مجلـــة كليـــة التربيـــة طرابلـــس
Journal of the Faculty of Education Tripoli

البيانات الضخمة وشبكات المركبات المعتمدة على الحوسبة السحابية: دمج التحليلات، والذكاء الطرفي، والتحسين القائم على البيانات

[1]Emadeddin A. Gamati
[1] Department of Computer Science, Faculty of Education, The University of Tripoli, Tripoli, Libya
[1]e.gamati@uot.edu.ly

Article history
Received: Month Nov, 2025 Accepted: Month Dec,2025

الملخص:

أدى تطور الشبكات المخصّصة للمركبات (VANETs) إلى منظومات قائمة على البيانات - إلى توليد كميات هائلة وغير مسبوقة من البيانات الناتجة عن أجهزة الاستشعار، والمركبات، ووحدات الطريق. إن إدارة هذه البيانات الضخمة للمركبات، ونقلها، واستثمارها بكفاءة تُعدّ عناصر أساسية لتحقيق رؤية التنقل الذكي والمتصل. يقدم هذا البحث دراسة شاملة حول شبكات المركبات المعتمدة على البيانات الضخمة (BDVNs)، وهو نموذج ناشئ يدمج بين الحوسبة السحابية، والذكاء الطرفي(Edge Intelligence) ، والتحليلات المعتمدة على الذكاء الاصطناعي بهدف تحسين الاتصال ،والتخزين، واتخاذ القرار. ويتناول البحث نماذج جمع البيانات، وأطر تكامل الحوسبة السحابية والطرفية، ومناهج التعلم الموزع مع التركيز على قابلية التوسع، وزمن الاستجابة، والخصوصية. وأظهرت نتائج التقييم المقارن أن البنى الهجينة القائمة على الدمج بين السحابة والطرف تقلل زمن التأخير الشامل بنسبة تقارب 35%، وتزيد كفاءة معالجة البيانات بنحو 40% مقارنة بالأنظمة المعتمدة على السحابة فقط. وأخيرًا، يسلط البحث الضوء على التحديات البحثية المفتوحة في مجالات التوحيد القياسي، والتشغيل البيني، وإدارة الثقة، ممهّدًا الطريق نحو شبكات مركبات ذاتية التنظيم مدعومة بالبيانات الضخمة في عصر الجيل السادس.(6G)

الكلمات المفتاحية: شبكات VANET، تحليلات البيانات الضخمة، الحوسبة السحابية، الذكاء الطرفي، الحوسبة الضبابية، التعلّم الاتحادي، إدارة البيانات، شبكات G V2X6.

Big Data and Cloud-Driven Vehicular Networks: Integrating Analytics, Edge Intelligence, and Data-Centric Optimization

A B S T R A C T:

The evolution of Vehicular Ad Hoc Networks (VANETs) into data-centric ecosystems has generated unprecedented data volumes from sensors, vehicles, and roadside units. Managing, transmitting, and exploiting this **vehicular big data** efficiently are essential for realizing the vision of smart and connected mobility. This paper presents a comprehensive study of **Big-Data-Driven Vehicular Networks (BDVNs)**—an emerging paradigm that merges cloud computing, edge intelligence, and AI-based analytics to optimize communication, storage, and decision-making. We examine data-collection models, cloud/edge integration frameworks, and distributed learning approaches, emphasizing scalability, latency, and privacy. Comparative evaluation shows that hybrid cloud-edge architectures reduce end-to-end delay by ≈ 35 % and increase data-processing efficiency by ≈ 40 % relative to cloud-only baselines. Finally, we identify open research challenges in standardization, interoperability, and trust management, paving the way for **autonomous big-data-enabled vehicular networks** in the 6G era.

Keywords— VANET, big data analytics, cloud computing, edge intelligence, fog computing, federated learning, data management, 6G V2X.

I. Introduction

Vehicular Ad Hoc Networks (VANETs) have evolved far beyond their initial purpose as simple broadcasting systems used to exchange safety-related messages. With the rapid integration of advanced sensing technologies—such as high-resolution cameras, LiDAR, radar, GPS, and onboard diagnostics—modern vehicles now operate as mobile data generators, producing gigabytes of heterogeneous information every hour. When aggregated across large vehicular fleets, these data streams form a rich and continuous source of real-time knowledge that can support traffic optimization, hazard detection, environmental assessment, and predictive mobility management [1].

However, traditional VANET architectures were not designed to handle such massive and fast-evolving data flows. Their decentralized nature limits the ability to process, store, and analyze data at scale, especially when low latency and high reliability are required. The emergence of Big Data technologies—distributed storage systems, scalable cloud platforms, and advanced analytics—has introduced new opportunities to transform VANETs into intelligent, data-centric ecosystems capable of supporting complex decision-making processes [2]. Yet relying solely on cloud infrastructures exposes significant limitations, including bandwidth constraints, unpredictable wide-area latency, and the inability to respond instantly to local events.

To overcome these constraints, research has increasingly shifted toward hierarchical architectures that integrate Edge and Fog Computing with centralized cloud resources. These multi-tier models push computation and analytics closer to the data sources, allowing vehicles and roadside units to perform time-critical tasks locally while offloading heavier workloads to cloud platforms when appropriate [3]. In parallel, the adoption of Machine Learning

(ML) and Federated Learning (FL) provides new mechanisms for extracting insights from vehicular data while preserving privacy and reducing network load.

Within this context, the present study offers a structured review of the architectural paradigms, analytics models, and optimization mechanisms that define Big-Data-Driven Vehicular Networks (BDVNs). The goal is to clarify how data collection, distributed intelligence, and hybrid cloud-edge coordination collectively enable scalable and reliable vehicular communication systems capable of supporting next-generation mobility applications.

II. Data Characteristics in Vehicular Networks

Understanding the nature of data generated within vehicular environments is essential for designing efficient Big-Data-Driven Vehicular Networks (BDVNs). Unlike conventional mobile networks, vehicular systems produce continuous, heterogeneous, and rapidly changing data streams that must be processed under strict latency and reliability constraints. These characteristics shape the requirements of sensing pipelines, resource allocation strategies, and the overall communication architecture.

A. The 5 V's of Vehicular Big Data

Vehicular data is commonly described through the lens of the *five V's*, which summarize the core challenges and opportunities associated with large-scale mobility datasets [4]:

- **Volume:** Connected vehicles collectively generate terabytes of sensory readings, telemetry, and event logs on a daily basis.
- **Velocity:** Data arrives at millisecond granularity, driven by high-frequency sensors and safety beacons.
- **Variety:** Inputs include numerical telemetry, LiDAR point clouds, high-definition video, environmental readings, and textual metadata.

- **Veracity:** Measurements are subject to sensor noise, positional drift, occlusion effects, and calibration inconsistencies.
- **Value:** Extracting meaningful patterns—such as risk indicators, congestion trends, or anomaly signatures—creates actionable intelligence for real-time decision-making.

These properties demonstrate why BDVNs require sophisticated mechanisms capable of managing high-throughput, multi-format data streams without compromising responsiveness.

B. Data Lifecycle

The lifecycle of vehicular data spans several interdependent stages, beginning with onboard sensing and followed by preprocessing, transmission, aggregation, analytics, and decision feedback. Each stage not only transforms the data but also influences system-wide performance. To maintain real-time responsiveness, the workflow must simultaneously minimize latency $(L_t)$and energy consumption (E_c), while ensuring that the extracted utility (U_d)meets the required threshold:

$$\min\,(L_t + \lambda E_c) s.t.\ U_d \geq U_{th}. \qquad \text{(Equation \#1)}$$

This formulation highlights the intrinsic trade-offs between speed, computational efficiency, and information quality. By quantifying these relationships, designers can evaluate alternative processing pipelines and determine which tasks should be allocated to vehicles, edge nodes, or cloud servers.

C. Challenges

restricts the amount of data that can be continuously uploaded, especially in dense environments where hundreds of vehicles share the same spectrum. Mobility introduces intermittent connectivity, leading to fluctuating link quality and variable end-to-end performance. Furthermore, privacy considerations—such as location confidentiality and sensor-data sensitivity—limit the feasibility of centralizing raw data.

These factors underscore the need for **hierarchical data-processing strategies**, where computational tasks are distributed across the vehicle → edge → cloud continuum [5]. By delegating low-latency decision functions to edge nodes and offloading long-term analytics to the cloud, BDVNs can maintain operational efficiency even under unpredictable network conditions. This multi-layer structure forms a foundational principle for scalable, privacy-conscious vehicular intelligence.

III. Big Data Architecture for VANETs

The architectural design of Big-Data-Driven Vehicular Networks (BDVNs) plays a central role in determining how efficiently data is collected, processed, and disseminated. As vehicular systems transition from simple message broadcasting to large-scale data analytics, it becomes necessary to adopt multi-layer frameworks capable of balancing latency, computational load, and communication overhead. This section outlines the major architectural paradigms—cloud-centric, fog-based, and edge-assisted—and explains how they operate together to support scalable vehicular intelligence.

A. Cloud-Centric Frameworks

Cloud computing provides elastic scalability for storage and analytics. Early BDVN frameworks stored vehicular logs in centralized Hadoop or Spark clusters [6]. Yet cloud latency (≥ 100 ms) and limited backhaul capacity make such systems unsuitable for safety-critical functions.

B. Edge and Fog Layers

Edge nodes located at RSUs perform local filtering and inference. Fog nodes coordinate groups of edges for regional aggregation. The **Cloud–Fog–Edge Tri-Layer Model** [7] distributes data as follows:

- **Edge:** immediate decision support, minimal latency.
- **Fog:** contextual aggregation and synchronization.
- **Cloud:** global analytics and long-term learning.

C. Data-Flow Optimization

Efficient distribution of data across the cloud–fog–edge hierarchy requires structured optimization rather than arbitrary assignment. The following objective function models how workloads should be balanced:

$$\min_{x_i} \sum_i (L_i + \alpha C_i),\text{s.t.} \sum_i x_i = D_{total} \qquad \textbf{(Equation \#2)}$$

where L_i represents the latency at layer i, C_i the processing or transmission cost, and x_i the portion of data routed to that layer.

Hybrid scheduling mechanisms built on this formulation have demonstrated up to **30% reductions** in response time compared with static allocation strategies [8].

To clarify its purpose, this optimization model captures how BDVNs can dynamically adapt to fluctuating vehicle density, link conditions, and computational loads. It provides a quantitative foundation for deciding which data should be processed locally at the edge and which should be escalated to fog or cloud layers. By grounding the architecture in formal optimization, BDVNs can maintain consistent performance even during periods of congestion or mobility-induced instability.

D. Edge-Assisted Data Caching

Caching plays a vital role in BDVNs by reducing redundant transmissions and decreasing access delays for frequently requested data such as map segments, traffic states, or predictive analytics outputs. Placing caches at the edge allows vehicles to retrieve critical information without relying on distant cloud servers.

Recent reinforcement-learning (RL)-based caching policies dynamically adjust cache contents according to observed vehicle mobility and request patterns. Studies report a **20% reduction** in cache-miss ratios and notable improvements in spectrum efficiency when RL policies are applied [9]. This demonstrates that intelligent caching is not merely an auxiliary function but a core enabler of low-latency, data-centric vehicular communication.

IV. Data Analytics and Learning Frameworks

A. Centralized Analytics

Cloud platforms such as Spark or Flink handle large-scale batch and stream analytics for traffic forecasting, route optimization, and anomaly detection [10].

B. Distributed and Federated Learning

Federated Learning enables collaborative training without raw-data sharing. Local gradients are computed on vehicles and aggregated at RSUs:

$$w_{t+1} = \sum_{i=1}^{N} \frac{n_i}{N} w_t^i. \qquad \text{(Equation \#3)}$$

FL cuts uplink load by ≈ 45 % while retaining > 95 % model accuracy [11].

C. Edge Intelligence and TinyML

TinyML models perform on-board inference for collision prediction and driver-behavior recognition within < 20 ms [12].

D. Reinforcement and Deep Learning for Data Routing

Deep-RL agents learn optimal forwarding and scheduling policies, achieving 25–40 % higher PDR under dense traffic [13].

E. Privacy-Preserving Analytics

Differential privacy and homomorphic encryption protect location data; hybrid encryption adds < 8 % overhead [14].

V. Comparative Evaluation and Performance Analysis

A. Comparative Overview

The comparative results summarized in the following table (Table 1) are derived from synthesizing findings reported in recent BDVN and cloud–edge integration studies.

Specifically, delay and processing-gain values reflect averaged outcomes from controlled simulation environments in prior research, including fog–cloud coordination frameworks [7], hybrid vehicular-cloud load-balancing models [8], and reinforcement-learning-based edge caching evaluations [9]. These studies adopt similar experimental conditions—5×5 km urban grids, 400–600 vehicles, and 10 Hz beaconing—making their results sufficiently comparable for aggregated presentation. The objective of the table is therefore not to report new measurements, but to unify representative performance indicators from established literature to illustrate the relative behavior of cloud, fog, edge, and hybrid architectures. Such consolidated comparison supports a clearer understanding of architectural trade-offs and highlights why hybrid designs consistently outperform single-layer configurations across latency, scalability, and privacy dimensions.

Architecture	**Delay (ms)**	**Processing Gain (%)**	**Scalability**	**Privacy**	**Limitation**
Cloud-only	> 100	Baseline	High	Low	Latency bottleneck
Fog-only	70	+20	Medium	Medium	Limited reach
Edge-only	40	+25	Low	High	Local isolation
Cloud + Edge	30–40	+40	High	High	Coordination overhead
Architecture	Delay (ms)	Processing Gain (%)	Scalability	Privacy	Limitation

Table1 : Comparative Evaluation and Performance

setup: 5 × 5 km grid, 500 vehicles, 10 Hz beaconing, IEEE 802.11p/5G NR.

C. Quantitative Case Studies

- **Case 1 – Cloud vs Hybrid:** Hybrid fog-cloud reduced delay from 92 → 58 ms, boosting throughput 37 % [7].
- **Case 2 – Federated Edge Analytics:** HFL-VNet achieved 96 % accuracy with 45 % less overhead [11].
- **Case 3 – RL Caching:** Hit ratio improved 0.73 → 0.88, backhaul load -22 % [9].
- **Case 4 – Digital Twins:** Predictive twins cut packet loss 18 % [16].

D. Analytical Discussion

The synthesis of empirical evidence indicates that **distributed intelligence at the edge** provides the greatest performance leverage. Hybrid cloud-edge orchestration achieves near-linear scalability until edge saturation, maintaining latency below 40 ms. RL caching and FL analytics jointly reduce backhaul traffic 30–50 %, enhancing spectral and energy efficiency. However, synchronization among heterogeneous nodes remains challenging under mobility and intermittent connectivity. Additionally, **context-aware prioritization**—transmitting only semantically relevant data—emerges as a key design principle. Employing **semantic compression and adaptive filtering** ensures that bandwidth serves high-value information, thereby improving both the reliability and interpretability of vehicular analytics. Future BDVNs must adopt intelligent relevance-driven communication policies rather than volume-driven transmission.

VI. Integration with 6G and Emerging Technologies

A. 6G-Enabled Big-Data V2X

6G introduces THz communication, AI-native control, and sub-millisecond URLLC, enabling real-time cooperation among vehicles. These capabilities support high-volume sensor streaming and predictive safety functions that cloud-only systems cannot meet. As such, 6G becomes the foundation for scalable, data-intensive V2X intelligence [17].

B. Blockchain and Data Trust

Lightweight blockchain consensus enables secure message validation with delays under 10 ms, strengthening provenance and authenticity in vehicular communication. This ensures protection against spoofing and false alerts while supporting privacy-controlled data sharing. Its integration enhances trust across multi-stakeholder mobility ecosystems [18].

C. Digital Twins and Semantic V2X

Digital-twin models mirror real vehicles and infrastructure to predict traffic states and network conditions proactively. Semantic V2X reduces communication load by transmitting essential meaning instead of raw data, improving spectral efficiency. Together, they shift vehicular systems from reactive to context-aware, anticipatory operation [19].

D. Integration Challenges

Combining 6G, blockchain, digital twins, and semantic communication faces interoperability gaps across heterogeneous protocols. Consensus mechanisms and semantic processing can conflict with URLLC timing requirements. Standardized interfaces and adaptive scheduling are needed to achieve seamless, secure, and coordinated deployments.

VII. Research Challenges and Future Directions

Despite substantial progress through hybrid cloud–edge architectures and intelligent analytics, critical challenges continue to define the BDVN research agenda.

Data governance and interoperability persist as fundamental issues. The lack of common vehicular-data ontologies and exchange formats limits large-scale cooperation among manufacturers, infrastructure operators, and regulators. International adoption of unified metadata frameworks (e.g., ISO 23793, ETSI ITS-G5 extensions) will be vital.

Scalable federated learning (FL) must evolve toward adaptive aggregation, asynchronous updates, and incentive mechanisms that ensure fairness across thousands of mobile nodes. Model-compression and knowledge-distillation techniques can mitigate communication overheads.

Energy-efficient edge computing requires the deployment of renewable-powered RSUs and resource-aware scheduling to align with sustainable ICT goals.

Cybersecurity and privacy remain paramount: BDVNs must integrate blockchain-anchored trust, differential-privacy mechanisms, and intrusion-detection algorithms resilient to data poisoning.

Finally, **city-scale digital-twin testbeds**—fusing real and simulated traffic—are crucial for validating algorithms under realistic conditions. Addressing these interdisciplinary challenges will define the roadmap toward scalable, secure, and sustainable vehicular-data ecosystems.

VIII. Conclusion

The transformation of VANETs into **Big-Data-Driven Vehicular Networks (BDVNs)** marks a milestone in intelligent transportation evolution. By uniting cloud scalability with edge immediacy, BDVNs enable real-time analytics, predictive control, and cooperative intelligence across vast vehicular infrastructures. Comparative evaluation in this study demonstrates that hybrid cloud–edge designs yield up to 40 % greater processing efficiency and ≈ 35 % latency reduction versus cloud-only counterparts. Integrating **Federated Learning (FL)** and **Reinforcement Learning (RL)** further enhances adaptability while maintaining privacy and data integrity.

Nevertheless, maturity requires addressing persistent gaps: harmonizing data standards, ensuring explainable AI, and embedding trust mechanisms via blockchain and secure edge orchestration. The forthcoming integration of **6G URLLC**, **semantic V2X**, and **digital-twin synchronization** will redefine vehicular communication by achieving sub-millisecond latency and proactive congestion prevention.

Ultimately, BDVNs represent the foundation for **autonomous, self-organizing, and trustworthy vehicular ecosystems**—where vehicles and infrastructure continuously sense, learn, and adapt collaboratively. Through the convergence of big-data analytics, distributed intelligence, and sustainable design, future transportation systems will become safer, greener, and more efficient, supporting the global transition to connected mobility in the 2030 era and beyond.

References

- X. Cheng et al., “Big Data Driven Vehicular Networks,” *IEEE Network*, vol. 32, no. 6, pp. 160–167, 2018.
- M. Gerla et al., “Internet of Vehicles: From Intelligent Grid to Autonomous Cars,” *IEEE World Forum on IoT*, pp. 241–246, 2019.

- L. Zhang and S. U. Khan, “Edge-Cloud Collaboration in Vehicular Networks,” *IEEE Commun. Surv. Tutorials*, vol. 23, no. 4, pp. 2560–2585, 2021.

- A. Tiwari et al., “Big Data Challenges in Vehicular Networks,” *IEEE Access*, vol. 9, pp. 102134–102150, 2021.

- Y. Duan et al., “Fog Computing for Smart Vehicles: A Survey,” *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 15675–15688, 2022.

- M. Rehman evolt al., “Vehicular Cloud Data Management Using Hadoop,” *J. Cloud Comput.*, vol. 10, Art. 32, 2021.

- C. Li et al., “A Fog-Cloud Collaborative Framework for Big-Data-Driven VANETs,” *IEEE Trans. Veh. Tech.*, vol. 72, no. 8, pp. 9871–9885, 2023.

- S. Hassan et al., “Dynamic Load Balancing for Hybrid Vehicular Clouds,” *IEEE Internet Things J.*, vol. 11, no. 5, pp. 7482–7495, 2024.

- K. Roy and H. Kim, “Reinforcement-Learning-Based Edge Caching for Vehicular Networks,” *IEEE Access*, vol. 12, pp. 88231–88245, 2024.

- N. Al-Sharif et al., “Stream Analytics for VANET Big Data,” *Expert Syst. Appl.*, vol. 234, Art. 121112, 2024.

- P. Wang et al., “Hierarchical Federated Learning for Vehicular Edge Intelligence,” *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 77–92, 2024.

- R. Abate et al., “TinyML for Edge-Intelligent VANET Devices,” *IEEE Embedded Syst. Lett.*, vol. 17, no. 1, pp. 24–27, 2025.

- S. Hussain et al., “Deep Reinforcement Learning for Data Routing in VANETs,” *IEEE Trans. Veh. Tech.*, vol. 73, no. 4, pp. 4412–4424, 2024.

- M. El-Moussa et al., “Differentially Private Analytics for Vehicular Big Data,” *IEEE Access*, vol. 12, pp. 99420–99433, 2024.

- A. Youssef et al., “Simulation Frameworks for Vehicular Big-Data Systems,” *IEEE Trans. Emerg. Topics Comput.*, vol. 12, no. 2, pp. 290–303, 2025.

- L. Jin et al., “Digital Twin-Based Vehicular Network Management,” *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 5, pp. 5211–5223, 2024.

- Z. Liu et al., “Toward 6G V2X Networks: Opportunities and Challenges,” *IEEE Commun. Surv. Tutorials*, vol. 25, no. 1, pp. 55–87, 2025.

- T. Zhao et al., “Blockchain-Enabled Data Sharing for Vehicular Edge Computing,” *IEEE Trans. Netw. Serv. Manag.*, vol. 21, no. 2, pp. 1420–1435, 2025.

- F. Nguyen et al., “Semantic V2X Communication for Efficient 6G VANETs,” *IEEE Access*, vol. 13, pp. 74122–74136, 2025.