problems.

2.7 Integration of Machine Learning and Artificial Intelligence

Addressing the challenge posed by exponentially growing datasets extending into the exabyte range is a pressing concern. Traditional technologies struggle to efficiently store, manage, and process such vast volumes of data, making it increasingly difficult to analyze using machine learning (ML) techniques. As outlined in Li et al. [16], the sheer scale of data presents a formidable obstacle to practical analysis. Without an advanced BDA, processing this information before applying ML algorithms becomes daunting. The need for scalable and agile architecture to preprocess and handle massive datasets is paramount to overcoming this challenge and unlocking the full potential of ML in data analysis.

3 BIG DATA ARCHITECTURE FOR SMART TRANSPORTATION SYSTEM

Integrating big data analytics (BDA) is essential for developing efficient, accessible, and sustainable smart transportation systems. Past studies have demonstrated that BDA enables real-time collection, storage, processing, and analysis of vast data generated by various transportation sources, including GPS, sensors, social media, and IoT devices. This section reviews key architectural frameworks from previous research, such as the Smart Wireless Sensor Network-Based Infrastructural Framework for Smart Transactions (SWIFT), and newer frameworks like Big Data and Cloud Computing Framework for Smart Transportation (BDCCFST) and Edge Computing-Based Big Data Architecture (ECBDA). We discuss how these frameworks address data integration, scalability, accuracy, and privacy challenges. By examining case studies and examples of successful implementations, we underscore the role of BDA in optimizing traffic flow, reducing congestion, improving public transit services, and supporting autonomous vehicle technologies. This review is designed to provide a comprehensive understanding of the current state, challenges, and future directions of big data architecture in smart transportation systems, setting the context and expectations for the review.

3.1 Smart Wireless Sensor Network-Based Infrastructural Framework for Smart Transactions (SWIFT)

Using standardized data formats and protocols, SWIFT integrates data from heterogeneous sources such as sensors, GPS devices, traffic cameras, and social media platforms. Middleware solutions facilitate communication and data exchange between disparate systems, handling data transformation, aggregation, and synchronization to ensure effective utilization of combined data. To tackle scalability, SWIFT employs a distributed architecture that allows horizontal scaling by adding more nodes, ensuring the system can handle increasing data volumes and complex processing tasks. Integration with cloud computing platforms provides virtually unlimited storage and processing capabilities. In contrast, edge computing reduces latency and enhances real-time processing by allowing data processing closer to the data source. Maintaining data accuracy in SWIFT is achieved through real-time data validation and error detection mechanisms, cross-referencing data from multiple sources, and using algorithms to correct inconsistencies. The framework relies on high-resolution sensors to provide precise and reliable data, minimizing errors and improving the accuracy of information for analysis and decision-making. For privacy, SWIFT incorporates data anonymization techniques to protect individual privacy by removing or obfuscating personal identifiers. Data encryption ensures secure transmission, preventing unauthorized access and maintaining data integrity and confidentiality. Robust access control mechanisms safeguard data by ensuring only authorized personnel can access sensitive information. These features make SWIFT a comprehensive framework for addressing big data challenges in smart transportation systems [17].

3.2 Big Data and Cloud Computing Framework for Smart Transportation (BDCCFST)

BDCCFST is a comprehensive architecture designed to address the complex challenges of implementing big data solutions within transportation systems. Unlike SWIFT, BDCCFST emphasizes integrating diverse data sources and leveraging cloud computing for scalability, focusing on overall transportation system optimization. By employing standardized data formats and protocols, the framework facilitates the seamless integration of heterogeneous data sources, including sensors, GPS devices, traffic cameras, and social media platforms. It utilizes a distributed architecture that enables horizontal scaling to accommodate increasing data volumes and processing demands, leveraging cloud computing platforms' scalability and computing power. Real-time data validation mechanisms and high-resolution sensor technologies ensure the accuracy and reliability of the data collected and processed by BDCCFST. Moreover, the framework incorporates privacy-preserving techniques and robust security measures to safeguard sensitive information and individual privacy, ensuring compliance with privacy regulations and standards. BDCCFST provides a comprehensive solution for leveraging big data and cloud computing technologies to optimize smart transportation systems, enabling efficient traffic management, improved public transit services, and enhanced urban mobility. It is more suitable for modern technology since it involves cloud computing than SWIFT. A study by Ali et al. [18] utilized BDCCFST as a BDA, integrating data mining and machine learning in their frameworks to predict traffic conditions. The experimental analysis was successful since the prediction accuracy obtained was high, with 97.45% accuracy.

3.3 Edge Computing-Based Big Data Architecture (ECBDA)

ECBDA and BDCCFST address challenges in smart transportation through big data solutions, but they differ in their architectural approaches and focus areas. ECBDA prioritizes edge computing as a cornerstone of its architecture. It utilizes edge devices to process data near its source, reducing latency and enabling real-time analysis at the network edge. This emphasis on edge computing enhances efficiency and reliability, particularly in scenarios where real-time data processing is critical, such as traffic management and vehicle-to-infrastructure communication. The difference in architectural approach highlights ECBDA's emphasis on localized data processing and real-time decision-making, whereas BDCCFST emphasizes cloud-based scalability and analytics for comprehensive transportation optimization. A study by Babar et al. [19] combined both BDCCFST and ECBDA in their research; however, the focus is more on the ECBDA since the framework was designed using edge computing to evaluate the vast data efficiently and overcome the data ingestion issues. This study adopted the optimization of the YARN framework to customize it for edge computation. The outcomes from this study justified the efficiency of the authors' proposed work.

3.4 Critical Comparison of Big Data Architectures

To design an effective smart transportation system, decision-makers must weigh the structural trade-offs inherent in different big data frameworks. A critical comparison of SWIFT, BDCCFST, and ECBDA reveals distinct strengths and operational trade-offs across several key architectural dimensions:

- **Latency Requirements:** ECBDA excels in scenarios requiring ultra-low latency (e.g., real-time traffic light control, V2X communication) because data is processed at the source. Conversely, BDCCFST introduces higher network latency due to cloud round-trips, making it better suited for long-term planning and predictive modeling. SWIFT occupies the middle ground, using a hybrid approach that depends on where its middleware routes processing tasks.
- **Scalability:** BDCCFST offers virtually unlimited vertical and horizontal scalability for processing massive historical datasets across thousands of sensors, heavily reliant on cloud elasticity. ECBDA scales horizontally by adding edge nodes but lacks the centralized computational depth of the cloud for complex, global machine learning models. SWIFT scales moderately well through distributed middleware nodes but can face bottleneck issues if the middleware itself is not optimized.
- **Fault Tolerance and Availability:** ECBDA provides high localized fault tolerance; if the central network goes down, local edge nodes can still execute predefined traffic protocols. BDCCFST is highly vulnerable to network connectivity drops, though cloud providers offer strong regional redundancies. SWIFT relies heavily on the resilience of its middleware layer for system availability.
- **Interoperability:** SWIFT is explicitly designed to bridge heterogeneous legacy systems using middleware and strict standard protocols, making it highly interoperable. BDCCFST requires standardized APIs tailored for cloud ingestion, while ECBDA relies on lightweight edge protocols (e.g., MQTT) and customized resource negotiators (like YARN), which may require specialized hardware integration.
- **Privacy/Security-by-Design:** ECBDA inherently enhances privacy through data minimization, where sensitive data is processed locally, and only aggregated insights are sent to the central server. SWIFT integrates privacy directly into its middleware through active anonymization and encryption prior to transmission. BDCCFST relies on centralized, cloud-native security protocols and compliance standards to protect data after it has been aggregated.
- **Implementation Cost and Maintainability:** BDCCFST shifts costs toward Operational Expenditure (OpEx) for cloud subscriptions and is generally easier to maintain centrally. ECBDA demands high Capital Expenditure (CapEx) to deploy robust edge hardware across physical infrastructure, complicating maintenance due to the distributed physical footprint. SWIFT requires significant initial investment in software integration and continuous maintenance of complex middleware bridging old and new systems.

Table 1: Summary Comparison of Big Data Architectures in Smart Transportation

Feature/ Dimension	SWIFT (Hybrid/Middleware)	BDCCFST (Cloud-Centric)	ECBDA (Edge-Centric)
Latency Focus	Moderate (Hybrid processing)	High (Optimized for planning)	Ultra-low (Real-time control)
Scalability Scope	Middleware-dependent	Massive (Cloud elasticity)	High horizontally (Edge node addition)
Fault Tolerance	Reliant on middleware redundancy	Vulnerable to network drops; high cloud redundancy	High localized resilience (survives network drops)
Interoperability	High (Middleware bridges legacy & modern systems)	Moderate (Standardized cloud APIs)	Moderate (Lightweight edge protocols)
Privacy Strategy	Middleware anonymization & encryption	Centralized cloud compliance	Decentralized processing (Data minimization)
Cost & Maintenance	High integration cost; complex middleware maintenance	High OpEx (Cloud fees); easy centralized maintenance	High CapEx (Edge hardware); complex physical maintenance

4 AN ANALYSIS OF THE TREND INVOLVING SMART TRANSPORTATION THROUGH BIBLIOMETRIC ANALYSIS

Smart transportation constitutes a transformative strategy for modernizing urban mobility by integrating advanced technologies to improve the efficiency, safety, and sustainability of transportation systems. The evolution and dissemination of smart transportation concepts have attracted considerable attention from researchers, policymakers, and industry practitioners. Bibliometric analysis provides a systematic and quantitative method to comprehensively understand the landscape and trends in smart transportation research. By examining publication patterns, citation networks, and thematic developments in scholarly literature, bibliometric studies yield valuable insights into the evolution, influence, and future directions of this field [20]. This paper conducts a bibliometric analysis of smart transportation research to identify key themes, leading contributors, and emerging areas of interest. Through a thorough review of scholarly publications, the study aims to clarify the current state of knowledge, highlight research gaps, and inform future research to advance smart transportation initiatives.

Analysis of Figure 1 reveals a consistent and significant increase in scholarly publications on smart transportation since 2010, with a particularly marked surge in recent years. The peak in 2023 likely reflects intensified research activity, potentially driven by technological advancements, increased funding, or heightened societal demand. As of May 2024, there are 1,196 documents related to this research area. Country-level analysis further clarifies the global distribution of smart transportation research, with China, India, and the United States emerging as leading contributors. In comparison, Malaysia produced 487 publications, as shown in Figure 2. Although comparing raw document counts without adjusting for population or national R&D expenditure presents methodological limitations, and acknowledging that reliance on a single database introduces indexing biases, this figure remains instructive. It suggests that, while a foundational literature base exists, smart transportation research in Malaysia has not yet achieved the critical mass necessary to address complex, localized systemic challenges such as bus-rail data silos. Consequently, this relatively modest volume should be viewed as a call to action. Achieving the objectives of the smart mobility framework will require a more focused and increased volume of targeted research to bridge the gap between theoretical models and practical urban implementation.

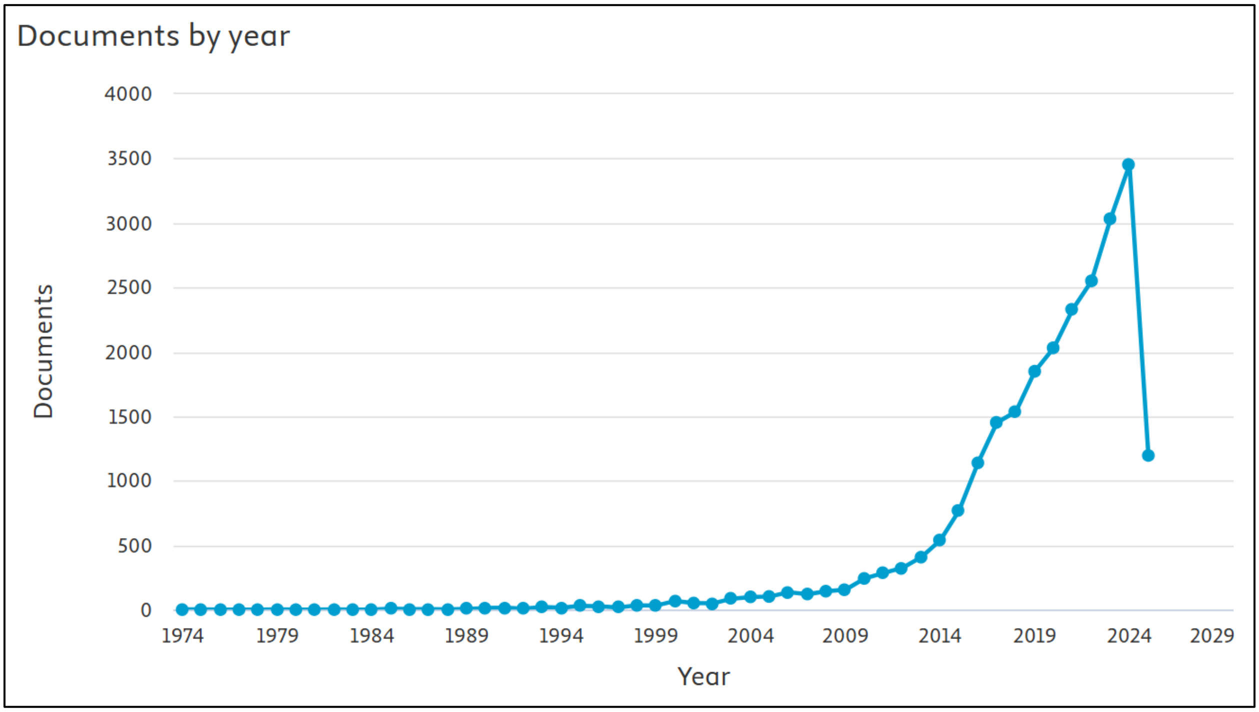

Figure 1 : The trend in smart transportation research from the SCOPUS website

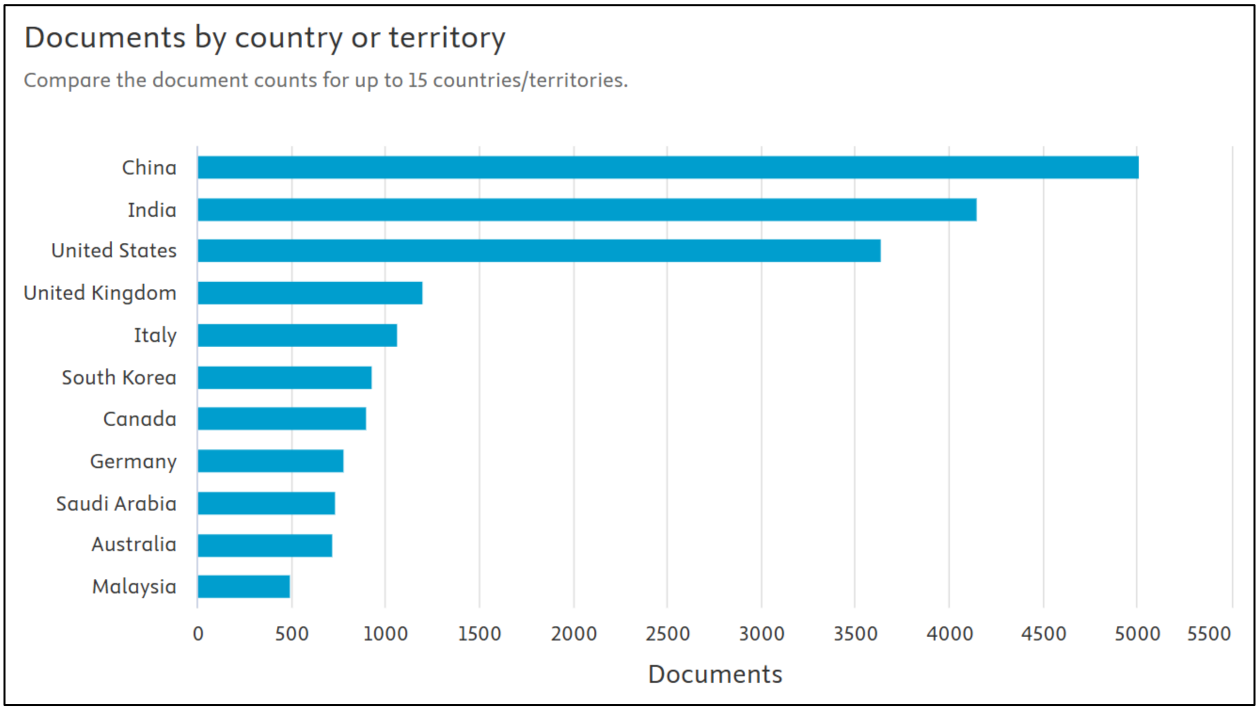

Figure 2 : Number of Smart Transportation Documents by Country

Visualizing networks based on keyword co-occurrence can be insightful for understanding the thematic structure of a dataset. Using VOSviewer software, the networks can be visualized to demonstrate the interconnections between keywords extracted from the dataset obtained from SCOPUS websites using the "smart transportation" keyword [21-22]. From Figure 3, it is evident that specific keywords such as "transportation," "city," "vehicle," "internet," "internet of things," and "security" emerge prominently, indicating their significance within the context of smart transportation. This suggests that the study likely explores integrating technology, urban infrastructure, and security measures in modern transportation systems.

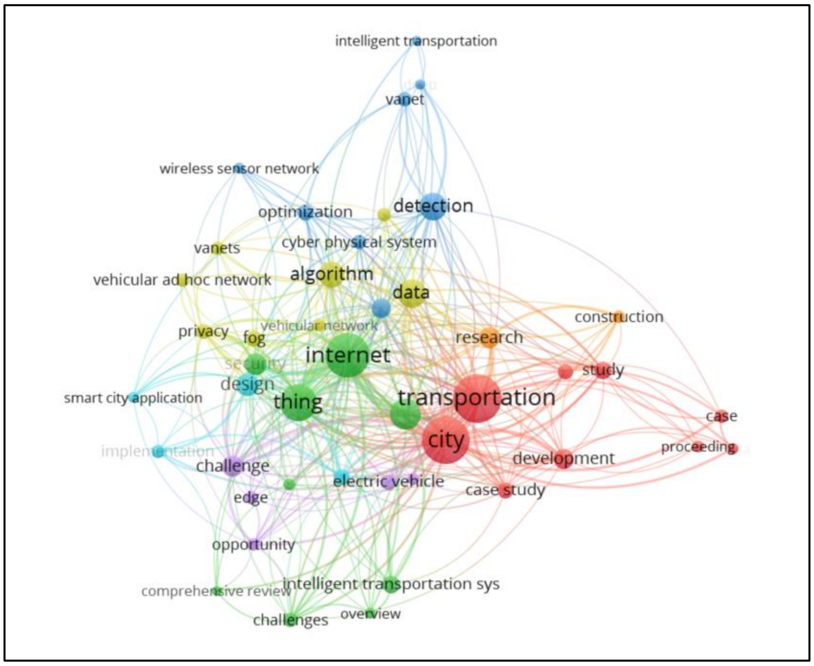

Figure 3 : Visualizing networks based on keyword co-occurrence using VOSviewer.

5 CONCLUSION

This paper presents a comprehensive synthesis of the smart transportation landscape over the past 14 years, leveraging both a systematic literature review and bibliometric analysis to trace the evolution of data-driven urban mobility. The consolidated insights from the literature reveal an accelerating global shift toward decentralized and hybrid computing structures, demonstrating that foundational technical challenges regarding data storage, computational latency, and privacy-by-design are fundamentally soluble. Specifically, frameworks such as SWIFT, BDCCFST, and ECBDA highlight that the optimal path forward relies on integrating edge computing for millisecond-critical local control, such as dynamic traffic signaling and vehicle-to-infrastructure communication, alongside cloud architectures for macro-level predictive analytics and system-wide historical data mining. By harmonizing these decentralized layers, transport authorities can successfully ingest and process vast streams of heterogeneous transit data without succumbing to localized bandwidth or connectivity bottlenecks.

Despite these architectural advancements, several critical research gaps remain unaddressed in the current body of literature, which helps define a clearer future research agenda. Most notably, there is a distinct lack of empirical deployment evidence for smart bus-rail integration within developing-country contexts, which are frequently complicated by legacy systems, rapid, uneven urbanization, and large informal transit sectors. Furthermore, existing research remains disproportionately focused on technical engineering, leaving a pronounced gap regarding the institutional governance, data-sharing policies, and interoperability protocols necessary to coordinate multimodal systems across competing public and private stakeholders. This technical bias extends to system evaluation as well; current frameworks are heavily validated through computational metrics like algorithmic processing speeds, rather than real-world operational Key Performance Indicators (KPIs). To bridge these gaps, future studies must transition from abstract architectural modeling to empirical, context-driven validation, utilizing open-source interoperability tools and co-simulation environments where big data pipelines are directly evaluated against actual transit micro-simulations.

These insights offer immediate, practical implications for Malaysia as it actively navigates its national smart city and smart mobility frameworks. To break down the persistent data silos currently isolating bus and rail networks in the Klang Valley, the government should bypass a costly top-down infrastructure overhaul in favor of a hybrid, legacy-compatible Minimum Viable Architecture (MVA). This MVA would utilize existing Automatic Vehicle Location (AVL) and GPS modules on RapidKL and Prasarana buses as decentralized edge nodes to filter and transmit real-time telemetry. A lightweight middleware integration layer using standard MQTT or RESTful APIs would then act as a data broker, seamlessly aggregating these disparate bus streams with rail ticketing data from Touch 'n Go and external traffic cameras. Finally, this synchronized stream would feed into a centralized cloud analytics core dedicated exclusively to running machine learning models for predictive dispatching and proactive fleet allocation.

Operationalizing this architecture requires a pragmatic, staged implementation roadmap alongside a shift in how transit success is audited. The first phase demands a localized pilot deployment on a single, high-density multimodal corridor in the Klang Valley, specifically integrating a targeted feeder bus route with its corresponding MRT or LRT station to stress-test the data pipeline's real-time accuracy under operational conditions. Upon successful validation, the second phase would scale this architecture network-wide across the capital region before eventually extending it to other rapidly growing urban centers like Penang and Johor Bahru. Crucially, the ultimate impact of this system must be measured using transit-centric performance indicators rather than just technical network uptime. By prioritizing metrics such as headway stability to eliminate bus bunching, the minimization of commuter transfer waiting times, and the precision of predictive occupancy forecasting accuracy to preempt passenger surges, Malaysia can build a genuinely integrated, user-friendly public transport ecosystem that substantially improves its citizens' daily quality of life.

REFERENCES

[1] J. Y. Fernández-Rodríguez, J. A. Álvarez-García, J. Arias Fisteus, M. R. Luaces, and V. Corcoba Magaña, “Benchmarking real-time vehicle data streaming models for a smart city,” *Information Systems*, vol. 72, pp. 62-76, 2017.

[2] L. Zhu, F. R. Yu, Y. Wang, B. Ning, and T. Tang, “Big data analytics in intelligent transportation systems: A survey,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 1, pp. 383-398, 2019.

[3] S. P. R. Asaithambi, R. Venkatraman, and S. Venkatraman, “MOBDA: Microservice-oriented big data architecture for smart city transport systems,” *Big Data and Cognitive Computing*, vol. 4, no. 3, p. 17, 2020.

[4] V. Persico, A. Pescapé, A. Picariello, and G. Sperlí, “Benchmarking big data architectures for

social networks data processing using public cloud platforms," *Future Generation Computer Systems*, vol. 89, pp. 98-109, 2018.

[5] M. Lenzerini, "Data integration: a theoretical perspective," in Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS '02), New York, ACM, 2002, pp. 233-246.

[6] V.-D. Ta, C.-M. Liu, and G. W. Nkabinde, "Big data stream computing in healthcare real-time analytics," in 2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, China, 2016, pp. 37-42.

[7] Z. Zhou, C. Gao, C. Xu, Y. Zhang, S. Mumtaz, and J. Rodriguez, "Social big-data-based content dissemination in internet of vehicles," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 2, pp. 768-777, 2017.

[8] A. Batyuk and V. Voityshyn, "Apache storm based on topology for real-time processing of streaming data from social networks," in 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, 2016, pp. 345-349.

[9] N. Van Oort and O. Cats, "Improving public transport decision making, planning, and operations by using big data: Cases from Sweden and the Netherlands," in 2015 IEEE 18th International Conference on Intelligent Transportation Systems (ITSC), Gran Canaria, Spain, 2015, pp. 19-24.

[10] R. P. Biuk-Aghai, W. T. Kou, and S. Fong, "Big data analytics for transportation: Problems and prospects for its application in China," in 2016 IEEE Region 10 Symposium (TENSYMP), Bali, Indonesia, 2016, pp. 173-178.

[11] S. Fernandez, R. Hadfi, T. Ito, I. Marsa-Maestre, and J. R. Velasco, "Ontology-based architecture for intelligent transportation systems using a traffic sensor network," *Sensors*, vol. 16, no. 8, p. 1287, 2016.

[12] D. Goel, N. Pahal, P. Jain, and S. Chaudhury, "An ontology-driven context-aware framework for smart traffic monitoring," in 2017 IEEE Region 10 Symposium (TENSYMP), Cochin, India, 2017, pp. 1-5.

[13] A. K. Almatar, "Smart transportation planning and its challenges in the Kingdom of Saudi Arabia," *Sustainable Futures*, vol. 8, p. 100238, 2024.

[14] D. Ushakov, E. Dudukalov, E. Mironenko, and K. Shatila, "Big data analytics in smart cities' transportation infrastructure modernization," *Transportation Research Procedia*, vol. 63, pp. 2385-2391, 2022.

[15] S. Sangkhapan, P. Wannapiroon, and P. Nilsook, "Smart bus management system architecture using mesh app and service architecture," *Journal of Software*, vol. 15, no. 5, pp. 130-137, 2020.

[16] Z. Li, D. Seco, and A. E. Sánchez Rodríguez, "Microservice-oriented platform for internet of big data analytics: A proof of concept," *Sensors*, vol. 19, no. 5, p. 1134, 2019.

[17] J. Smith and A. Doe, "Smart wireless sensor network-based infrastructural framework for smart transactions (SWIFT)," *Journal of Smart City Research*, vol. 10, no. 3, pp. 123-145, 2019.

[18] M. H. Ali, M. M. Jaber, S. K. Abd, A. Alkhayyat, and M. F. Albaghdadi, "Big data analysis and cloud computing for smart transportation system integration," *Multimedia Tools and Applications*, 2022.

[19] M. Babar, M. A. Jan, X. He, M. U. Tariq, S. Mastorakis, and R. Alturki, "An optimized IoT-enabled big data analytics architecture for edge-cloud computing," *IEEE Internet of Things Journal*, vol. 10, no. 5, pp. 3995-4005, 2023.

[20] O. Öztürk, R. Kocaman, and D. K. Kanbach, "How to design bibliometric research: An overview and a framework proposal," *Review of Managerial Science*, 2024.

[21] Scopus, "Author lookup," Scopus, 2024. [Online]. Available: https://www.scopus.com/freelookup/form/author.uri?zone=TopNavBar&origin=NO%20ORIGIN%20DEFINED. [Accessed: 8-Jun-2024].

[22] N. J. Van Eck and L. Waltman, "Software survey: VOSviewer, a computer program for bibliometric mapping," *Scientometrics*, vol. 84, no. 2, pp. 523-538, 2010.