International Journal of Emerging Trends in Computer Science and Information Technology
ISSN: 3050-9246 | https://doi.org/10.63282/3050-9246.IJETCSIT-V6I1P117
Eureka Vision Publication | Volume 6, Issue 1, 148-152, 2025

Original Article

A Framework For Real-Time Root Cause Analysis In Connected Vehicle Iot Data Streams Using Aiops

Naresh Kalimuthu
Independent Researcher.

Received On: 20/02/2025 ***Revised On: 08/03/2025*** ***Accepted On: 19/03/2025*** ***Published On: 21/03/2025***

Abstract *- The development of connected vehicles is generating unprecedented volumes of IoT data, which in turn impacts the functionalities of modern transportation systems. This puts pressure on conventional IT Operations Management (ITOM) frameworks. In this paper, we propose a novel, multi-layered architecture that integrates AIOps (Artificial Intelligence for IT Operations) for real-time fault prediction and autonomous Root Cause Analysis (RCA) within the context of a connected vehicle ecosystem. The architecture integrates vehicle onboard diagnostics, edge computing, and cloud computing to manage efficient data and workload analytics. It contains a hybrid predictive engine that applies lightweight statistical models for low-latency anomaly detection and advanced cloud deep learning models for recognizing complex failure patterns. For diagnostics, we propose a graph-based RCA engine that dynamically models the V2X system's interrelationships to determine the rapid and precise origin of failures. We address the challenges of latency, data scalability, and model explainability, proposing solutions for each. This study aims to propose an operational intelligence framework for connected mobility solutions.*

Keywords *- AIOps, connected vehicles, fault prediction, Internet of Things (IoT), machine learning, root cause analysis (RCA), V2X communications.*

1. Introduction

The automotive sector is undergoing a dramatic shift in its paradigm from classical mechanical engineering to software and distributed computing. Today's vehicles are equipped with dozens of Electronic Control Units (ECUs) and hundreds of sensors. They also have intra-vehicle networks like CAN, FlexRay, and automotive Ethernet, which incorporate vehicles as cyber-physical systems. This change further integrates vehicles into the IoT ecosystem and turns them into crucial data-generating nodes. however, this shift in technology introduces operational challenges to the automotive and IT sectors that have never been seen before, in terms of complexity and scale. Merging these two worlds introduces a new category of cyber-physical operational problems. The latest issue is that a software bug in a cloud-based service could cause a vehicle to malfunction critically, and a vehicle malfunction could, in turn, compromise the cloud service. Approaches from each of the two domains alone are not enough to handle this new, closely connected reality. The complexity of the issue is staggering. For every hour of operation, the high-resolution cameras, LiDAR, and other advanced sensors are projected to yield raw data of three to six terabytes per vehicle[4]. When scaled to a fleet of millions of vehicles, the volume, velocity, and variety of data become profoundly overwhelming. From a single vehicle, there is already a significant data issue.

The challenge becomes even greater with the newer concept of Vehicle-to-Everything (V2X) communications that form a self-organizing and dynamic network of interactions with other vehicles (V2V), Infrastructure (V2I), pedestrians (V2P), and cloud services (V2C). The resultant ecosystem of V2X is a distributed system of unprecedented scale and dynamism. Such an unchecked system creates the paradox of losing control, risking overlooked issues, and compromising reliability and safety. Such frameworks are impossible to provide with the current IT Operations Management (ITOM) systems that are reliant on manual, inflexible, and rule-bound reactive systems. The impact of V2X changes focus not only on the system, but also on operational effectiveness and cost efficiency. Regarding the areas of connected and autonomous mobility, system failures extend beyond the typical considerations of mobility service disruptions. Autonomous connected mobility service failures can lead to costly recalls and pose a risk to human safety. Predicting and mitigating operational anomalies, understanding the precise root cause of deviations, and diagnosing irregularities in real time are mission-critical for operations.

2. End-to-End Working of the AIOps Framework

Prior to the examination of individual operational challenges, an overview of the complete AIOps framework operational sequence is warranted. The architecture implements a stratified distribution of cognitive functions, tiered into three interdependent layers: the vehicle, the network edge, and the cloud. This stratification is a design imperative that balances latency constraints, bandwidth limitations, and computational overhead, ensuring each processing function is placed where it incurs the least total burden.

- **Execution commences at Layer 1:** The Vehicle (On-Board Intelligence). At this tier, a two-step sequence of real-time data triage is executed. First,

incoming high-bandwidth streams from resident sensors and vehicle data buses (e.g., Controller Area Network (CAN), Local Interconnect Network (LIN)) are subjected to recursive filtering, wherein a calibrated set of lightweight anomaly detection models parametrize signal behavior. Deviations exceeding empirically defined thresholds are tagged; all remaining data are suppressed. Rather than transmitting the voluminous raw datasets, only salient, reduced messages consolidated anomaly alerts and statistical summaries are then transmitted to the succeeding edge node. This onboard filtering is an essential first step in managing the large volume of data.

- **Data proceeds from Layer 2:** The Edge (Localized Fleet Coordination). Deployed on edge servers within the 5G mobile network (e.g., Multi-access Edge Computing nodes), this stratum assimilates inputs from numerous nearby vehicles. It executes event correlation across the local fleet to surface emergent phenomena affecting multiple automobiles such as sudden surface ice and concurrently deploys intermediate-complexity predictive models to enable low-latency, vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) safety services.
- **Subsequently, information is conveyed to Layer 3:** The Cloud (Global Fleet Intelligence). This central hub provides the required storage and compute capacity to sustain prolonged, fleet-wide analytics. Here, intricate deep learning architectures ingest petabytes of legacy telemetry to detect subtle performance declines and predict the failure of individual components. This layer also executes domain-general, root-cause exploration and delivers validated findings that are capable of driving both preemptive product recalls and future vehicle designs.

3. Challenges in AIOps for Connected Vehicles

The remainder of this section examines the key challenges to implementing the AIOps architecture within the ecosystem of connected automobiles.

3.1. Managing Data Volume and Velocity at the Edge

The edge nodes in the framework's layered architecture continue to serve as major hotspots of data concentration within the system. For instance, an edge server in charge of a metropolitan area may be receiving aggregated data streams from hundreds or thousands of vehicles at the same time, creating a localized big data challenge. It poses the challenge of processing an exceptionally high volume and velocity of data in real-time, all while constrained by the resources of edge devices.

3.2. Minimizing Latency For Predictive Safety Functions

Active safety and collision avoidance applications require an extremely optimized pipeline for a diverse set of streams, as sensor data collection and insight or warning prediction must be under a few milliseconds. Achieving these ultra-low latencies requires optimizing every stage of the processing pipeline: from data collection and model execution, to data transmission and insight generation.

3.3. Model Accuracy, Drift, and Explainable AI (XAI)

The machine-learning models employed in the framework determine its effectiveness. These models, however, are not fixed. Their accuracy can deteriorate over time and with changes in the operating environment this is referred to as “concept drift.” In addition, the “black box” problem associated with complex models, particularly deep neural networks, is a significant concern in safety-critical domains where every automated decision must be transparent, auditable, and trustworthy.

3.4. Interoperability and framework scalability in a multi-OEM context

The overarching goal of the framework is to achieve fleet-wide scale, multi-telecommunication, and multi-OEM vehicle interoperability. A lack of standards for data and communication in the automotive industry currently restricts that vision. The adoption barrier is not merely technical. It is also socio-technical; concerns of data sovereignty, liability, and trust are at its center.

4. Mitigation Strategies

The strategies listed below aim to mitigate the challenges discussed in the previous section.

4.1. Managing Data Volume and Velocity at the Edge

To address the localized big data challenge, a more comprehensive approach to data reduction at the edge is required in order to avoid being overwhelmed. Considering that each vehicle produces data at a staggering rate of 6 TBs an hour, raw data transmission is out of the question. The core strategy is to implement an intelligent, adaptive data reduction pipeline that captures crucial data while drastically slicing the volume. This consists of several complementary techniques. To accomplish this objective, a combination of data filtering, aggregation, and predictive methods could be employed. These methods include filtering data points based on specific criteria to remove those that are more irrelevant than relevant, as well as aggregating data over periods of time or across multiple sensors. One form of aggregation calculation is the mean and variance of a sensor reading over a second, instead of every millisecond. Predictive solutions that are based on data can refrain from transmission until the actual reading diverges from a predicted reading by a considerable margin, especially during stable conditions.

Next, compressing the data is vital. This may include lossless algorithms such as Huffman or Lempel-Ziv-Welch (LZW), which retain the original data without any alterations, as well as lossy algorithms that achieve higher compression ratios by eliminating less critical data. Modern algorithms, like Zstandard (or Zstd) can be applied to reduce the data footprint before transmission from the vehicle or processing at the edge node. Finally, the reduction of dimensionality techniques can transform large volumes of sensor data into lower-dimensional sensor data that retains

the rich data. This is accomplished with feature extraction. Some linear methods, such as Principal Component Analysis (PCA), can be used to extract the most essential parts of data, but more sophisticated non-linear techniques tend to perform better on complex data sets like those coming from vehicles. For example, Autoencoders, a type of neural network, may be trained on normal operating data to extract a compact representation of the healthy state of the system.

This compact representation enables efficient anomaly detection at the edge any data that cannot be accurately reconstructed may indicate a fault. In some scenarios, a lightweight neural network may even be able to run on the vehicle's embedded systems to encode raw data into a compact, task-oriented format designed for AI processes before reaching the edge, ensuring that the data reduction is aligned with the specific AI processes that will be executed. Finally, these techniques are context-aware and adaptive. For example, the system may operate in a low-fidelity sampling mode during normal driving conditions; however, for certain vehicles or regions where anomalies are suspected, the system may increase the data resolution, sampling rate, or both. This adaptive data acquisition enables the system to dynamically adjust the data density to where computational and network resources are most limited, providing a scalable and efficient data management solution at the edge.

4.2. Ultra-Low Latency Assurance for Safety-Critical Predictions

Synergistic optimization of hardware, models, and network fabric is crucial for achieving millisecond-level end-to-end latency in safety-critical applications. The key to low-latency inference is hardware acceleration. Specialized hardware designed for the parallel processing of neural networks is necessary, not only on edge servers but also on in-vehicle gateways. These accelerators utilize parallel computing techniques, including Graphics Processing Units (GPUs), Field-Programmable Gate Arrays (FPGAs), Application-Specific Integrated Circuits (ASICs), and Tensor Processing Units (TPUs). For example, Nvidia's Jetson Nano consumes 5-10 watts of power while delivering hundreds of GigaFLOPS, making it an energy-efficient system on a chip. This enables the deployment of advanced AI models within vehicles or nearby edge nodes. These devices greatly cut down the time needed to compute models, which is essential for real-time decision-making.

In parallel with hardware acceleration, the AI models undergo aggressive optimization. Typically, a large and sophisticated deep learning model trained in the cloud cannot be deployed directly to the edge due to the resources it requires. Hence, it is important to consider techniques that lower the resources in terms of computation and memory. One of the techniques is Quantization, in which the precision of the model's weights and activations is reduced. For example, transforming 32 floating-point numbers to 8-bit integers. This technique not only shrinks the size of storage but also improves the speed of lower-precision arithmetic hardware computation. Another important technique is Pruning, which is the process of removing connections, also known as redundant or non-critical weights within the neural network. This process leads to a model that is "sparse" and smaller. A sparse model improves inference speed while consuming less memory and computation. These techniques can be combined. Additionally, advanced methods such as quantization-aware training can minimize the accuracy loss that might result from these optimizations.

Lastly, advanced network guarantees enable the controlled certainty of latency, a capability provided through 5G network slicing. This is a foundational capability of 5G architecture that allows the segmentation of a single, physical network into several virtual, end-to-end networks. A network slice can be set up for Ultra-Reliable Low-Latency Communications (uRLLC) [6]. This uRLLC slice is provisioned with specific Quality of Service (QoS) requirements to ensure a minimum of one millisecond latency, ultra-reliable throughput, and high bandwidth, high reliability, as well as ultra-low latency. Through network slicing, network resources are reserved, ensuring advanced reliability for safety applications. This allows critical data packets to bypass congestion caused by non-urgent traffic, thereby guaranteeing deterministic network performance crucial for the function of autonomous vehicles.

4.3. Model Explainability and Drift

A dual approach of a strong operational pipeline to address model erosion, along with the implementation of explainability techniques to clarify model outputs, is needed to maintain the performance and integrity of the framework's AI models for a prolonged period. The first component of the strategy addresses model drift, which is the decline in a model's accuracy over time as it encounters real-world data that diverges from the data on which it was trained.

There are two primary forms of drift: Data drift refers to a change in the statistical characteristics of the input data (e.g., the introduction of a new sensor type across the fleet), and concept drift, which is a change in the relationship between the inputs and outputs (e.g., the emergence of a new, previously unseen failure mode). To address this, a thorough Machine Learning Operations (MLOps) pipeline is essential. This pipeline encompasses the entire lifecycle of the machine learning and acts as a proactive, systematic safeguard against drift. Its key functions include:

- **Continuous Monitoring:** The MLOps pipeline checks the performance of the deployed models against the set metrics (accuracy, precision, recall) and measures data drift between distributions using statistical techniques like KL Divergence.
- **Automated Retraining:** If significant data drift occurs, performance metrics surpass set thresholds, or monitored KPIs are neglected, the pipeline initiates an automated retraining sequence triggered by a fresh model. This new model is trained on a corpus of recent data.
- **Automated Redeployment:** Once the newly trained model has undergone validation and is confirmed to surpass the previously deployed model, the MLOps pipeline automates its

deployment to the edge and cloud. This structure establishes a self-sustaining system of monitoring, retraining, and improvement that sustains the accuracy and reliability of the models over time.

The latter part of the strategy addresses the "**black box**" issue. "Black box" describes a system that may be functioning correctly but lacks a way to conduct a transparent audit of its performance. In an automotive application, having accurate predictions is necessary but not enough; predictive reasoning must also be understandable and auditable. This is the reason for Explainable AI (XAI), which aims to give human-level understanding of AI-based decisions.

Although the framework's graph-based RCA engine is capable of visualizing the failure path and, thus, provides a certain level of explainability, deeper XAI analyses can be used to interpret the deep learning models in the predictive engine. Some of the key XAI approaches are:

- **Model-Agnostic Local Explanations**: Local Interpretable Model-Agnostic Explanations (LIME) [8] and other similar approaches explain a prediction via a simpler model, which is interpretable in its simpler form, built in the vicinity of the prediction of interest, and is, however, faithful to the more complicated model.
- **Feature Contribution Analysis**: With methods such as SHAP (SHapley Additive exPlanations), the contribution of each input feature to a specific prediction shows which sensor readings or data points pushed the model's output in a certain direction[9].
- **Feature Importance and Visualization**: Engineers can gain a deeper understanding of how a model behaves globally by visualizing the marginal effect or its prediction using techniques like Partial Dependency Plots.

With the application of the aforementioned methods of explanation, validation of each model allowed for informed inferences and decisions, thus enabling engineers to place trust in inferences and predictions in a high-stakes context.

5. Framework Scalability and Interoperability in a Multi-OEM Environment

The AIOps framework can be deployed across a fleet of vehicles, provided that socio-technical and technical fragmentation barriers to data sharing are resolved. This must be addressed through a two-pronged mitigation approach centered on the standardization of technology and a governance system based on collaboration. First, to achieve technical interoperability, the framework must address the ecosystem of competing vehicle manufacturers that employ a variety of proprietary communication interfaces and data exchange paradigms. A competing solution would need to focus on proprietary standards. V2X communication standards essential to the framework's modular, API-driven architecture include the IEEE 1609 WAVE (Wireless Access in Vehicular Environments) protocol suite and the SAE J2735 message set standard, which, together, provide secure and interoperable communication for DSRC (Dedicated Short-Range Communications) based systems. In Europe, the equivalent standard is the ETSI ITS-G5. With these harmonized standards, the framework will achieve a minimum communication interface standard for vehicle diversity and model year. More difficult, the second problem is to enable the sharing of data in an ecosystem where data is sensitive and highly desired.

OEMs rarely share telemetry and diagnostic data due to concerns over intellectual property, competitive advantage, and liability. The most effective mitigation strategy would be to form an industry-wide consortium or a trusted, neutral third-party data-sharing platform. Such a governing body would be responsible for core governance strategies and have expansive obligations, providing critical support for an organization or consortium dealing with massive amounts of data. Their core obligations would include the following:

- **Defining Governance Policies:** The governing body would establish and enforce policies for data access and data stewardship, creating a balanced, competitive environment for all participants
- **Enforcing Data and Security Standards:** The consortium would implement consistently the application of a common data ontology, thereby enforcing ecosystem-wide data security and requiring rigorous ecosystem-wide security governance. This also covers the management of a common PKI for all authenticated vehicles and infrastructure nodes, protecting against data poisoning and ensuring system-wide message integrity.
- **Ensuring Privacy:** The consortium would ensure the application of privacy safeguards with regard to the use of GUIDs (Globally Unique Identifiers) and compliance with data sharing and consent rules.
- **Managing Legal and Financial Frameworks:** The body would establish common templates that grant intellectual property rights, outline liability for cross-system failures, specify shares, and detail all associated costs for data and access.

Such a collaborative framework is an essential non-technical complement to the technical architecture, creating the trusted environment necessary to unlock the full safety and efficiency benefits of a truly interconnected vehicular ecosystem.

6. Recommendations

After analyzing the challenges of an AIOps framework for connected vehicles and strategies to mitigate them, we have identified key recommendations to guide organizations working towards an AIOps framework for connected vehicles.

6.1. First safety prioritization for network architecture

The network is a safety-critical backbone component and must be treated as one. A safety-critical network, as a best-effort minimal approach, requires going beyond best effort. It is best to implement proactive approaches, such as the required implementation of network slicing on 5G, which

will guarantee the creation of separate, ultra-reliable, low-latency virtual networks. It is also recommended to utilize edge hardware acceleration, leveraging GPUs and FPGAs, to ensure the allocated latency for critical predictive models is met for stringent computations.

6.2. Set Up a Separately Managed Trustworthy AI and MLOps Practice:

The AI models are an integral part of the AIOps engine and should not be treated as a static software asset. We propose establishing a dedicated MLOps unit that will be accountable for the entire lifecycle of the predictive models. This unit should be responsible for model diagnostics, automated retraining, and model redeployment due to drift, as well as the XAI (Explainable AI) integration pipeline. Every predictive model that is used in safety-critical functions must have an explainability component integrated (for example, via SHAP or LIME) to allow for transparency and auditability of all automated decisions for human controllers and regulators.

6.3. Advocate for Establishing a Sector-Wide Consortium for Data Sharing

The most critical obstacle to a fully functional fleet-wide AIOps platform is not technical, but rather, organizational. In order to realize the full potential of the graph-based RCA engine, ecosystem-wide data accessibility is a necessity. As a result, we recommend that leading OEMs, telecommunication providers, and technology firms form a neutral, third-party consortium that will oversee the controlled and anonymized sharing of operational data. This consortium could formulate standardized data ontologies, security hardware and software infrastructure (PKI), and legal frameworks that would enable the establishment of a trusted data exchange ecosystem, transforming a competitive obstacle into a collaborative effort that bolsters collective data safety and system reliability.

7. Conclusion

This document has theoretically developed a complete, multi-layered AIOps framework for real-time fault prediction and automated root cause analysis for the connected vehicle ecosystem. It can be concluded from the study conducted that a framework of this type is not only possible, but it is also crucial in managing the operational complexities and ensuring the safety of the next-generation intelligent transportation systems. The proposed framework is fully responsive and incorporates actionable answers to the primary research questions that underpin the rationale for this research. Altogether, this framework provides a comprehensive and resilient blueprint for the future of automotive operations by integrating a layered architecture, a hybrid intelligence engine, a graph-based diagnostic system, and strategically addressing key non-technical problems with actionable recommendations.

References

[1] Akoglu, L., Tong, H. & Koutra, D. Graph based anomaly detection and description: a survey. Data Min Knowl Disc 29, 626–688 (2015). https://doi.org/10.1007/s10618-014-0365-y

[2] M. Ammerman, The Root Cause Analysis Handbook: A Simplified Approach to Identifying, Correcting, and Reporting Workplace Errors. Boca Raton, FL: CRC Press, 2001.

[3] Brandon, Alvaro & Solé-Simó, Marc & Huélamo, Alberto & Solans, David & Pérez, María & Muntés-Mulero, Victor. (2019). Graph-based Root Cause Analysis for Service-Oriented and Microservice Architectures. Journal of Systems and Software. 159. 110432. 10.1016/j.jss.2019.110432.

[4] T. Tejpal, "The Data Deluge: What do we do with the data generated by AVs?," Siemens Software, Nov. 14, 2019. [Online]. Available: https://blogs.sw.siemens.com/thought-leadership/the-data-deluge-what-do-we-do-with-the-data-generated-by-avs/

[5] L. D. Xu, W. He and S. Li, "Internet of Things in Industries: A Survey," in IEEE Transactions on Industrial Informatics, vol. 10, no. 4, pp. 2233-2243, Nov. 2014, doi: 10.1109/TII.2014.2300753.

[6] W. Shi, J. Cao, Q. Zhang, Y. Li and L. Xu, "Edge Computing: Vision and Challenges," in IEEE Internet of Things Journal, vol. 3, no. 5, pp. 637-646, Oct. 2016, doi: 10.1109/JIOT.2016.2579198.

[7] W. P. Popovski et al., "Wireless Access for Ultra-Reliable Low-Latency Communication: Principles and Building Blocks," in IEEE Network, vol. 32, no. 2, pp. 16-23, March-April 2018, doi: 10.1109/MNET.2018.1700258.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier," in Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778

[9] S. M. Lundberg and S. -I. Lee, "A Unified Approach to Interpreting Model Predictions," in Proc. 31st Int. Conf. on Neural Information Processing Systems (NIPS), 2017, pp. 4768–4777.

[10] Sandeep Rangineni Latha Thamma reddi Sudheer Kumar Kothuru , Venkata Surendra Kumar, Anil Kumar Vadlamudi. Analysis on Data Engineering: Solving Data preparation tasks with ChatGPT to finish Data Preparation. Journal of Emerging Technologies and Innovative Research. 2023/12. (10)12, PP 11, https://www.jetir.org/view?paper=JETIR2312580

[11] Swathi Chundru et al., "Architecting Scalable Data Pipelines for Big Data: A Data Engineering Perspective," IEEE Transactions on Big Data, vol. 9, no. 2, pp. 892-907, August 2024. [Online]. Available: https://www.researchgate.net/publication/387831754_Architecting_Scalable_Data_Pipelines_for_Big_Data_A_Data_Engineering_Perspective.

[12] S. Panyaram, "Connected Cars, Connected Customers: The Role of AI and ML in Automotive Engagement," International Transactions in Artificial Intelligence, vol. 7, no. 7, pp. 1-15, 2023.

[13] Mohanarajesh Kommineni. (2022/9/30). Discover the Intersection Between AI and Robotics in Developing

Autonomous Systems for Use in the Human World and Cloud Computing. International Numeric Journal of Machine Learning and Robots. 6. 1-19. Injmr.