# Big Data Framework for Monitoring Real-Time Vehicular Traffic Flow

Nawar A. Sultan
*College of computer science and Mathematics*
*University of Mosul*
Mosul, Iraq
nawar.20csp62@student.uomosul.edu.iq

Rawaa Putros Qasha
*College of computer science and Mathematics*
*University of Mosul*
Mosul, Iraq
rawa_qasha@uomosul.edu.iq

2023 International Conference on Engineering, Science and Advanced Technology (ICESAT) | 979-8-3503-4440-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICESAT58213.2023.10347303

***Abstract*—The relatively high rate of traffic accidents in Iraq shows the necessity of working on the driver's actions monitoring through the use of vehicle flow data to improve the road safety. Based on this situation, many tools and technologies such as sensors, cameras, and data management can be utilized to monitor traffic conditions and provide real-time information to drivers and transportation authorities. The primary challenges are collecting, processing, analyzing, and visualizing the huge volume of data produced by vehicles and devices. To address these challenges, we proposed and implemented a big data framework for monitoring the data flows generated by vehicles in the city environment. Among the various data generated by vehicles, our framework monitors the latitude and longitude values of the global positioning system (GPS) and speed. The framework's architecture is scalable and fault-tolerant which makes it suitable for handling large-scale data flows generated by many connected vehicles. The results show that it allows for increased throughput, high availability, and fault tolerance and provides full-text search. This framework has been implemented using several big data platforms and tools such as Apache Kafka and Elasticsearch. In addition, the framework's services have been packaged in the container-based virtualization environment to support the reusability and portability of the framework.**

***Keywords—Traffic data flow, big data, Apache Kafka, Elasticsearch, Container technology***

## I. Introduction

According to the statistics for 2022 in Iraq, there were 11,523 accidents, of which 3,079 were fatal, for a fatality rate of 26.7%, according to data from the Central Statistical Organization at the Ministry of Planning of Iraq. The main cause of the accidents was driver error, accounting for 79.2% of the total accidents, while vehicle-related factors caused 8.1%, and road conditions caused 6.2%. Other factors contributed to 6.5% of the accidents. The most common type of accident was collisions, accounting for 56.3% of the total, followed by pedestrian accidents at 32.3% and overturning at 9.5% [1]. This indicates the driver's need for more awareness to comply with traffic regulations, such as the maximum speed limit for vehicles. Based on these conditions, the evidence-based way of making decisions can be used in many research areas into big data and analytics. Big data can make transportation systems safer and more sustainable [2]. Big data refers to large amounts of structured and unstructured data generated and collected by various sources, such as sensors, cameras, and other devices [3].

Many cities have deployed cameras, roadside sensors, and wireless sensor networks to monitor traffic and improve safety. This technology gathers much traffic data, helping transportation departments analyze local traffic flow. Data can reveal important traffic patterns, congestion, and collision or near-miss causes through historical and streaming data analysis. Big data analytics, such as machine learning, can filter traffic data and extract useful information to help the transportation authority prevent accidents and make smart judgments [4].

One of the primary advantages of using big data in vehicles is the capacity to analyze massive volumes of data to get insights into how vehicles are used, how they perform, and how they may be tuned for better performance and efficiency. Big data analytics, for example, can detect patterns in driving behavior, such as forceful braking or acceleration, which can then be utilized to enhance driver behavior and prevent accidents [5].

This research aims to identify big data characteristics in traffic data monitoring and provide a cohesive understanding of big data processing and analytical approaches in the traffic data flow, supporting real-world applications, including route monitoring, travel planning, road capacity management, traffic flow monitoring, decision-making, and propose the framework to discover the most frequently used data processing and analytical techniques in the literature. However, before processing the data using any of the frameworks, the first step is to transfer the data to the data center where sufficient resources are available for analysis. This transfer mechanism needs to be explored more.

This paper is structured as follows: the next section presents some recent related works. Then describes the framework architecture, design, and tools used. After that, showing the framework implementation. The fifth section discusses performance evaluation and results. Finally, section six shows the conclusions.

## II. Related Work

We briefly discuss relevant works on this topic. These works demonstrate the potential of big data framework techniques for monitoring real-time vehicular traffic flow. By analyzing large volumes of data from multiple sources, these frameworks can provide valuable insights into traffic patterns and help improve traffic management and safety.

Torre-Bastida and others in [6] reviewed recent advances, trends, and challenges in applying big data for transportation and mobility. They highlight the potential of big data in transforming transportation systems through improved efficiency, safety, and sustainability and discuss various data sources, such as traffic sensor data, GPS data, social media data, and mobile phone data, that can be used to extract valuable insights. The authors also discussed challenges in big data transportation research, including data quality, data privacy, and data integration, and proposed future research directions for addressing these challenges.

Reddy et al. [7] found that the GPS data could be used to monitor vehicle speed and identify instances of speeding accurately. They also developed a categorization system for drivers based on their driving behavior, including aggressive, cautious, and average drivers.

M Abdelsalam and T Bonny [8] provided a comprehensive overview of a speed-limiting system for improving road safety in the IoV environment. While the system has some limitations and challenges, the authors believe that it has the potential to make a significant impact on road safety and should be considered a valuable solution for reducing the number of road accidents.

A Najmurrokhman and A Daelami [9] describe the design and implementation of a vehicle speed recorder using a GPS tracker and an Internet-of-Things (IoT) platform. The system is designed to provide real-time monitoring and recording of vehicle speed and location data. The GPS tracker collects data on the vehicle's location and speed, transmitting to the IoT platform via a wireless network. The data is processed and analyzed on the platform, and users can access the data via a web-based dashboard. The paper discusses the hardware and software components of the system, as well as the implementation and testing process. The authors also evaluate the system's performance, including its accuracy and reliability. Overall, the system is designed to provide a cost-effective and efficient solution for vehicle speed monitoring and recording.

Gamboa-Venegas and Carlos [10] suggest that integrating GPS navigation data into traffic simulation models can help to optimize traffic management and improve the accuracy of traffic predictions. The authors suggest that their approach has the potential to be applied to other areas and can contribute to the development of more efficient and effective traffic management strategies.

C Bachechi et al [11] discussed using big data analytics and visualization in traffic monitoring. The authors highlight the benefits of using these tools to collect and analyze real-time traffic data from various sources, such as traffic cameras and GPS sensors. The paper also provides examples of how big data analytics and visualization can be used in traffic monitoring, including predictive analytics, incident detection, route optimization, and public transportation optimization. Overall, the paper emphasizes the importance of using these tools to improve traffic management and the transportation experience for commuters.

Rodolfo Metulini and Maurizio Carpita [12] discussed using mobile phone big data to model and forecast traffic flows in flooding risk areas. The study aims to support data-driven decision-making by providing accurate and timely information on traffic patterns during floods. The authors propose a method that uses data from mobile phones to create a traffic flow model, which is then used to forecast traffic conditions during a flood. The proposed method is tested in a case study in Guangzhou, China, and the results show that it effectively predicts traffic conditions during a flood. The authors conclude that the method has the potential for use in other cities and can support decision-making for flood management and evacuation planning.

These studies demonstrate the continued development and application of big data frameworks for monitoring traffic data flow, emphasizing real-time monitoring and prediction utilizing data sources and advanced analytics techniques. Utilizing edge computing, cloud computing, and IoT architectures demonstrates the potential for scalable and effective traffic management systems. Our framework proposal used Apache Kafka for data streaming, Elastic Search for full-text search, and online monitoring of traffic data flow and data storage for other processing and analyses.

## III. The Framework Architecture

To implement the framework, it would be necessary to have a robust infrastructure that can handle the large volumes of data generated by vehicles. This would include hardware and software components such as sensors, data processing tools, and data storage systems. Additionally, it would be necessary to have trained personnel to analyze the data and interpret the results [9].

### *A. The Design and Used Tools*

Based on the aforementioned challenges and the requirements, we designed the framework to provide a high-performance big data monitoring system for data traffic flow. As illustrated in Fig.1, our framework is an integrated system with multiple components involved.

It contains the following services/components:

- Connected vehicles: Gateway service (API service where clients/vehicles interact with our framework for sending data from applications (Android application) or sensors via a Virtual Private Network (VPN), implemented with Golang or Python, and deploy as a container in Docker or Kubernetes).
- Access layer: a cluster of the server computers (local computers or cloud service like Amazon EC2 instance).
- Apache Kafka Layer: to collect, process, and analyze real-time streaming data.
- Logstash layer: collects, parses, and transforms data from different sources and sends it to various destinations.
- Elasticsearch Layer: to provide full-text search, log analysis, monitoring, and machine learning. It can handle large volumes of data and can be distributed across multiple nodes, allowing for high availability and fault tolerance.
- Kibana (interface layer): provides a web-based interface to explore and analyze data using various interactive visualizations such as bar charts, line charts, pie charts, tables, maps, and more. It creates custom dashboards and reports to share insights with others. we can perform various operations on data such as filtering, aggregating, and sorting. Kibana also provides a powerful search interface to search and analyze data in real-time [13].

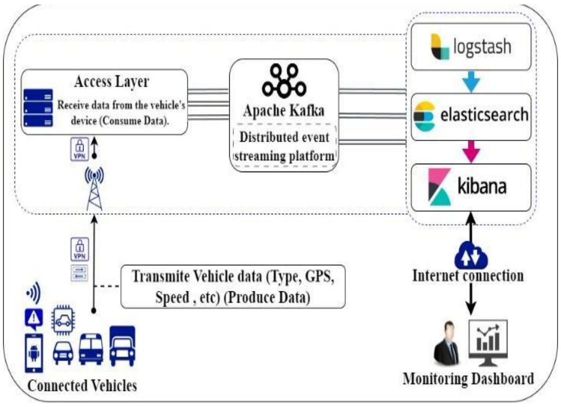

Access Layer
Receive data from the vehicle's device (Consume Data).
VPN
Apache Kafka
Distributed event streaming platform
logstash
elasticsearch
kibana
VPN
Transmite Vehicle data (Type, GPS, Speed , etc) (Produce Data)
Internet connection
Connected Vehicles
Monitoring Dashboard

Fig. 1. Overall Framework Structure.

### B. *The Framework Stages*

We can scope the characteristics according to the techniques used to build our framework, as follow:

- Data Collection: Collecting data from various sources such as vehicles, infrastructure, and sensors [14]. Using Apache Kafka can create a scalable and reliable data pipeline that can handle the large volumes of streaming data generated by different sources. This pipeline can be used to collect data from various sources, including sensors, GPS trackers, and mobile devices. The data can then be processed and analyzed in real-time using a variety of tools, including machine learning algorithms and predictive analytics [15].
- Data Processing: Analyzing the collected data to extract meaningful insights and patterns [16]. By using the Logstash configuration file we can define the input, filter, and output, we need to create a Logstash configuration file that specifies these components and their configurations. Logstash provides a flexible and powerful way to process data [13].
- Data Storage: Storing the processed data in a database or warehouse for easy access and retrieval [17]. Elasticsearch is a powerful, distributed search and analytics engine that can be used for data storage. It provides a highly scalable and flexible way to store and search data, and its powerful search and aggregation capabilities make it ideal for use cases such as traffic data flow, log analytics, e-commerce, and security analytics [13].
- Data analysis and visualization: Analyzing stored data to find patterns and trends that can be used to improve traffic flow management [18]. Presenting the analyzed data meaningfully to help stakeholders understand the traffic flow patterns and identify areas for improvement [11]. As mentioned early, Kibana is an open-source tool that can be used for data analysis and visualization. It provides a flexible and powerful way to analyze and visualize data, and its wide range of visualization types and customization options make it suitable for various use cases, especially for traffic data flow [13].

## IV. The Framework Implementation

This framework can be applied in many applications that require dealing with big data and need robustness in implementation and availability of services. This is what our framework provides, as it can be adapted for monitoring big data of traffic flow.

As a use case for our framework, vehicle traffic data flow tracking is presented to collect data on public roads. The proposed system was developed to collect vehicle data using an Android application that is installed on a mobile device or a vehicle dashboard unit that supports the Android system. And through this application, the vehicle information across a Virtual Private Network (VPN) via the internet (4G or 5G), including the current speed, is sent to the towers, local stations, and the Road Side Unit (RSU) on edge computing. Depending on this VPN, all nodes communicate on the framework network.

All devices that contain the application (mobile, or vehicle dashboard unit) send their information (producing) to the framework.

As mentioned earlier, Apache Kafka is used, which provides a set of brokers that receive and store vehicle information. Fig. 2 shows Kafka's topics and partitions. In the proposed framework, a special topic, called (Data Collector Topic), contains 1000 partitions to deal with the huge amount of information and huge data generated by the vehicles.

After that, connect Logstash with Kafka by configuring the input plugin for Kafka in the Logstash configuration file, then configured the Elasticsearch output plugin with the Elasticsearch details, such as the host, port, index, and document type. The configuration is shown in Fig. 3.

Next, start Logstash and monitor the logs to ensure data flows from Kafka to Elasticsearch through Logstash.

By connecting Logstash with Kafka and Elasticsearch, we create a powerful data pipeline that allows us to consume data from Kafka topics, transform and filter the data using Logstash, and send the data to Elasticsearch for indexing and analysis. This pipeline can be used for various use cases, such as real-time data processing, log analysis, and monitoring.

Finally, connect Kibana to Elasticsearch to access the data that we have indexed. You can do this by specifying the Elasticsearch URL and port in the Kibana configuration file. By connecting Kibana to Elasticsearch, we can create visualizations and dashboards in Kibana to analyze, visualize data and user interface to create various types of visualizations, such as bar charts, line charts, and maps, and combine them into dashboards for a unified view for traffic data flow.

As a result, our framework can ingest, process, store, and visualize data in a scalable and efficient way.

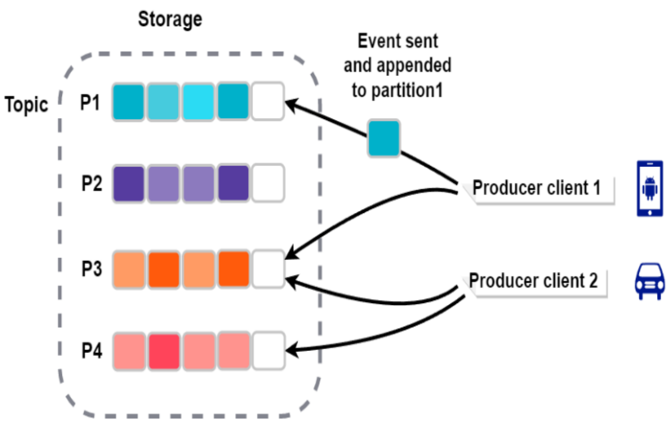
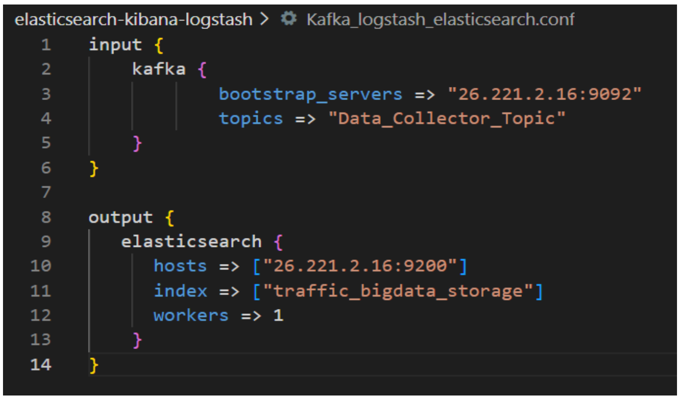


Fig. 2. Partitioned Kafka topic.



elasticsearch-kibana-logstash > ⚙ Kafka_logstash_elasticsearch.conf

```
input {
    kafka {
        bootstrap_servers => "26.221.2.16:9092"
        topics => "Data_Collector_Topic"
    }
}

output {
   elasticsearch {
      hosts => ["26.221.2.16:9200"]
      index => ["traffic_bigdata_storage"]
      workers => 1
   }
}
```

Fig. 3. Logstash config file.

## V. Performance Evaluation and Results

The framework was deployed on two server clusters (8 CPU core i7 and 32GB of RAM). It is set up to run with Kafka, Zookeeper, and Elasticsearch as the database. The Vehicle Data Producer Simulation Application (VDPSAPP) is implemented with Golang and deployed the simulation application in the Docker hub.

To evaluate the performance of the framework, some of the commonly used performance metrics can be used, such as:

Throughput: Throughput is the number of requests that the system can handle per unit of time [19]. In the context of real-time traffic monitoring, the throughput can be measured in terms of the number of vehicles that can be tracked per second or minute.

Scalability: Scalability is the ability of the system to handle increasing amounts of data and traffic [11]. In the context of real-time traffic monitoring, scalability can be measured in terms of the number of vehicles that can be tracked as the traffic volume increases.

Accuracy: Accuracy is the ability of the system to correctly identify and track vehicles [11]. In the context of real-time traffic monitoring, accuracy can be measured in terms of the number of correct vehicle identifications and tracking.

Availability: Availability is the ability of the system to remain operational and responsive even during high traffic loads or system failures [20]. In the context of real-time traffic monitoring, availability can be measured in terms of the percentage of time that the system is operational and responsive.

We have conducted a performance evaluation of our framework based on the following:

- Accuracy and reliability: The results show our framework provides accurate and reliable information about traffic flow, with a low error rate, high confidence, and fault tolerance, can handle noisy and incomplete data, and adapt to changing traffic conditions over time. when running 100 VDPSAPP in a docker container in one of the servers, the results show the stability of our framework in different time snapshots (in 1 Second, the number of messages hits 2000, see Fig. 4. In 1 Minute, the number of messages hits 150000, see Fig. 5. in 1 Hour, the number of messages hits 2000000, see Fig. 6).

- Performance: The results show our framework can handle high volumes of data and provide real-time analysis and visualization with low latency. It is scalable and handles increasing data volumes over time. We checked the framework in four different workloads; first, we ran 1 docker container of VDPSAPP in one of the servers, then ran 2, then 5, and finally ran 10. By calculating the throughput for the four statuses of workload in each second, we get results as shown in Fig. 7, Fig. 8, Fig. 9 and Fig. 10.

By evaluating a vehicular big data framework against these factors, you can assess its suitability for monitoring data traffic flow. also, full-text search capabilities in the Kibana dashboard can be used to find relevant data and identify potential issues quickly. For example, we could search for instances where traffic flow was particularly slow or where accidents occurred, allowing us to take action to alleviate congestion and improve safety.

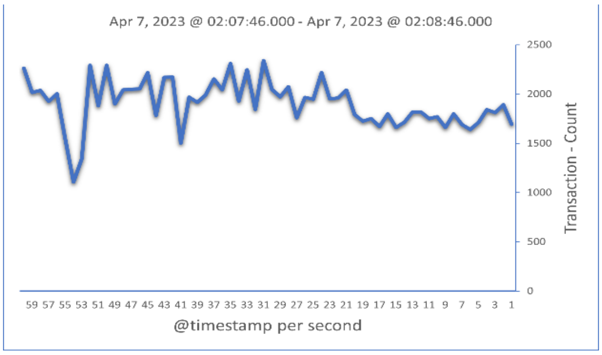


Fig. 4. Accuracy and reliability: per second.

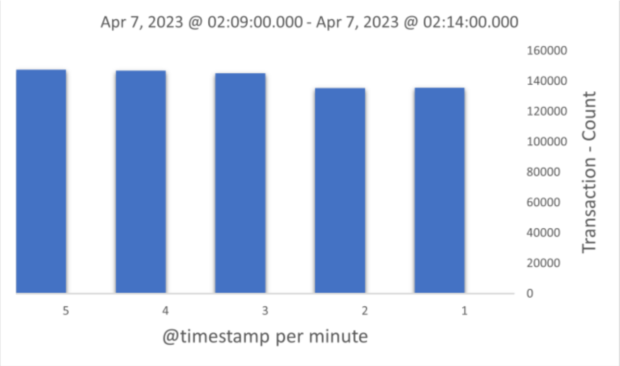


Fig. 5. Accuracy and reliability: per minute.

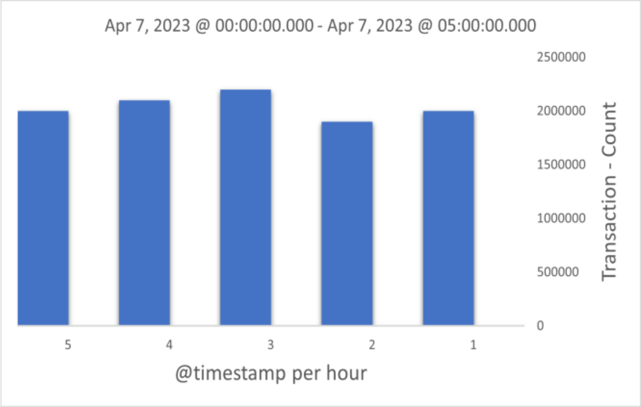


Fig. 6. Accuracy and reliability: per hour.

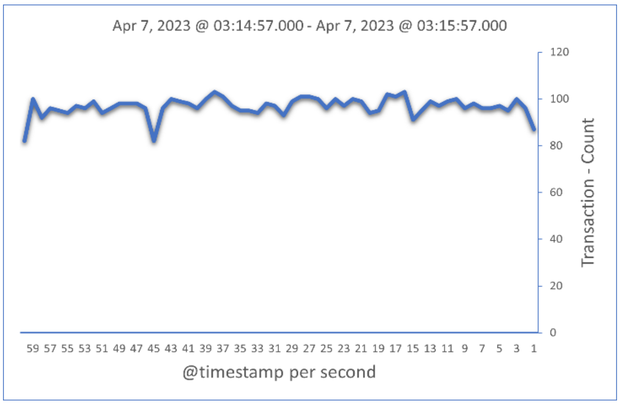


Fig. 7. Performance: 1 workload.

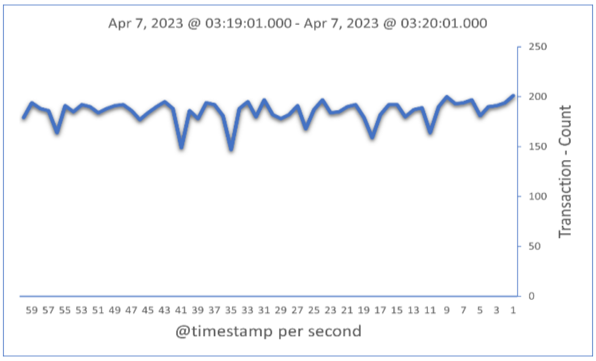


Fig. 8. Performance: 2 workloads.

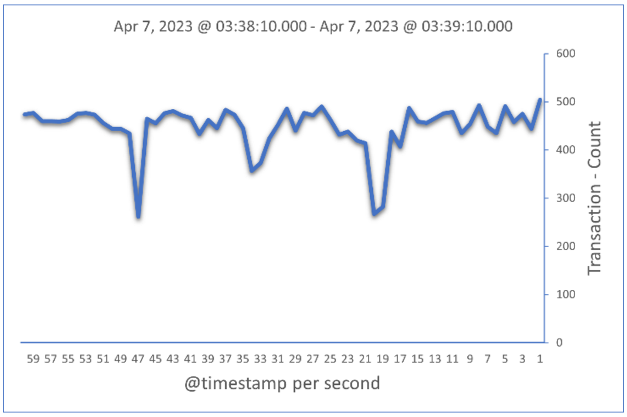


Fig. 9. Performance: 5 workloads.

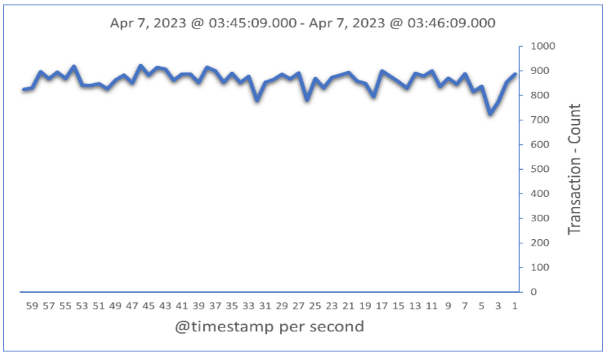


Fig. 10. Performance: 10 workloads.

## VI. Conclusions

In this research, we presented our proposed framework that provides collecting, managing, and analyzing the data generated by vehicles, which can help organize and process a large volume of data efficiently and effectively. The results showed that our proposed framework which is based on Apache Kafka and Elasticsearch can handle a large volume of data generated by vehicles and can be distributed across multiple servers for increased throughput, high availability, real-time processing, and fault tolerance. It also provides full-text search, making searching and analyzing the data more flexible.

Future smart vehicles will be increasingly connected and share information with smart city facilities. As a result, the vehicles will generate a significant amount of data (big data), which can be utilized to make informed decisions, improve the transportation industry's efficiency, reduce accidents, and enhance overall mobility. Effective management and analysis of this data will be critical to achieving these goals.

## References

[1] A. A. Aldhalemi and F. A. Abidi, “Road traffic injuries in Iraq related to the sustainable development goals: A retrospective study,” in AIP Conference Proceedings, vol. 2776, no. 1, 2023.

[2] S. Kaffash, A. T. Nguyen, and J. Zhu, “Big data algorithms and applications in intelligent transportation system: A review and bibliometric analysis,” Int. J. Prod. Econ., vol. 231, 2021.

[3] L. Zhu, F. R. Yu, Y. Wang, B. Ning, and T. Tang, “Big data analytics in intelligent transportation systems: A survey,” IEEE Trans. Intell. Transp. Syst., vol. 20, no. 1, pp. 383–398, 2018.

[4] A. Neilson, B. Daniel, and S. Tjandra, “Systematic review of the literature on big data in the transportation domain: Concepts and applications,” Big Data Res., vol. 17, pp. 35–44, 2019.

[5] T. Rajeshkumar, S. Preethi, R. S. Rubini, and V. Yamini, “Speed Detecting and Reporting System using Gps/Gprs and Gsm,” Int. J. Pure Appl. Math., vol. 118, no. 20, pp. 73–79, 2018.

[6] A. I. Torre-Bastida, J. Del Ser, I. Laña, M. Ilardia, M. N. Bilbao, and S. Campos-Cordobés, “Big Data for transportation and mobility: recent advances, trends and challenges,” IET Intell. Transp. Syst., vol. 12, no. 8, pp. 742–755, 2018.

[7] N. R. Reddy and S. Subhani, “Monitoring Vehicle Speed using GPS and Categorizing Driver,” International Journal of Scientific Research in Computer Science and Engineering, Vol.7, Issue.5, pp.14-21, 2019.

[8] M. Abdelsalam and T. Bonny, "IoV Road Safety: Vehicle Speed Limiting System," International Conference on Communications, Signal Processing, and their Applications (ICCSPA), Sharjah, United Arab Emirates, pp. 1-6, 2019.

[9] A. Najmurrokhman, Kusnandar, A. Daelami, U. Komarudin and M. Imanudin, "Design and Implementation of Vehicle Speed Recorder using GPS Tracker and Internet-of-Things Platform," International Conference on Artificial Intelligence and Computer Science Technology (ICAICST), Yogyakarta, Indonesia, pp. 152-156, 2021.

[10] C. Gamboa-Venegas, “Optimization of traffic simulation using GPS

navigation records,"M.S. thesis, School of Computing, Costa Rica Institute of Technology, Costa Rica, 2021.

[11] C. Bachechi, L. Po, and F. Rollo, "Big data analytics and visualization in traffic monitoring," Big Data Res., vol. 27, p. 100292, 2022.

[12] R. Metulini and M. Carpita, "Modeling and forecasting traffic flows with mobile phone big data in flooding risk areas to support a data-driven decision making," Ann. Oper. Res., pp. 1–26, 2023.

[13] A. Talaş, F. Pop, and G. Neagu, "Elastic stack in action for smart cities: Making sense of big data," 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP), pp. 469–476, 2017.

[14] Y. Ren, T. Wang, S. Zhang, and J. Zhang, "An intelligent big data collection technology based on micro mobile data centers for crowdsensing vehicular sensor network," Pers. ubiquitous Comput., pp. 1–17, 2020.

[15] A. Bandi and J. A. Hurtado, "Big data streaming architecture for edge computing using kafka and rockset," 5th International Conference on Computing Methodologies and Communication (ICCMC), pp. 323–329, 2021.

[16] A. Arooj, M. S. Farooq, A. Akram, R. Iqbal, A. Sharma, and G. Dhiman, "Big data processing and analysis in internet of vehicles: architecture, taxonomy, and open research challenges," Arch. Comput. Methods Eng., vol. 29, no. 2, pp. 793–829, 2022.

[17] A. W. Malik, I. Mahmood, N. Ahmed, and Z. Anwar, "Big data in motion: A vehicle-assisted urban computing framework for smart cities," IEEE Access, vol. 7, pp. 55951–55965, 2019.

[18] T. Kolajo, O. Daramola, and A. Adebiyi, "Big data stream analysis: a systematic literature review," J. Big Data, vol. 6, no.47 , 2019.

[19] N. Ben Aoun, "A Scalable Big Data Framework for Real-Time Traffic Monitoring System," Journal of Computer Science, vol 18, no.9, pp. 801-810, 2022.

[20] K. Khazukov et al., "Real-time monitoring of traffic parameters," J. Big data, vol. 7, no. 1, pp. 1–20, 2020.