Distributed Big Data Architecture for Smart Transportation Using Hadoop, Kafka, and Apache Flink

Nambua Ladslaus Mnyone and Noor Amin *

Posted Date: 17 March 2026

doi: 10.20944/preprints202603.1269.v1

Keywords: big data architecture; smart transportation; Artificial Intelligence

 doi:10.20944/preprints202603.1269.v1

Article

Distributed Big Data Architecture for Smart Transportation Using Hadoop, Kafka, and Apache Flink

Nambua Ladslaus Mnyone and Noor Ul Amin *

Taylor's University Subang Jaya, Malaysia

* Correspondence: nooraminnawab@gmail.com

Abstract

Smart transportation systems generate large amounts of data from sources such as GPS devices, IoT sensors, cameras, and connected vehicles. Managing and processing this data efficiently is important for improving traffic flow, reducing congestion, and enhancing road safety. Traditional centralized systems often struggle to handle the volume, velocity, and variety of transportation data. Therefore, distributed big data technologies are required to support scalable and efficient data processing. This paper presents a distributed big data architecture for smart transportation using technologies such as Hadoop, Apache Kafka, and Apache Flink. Hadoop provides distributed storage and batch processing for large historical datasets, while Kafka enables reliable real-time data streaming from multiple sources. Apache Flink supports real-time stream processing and event detection for traffic monitoring and incident management. The proposed architecture integrates these technologies to enable efficient data collection, processing, and analysis in intelligent transportation systems. The study also discusses the role of data analytics, edge computing, and machine learning in improving traffic management. Results from the analyzed dataset show improvements in emergency detection, response time, accident reduction, and congestion management when advanced data processing techniques are applied.

Keywords: big data architecture; smart transportation; Artificial Intelligence

Introduction

Modern cities are growing quickly and the number of vehicles on roads is increasing every year. Because of this, transportation systems generate a very large amount of data. This data comes from many sources such as GPS devices, traffic cameras, road sensors, mobile applications, and connected vehicles [1]. These technologies help cities monitor traffic conditions and understand how transportation systems work

Managing this large amount of transportation data is not easy [2]. Traditional systems often store data in centralized databases, which can become slow and inefficient when the data volume increases. These systems may also struggle to process real-time traffic information, which is very important for detecting accidents, reducing congestion, and improving road safety [3–5].

To solve this problem, distributed big data technologies are used in modern smart transportation systems. Distributed systems allow data to be stored and processed across many machines instead of a single server. This improves scalability, reliability, and processing speed when dealing with large datasets [6,7].

Technologies such as **Hadoop**, **Apache Kafka**, and **Apache Flink** play an important role in building these distributed transportation systems [8,9]. Hadoop provides distributed storage and batch data processing, Kafka enables real-time data streaming from multiple sources, and Flink

 doi:10.20944/preprints202603.1269.v1

supports real-time data analysis with low latency. By combining these technologies, transportation authorities can process both historical and real-time data to improve traffic management and support intelligent transportation systems [10–12].

This paper discusses the architecture and components of a distributed big data system for smart transportation. It explains how technologies such as Hadoop, Kafka, and Flink can work together to collect, process, and analyze transportation data efficiently.

Methodology

Hadoop

Hadoop is a framework for distributed storage and processing that allows organizations to store and analyze vast amounts of structured and unstructured data. It consists of the Hadoop Distributed File System for storage and the MapReduce programming model for data processing [13,14]. In transportation, Hadoop is useful for handling large datasets such as historical traffic data, vehicle tracking logs and public transit records.it enables scalable data processing and helps authorities and organizations derive meaningful insights for urban planning and transportation.Management [15]. Apache Hadoop is a foundational big data framework that enables the distributed storage and processing of massive datasets across clusters of computers. Its architecture is designed for scalability and fault tolerance, making it a cornerstone in handling extensive transportation data [16–18].

Large volumes of data may be stored which is thanks to Hadoop's distributed file system (HDFS), which is scalable. The namenode, also known as the masternode, provides the data, which is then distributed or divided among several nodes, which are the datanodes . Performance levels are maintained by this design, which ensures that the system may scale horizontally by adding additional nodes as data quantities increase as a result of an increase in data sources, vehicles, and infrastructure [19–21]. This scalability is essential for storing historical data in the transportation industry as data volumes rise. In big clusters, HDFS divides files into blocks and distributes them among multiple nodes. Furthermore, even in the event of a node failure, HDFS facilitates data movement between the nodes, allowing the system to continue functioning.

Figure Hadoop Logo

Hadoop Architecture and Components

 doi:10.20944/preprints202603.1269.v1

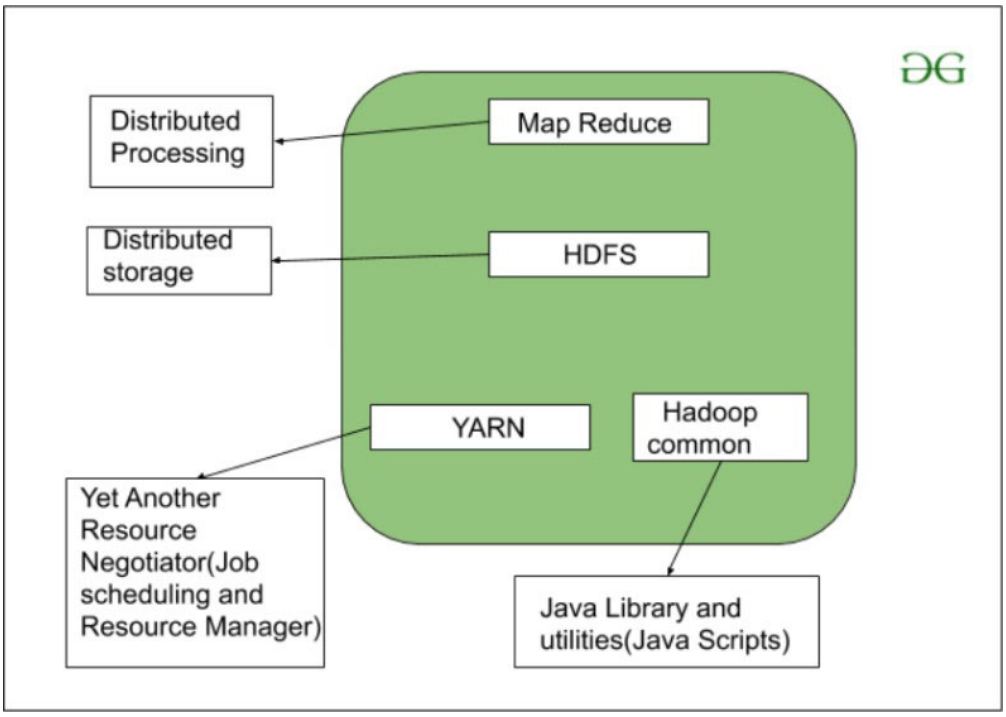
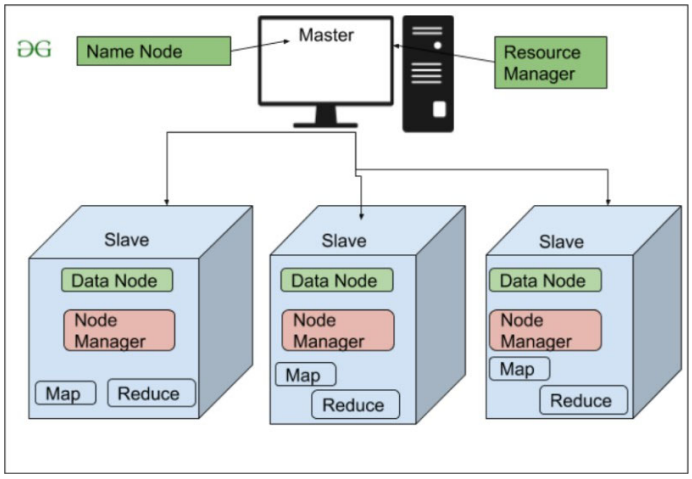


Figure Hadoop Components

2. Hadoop Distributed File System (HDFS) – Storage Layer

HDFS (Hadoop Distributed File System) is a framework which is fault-tolerant and scalable storage system designed to handle a particularly large-scale dataset. It operates on a master-slave architecture, where the Namenode or the Masternode is responsible for the management of metadata, while Datanodes store actual data(Malidev, 2020) [22].



Figue showing Hadoop distributed file system architecture

One of the major features of HDFS is its distributed storage capacity. The data is divided into blocks and distributed in many machines, or nodes [23–25]. This structure increase's fault tolerance, as each data block is repeated by default, three copies are stored in different nodes. This means that if one node fails, another copy of the data from the three stored copies ensures its availability. In addition to that, HDFS is highly scalable meaning that new machines can be added to the cluster without affecting the overall performance [26–29].

Its importance in Smart Transportation:

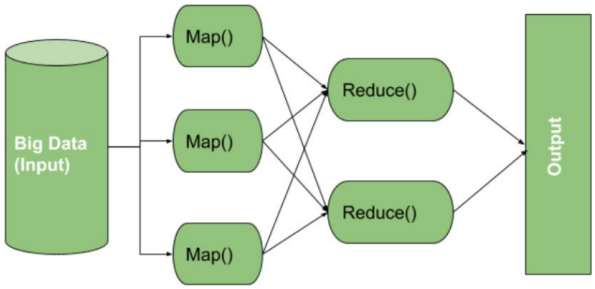
 doi:10.20944/preprints202603.1269.v1

HDFS is very important and has a crucial role in traffic management in smart cities, which produce daily traffic data from sources such as sensors, GPS and monitoring cameras. HDFS ensures that this data is safely stored, remains accessible, and is protected from damage due to hardware failures [30].

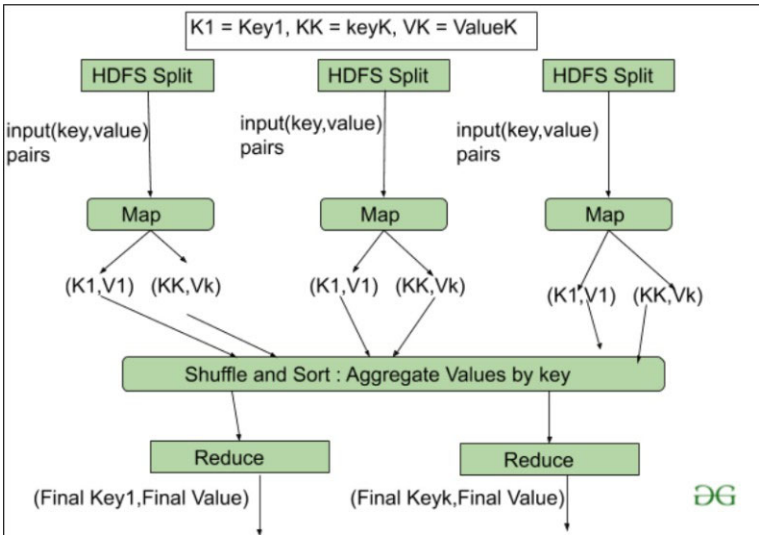
For example, in a smart transport system, historical traffic data can be stored in HDF. This data can later be used for traffic analysis and congestion prediction, which helps in improving traffic management and plan [31–33].

3. MapReduce – Processing Layer

Mapreduce is a Hadoop's data processing model which is especially designed to handle large processing functions, this is done by breaking them into small parallel tasks. This model increases efficiency and speed when dealing with broad or large dataset [34,35].



The process of mapreduce consists of three main stages. The first is the map phase this is where the dataset is divided into small segments that is processed simultaneously in different nodes [36,37].Next, there is the shuffle and sort phase whereby the results of these various nodes are collected. Finally, in the reduce phase, the data is processed and combined to produce the final output (Malidev, 2020).



Its importance in Smart Transportation:

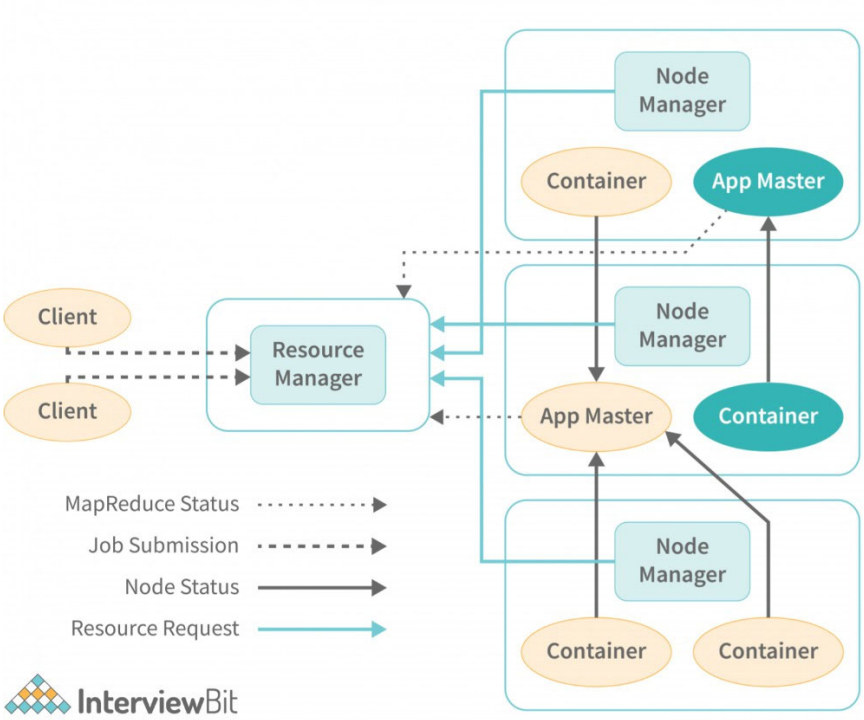
Mapreduce is particularly important for traffic monitoring, as city-wide traffic flows, accident reports, and analyzing vehicle movement data requires parallel processing [38,39]. This ability ensures that the processing remains sharp and efficient whenever billions of GPS data points are dealt with [40]. For example, a city traffic control system may use mapreduce to analyze vehicle density per hour on major highways. This analysis helps in generating reports and alerts for traffic management officers, which facilitates better decision making and resource allocation.

 doi:10.20944/preprints202603.1269.v1

4. Yet Another Resource Negotiator (YARN) – Resource Management Layer

Yarn, or yet another resource -negotiator is the resource manager of Hadoop which plays an important role in efficiently allocating CPU, memory and other resources for various tasks within the Hadoop cluster. This layer is very cruicial since it ensures that resources are effectively used in many applications.The elements of yarn include resource manager(one per cluster) , application master (one per application) and node managers (one per node) (*Hadoop Architecture in Big Data Explained: A Complete Guide with Its Components*, n.d.).

One of the major tasks of yarn is to manage several applications running simultaneously in the hadoop cluster. This ensures proper resource distribution between all functions, which helps prevent hurdles that can slow down processing. Additionally, the yarn allows for dynamic scaling, which means that it can accommodate resource allocation on the basis of current assignment, optimize performance as changing performance.



Its importance in Smart Transportation:

In the context of traffic management, yarn is particularly important because many traffic-related tasks in a city are running in parallel, such as real-time congestion analysis, vehicle tracking and detection of accidents. By managing these tasks, the yarn ensures that each receives the necessary computing power without negatively affecting the performance of the overall system. For example, a traffic analytics can use platform yarn to effectively balance computing resources between real - time GPS tracking and historical congestion analysis. This balance allows timely insight and better traffic management decisions.

5. HBase – NoSQL Database for Real-Time Data

HBASE is a high-performance NOSQL database that is designed or created to provide real-time access to structured data that is stored within the Hadoop system. Unlike the traditional relationship database, HBASE is capable of efficiently handling unnecessary and semi-composed data. This flexibility is especially well suited for applications that require rapid data processing.

 doi:10.20944/preprints202603.1269.v1

One of the standout features of HBASE is the ability to support millions of readings and writing operations per second. This high throughput is required for real -time applications, which allow for quick data updates and recovery. In addition to that, HBASE appoints a column-based storage model, which optimizes it for large-scale data recovery, making it an ideal option for managing wide datasets.

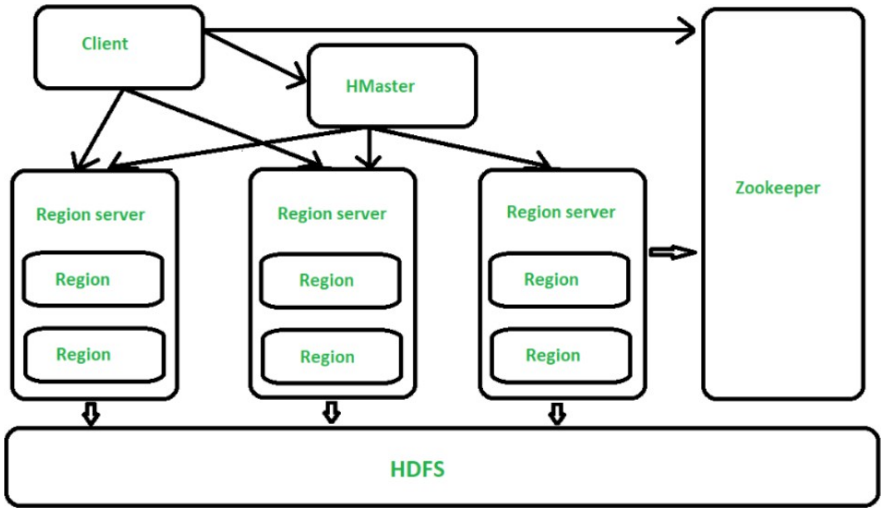


Figure Architecture of HBase(*Architecture of HBase*, 2018)

Its importance in Smart Transportation:

In the context of traffic monitoring or transportation management at large, HBASE plays a very important role. Real-time traffic data, including vehicle speed, road closure and congestion levels, should be rapidly accessed and updated in order to provide accurate information. HBASE facilitates immediate data recovery, ensuring that the traffic reports remain current and reliable.

For example, navigation systems such as Google Maps uses HBASE to store live GPS locations of thousands of vehicles. When a user request real -time traffic updates, HBASE can quickly recover the latest congestion data, allowing the system to provide timely and relevant information to users navigating the roads.

6. Hive – SQL-Based Data Warehouse

The hive is a powerful data warehouse tool that enables different users to query large datasets using syntax like SQL such as as HiveQL. This feature makes it accessible to people familiar with SQL, allowing easy interaction with large data without the need for comprehensive programming skills.

One of the major features of the hive has is that it has the ability to convert SQL questions into mapreduce jobs, which allows efficient execution on large datasets. In addition, the Hive supports both structured and semi-structured data which makes it versatile for various types of data analysis. This capacity or capability is particularly beneficial for analysts and data scientists who prefer to work with familiar SQL Syntax rather than delaying complex programming languages.

Hive it is particularly important for the planners and researchers of the city, who require direct methods to query and analyze historical traffic data.

 doi:10.20944/preprints202603.1269.v1

Table 1. Simplified explanation on Hadoop components.

Hadoop Component	Description
HDFS (Hadoop Distributed File System)	A fault-tolerant storage system that splits data into blocks and distributes them across multiple machines for reliability and scalability.
MapReduce	A processing model that breaks large tasks into smaller parallel jobs for faster computation.
YARN (Yet Another Resource Negotiator)	Manages resources (CPU, memory) across multiple applications to optimize performance
HBase	A NoSQL database for real-time access to large structured datasets, supporting fast read/write operations.
Hive	A data warehouse system that allows querying large datasets using SQL-like syntax (HiveQL).

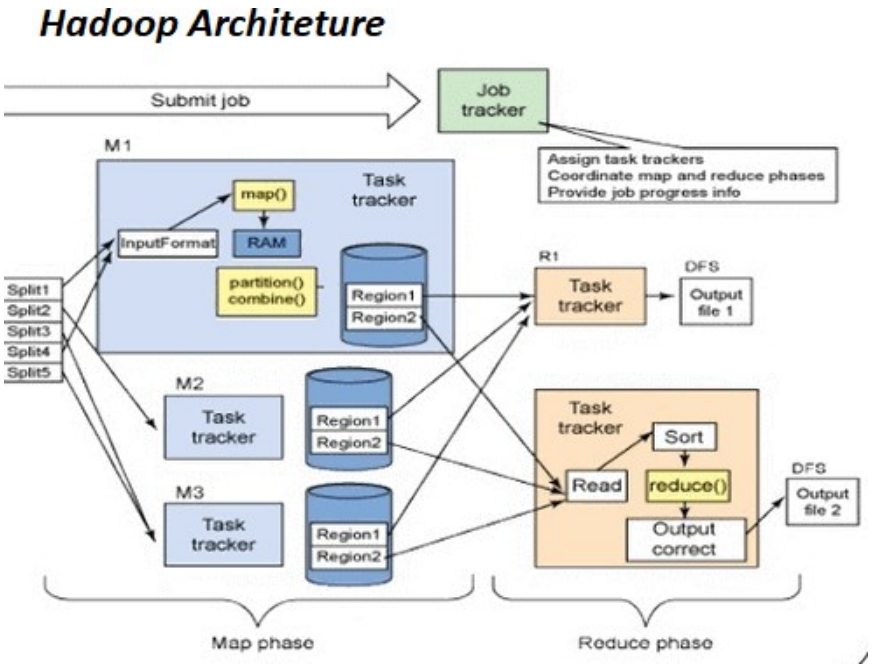


Figure of Hadoop Architecture

Justification for Using Hadoop in Traffic Monitoring Systems

- Cost-Effective Storage

Cost-Effective storage is an important aspect of management in huge amounts of unstructured data generated by transport systems. These systems produce data from various sources including GPS devices, road sensors, cameras, weather reports and social media. Traditional relationship databases, which are SQL-based, often require expensive high-end servers and structured data formats. This makes them unable to handle large amounts of large data produced in the transport sector.

 doi:10.20944/preprints202603.1269.v1

Hadoop provides a low-cost solution by allowing the data to be stored, in a distributed, scalable and fault-tolerant manner in many machines. Hadoop is one of the ways to obtain this cost - effective storage, using commodity hardware. Instead of relying on expensive enterprise-grade servers, Hadoop clusters can work on low cost, off-the-shelf machines. This significantly reduces the initial investment required for data storage.

Additionally, Hadoop uses a scalable distributed file system known as HDFS. In this system, the data is broken into chunks and stored in several nodes, which helps reduce storage costs further. Another important aspect is the use of compression and data dedication. Redundant and duplicate data, which are common in transport logs, can be removed or compressed, storage overheads can be minimized.

For example, consider a smart city transportation system that collects GPS logs from thousands of vehicles every second. Instead of relying on expensive relationship databases, the system can stool raw GPS data in HDFS at a fraction of the cost. When traffic planners need to reach the historic movement pattern, the data remains easily accessible without the need for expensive storage upgrade. This shows how Hadoop's cost-effective storage solution can benefit transportation systems.

- Batch Processing for Historical Data Analysis

Hadoop's batch processing is very important for smart cities as it is used to analyze historical traffic patterns, identify congestion trends and also plan new road infrastructure. Traditional databases are designed for real -time transactions, but they are not equipped to handle the huge amounts of data that is required for long term traffic analysis. Hadoop's mapreduce model allows for the parallel processing of very large datasets, while tools such as Hive and PIG enable analysts to use queries such as SQL to extract insight or different information from data. This approach is particularly useful to train the AI model for predictive analysis, using previous traffic data to estimate the future congestion.

For example, a city government can analyze a five-year traffic data to identify extreme congestion time and places, eventually reporting decisions about new road expansion or alternative routes. Without Hadoop, processing this amount of data would be an important challenge, but the efficient batch processing of Hadoop allows for analysis of petabytes of historical traffic logs, which supports the long -term plan.

- Fault Tolerance for Uninterrupted Data Availability

Fault tolerance is essential for smart traffic monitoring systems, which depends on continuous data collection from various sources, including monitoring cameras, weather sensors, IoT devices and GPS tracking systems. System crash, hardware failures, or power outage should not result in data loss or downtime. The architecture of Hadoop is designed to ensure uninterrupted data availability through several major features.

HDFS replication stores several copies which by default is three, of each data block in various machines, ensuring data excesses. If a node fails, the system automatically retrieves data from another copy. In addition, Hadoop's self-healing architecture allows yarn to recreate or reallocate healthy nodes in the event of node crash, reduce system interruption. Distributed processing further increases mistake tolerance by distributing functions in many machines, reduces the risk of total system failure.

For example, a real -time traffic monitoring system, which collects data from the one hundred thousand IOT sensor, will also be able to continue operations, even if a server stores a server fails a server fails, thanks to the automatic recovery of backup copies of Hadoop. Without Hadoop's fault tolerance it is possible to lose valuable traffic data and potentially affect road safety and emergency response.

 doi:10.20944/preprints202603.1269.v1

1. Apache Kafka

The Apache Kafka is a distributed streaming technology that is made to stream data from Internet of Things devices in real time. Large data streams may be gathered, then stored, and processed in real time thanks Apache Kafka. For smart transport systems that depend on real-time data from several sensors and devices, Kafka is an excellent choice due to its capacity to manage enormous quantities of data streams (Kreps, 2022). Because it captures, stores, and duplicates massive amounts of real-time data flows, Kafka is a crucial component of the big data system even if it is not strictly a processing framework.

Figure Kafka Logo

Kafka can be thought of as a fast data pipeline that is meant to collect and transmit data in real time from several sources. This consists of Internet of Things sensors that gather information from parking spots, traffic signals, and automobiles. GPS devices that track the location of cars and passengers in real time are also used. Kafka can also keep track on social media accounts to get up-to-date information on accidents, road closures, and other delays in transit.

One of Kafka's primary characteristics is its high efficiency, which allows it to process enormous amounts of large data by handling millions of messages per second. Because of its scalability, the distributed system can accommodate increasing data sizes by adding more nodes.

By distributing data among multiple nodes, Kafka also offers fault tolerance, ensuring that data is accessible even in the event of a node failure. Last but not least, it is perfect for real-time applications because to its real-time data transmission capabilities, which provide minimal latency.

Kafka Architecture and Components

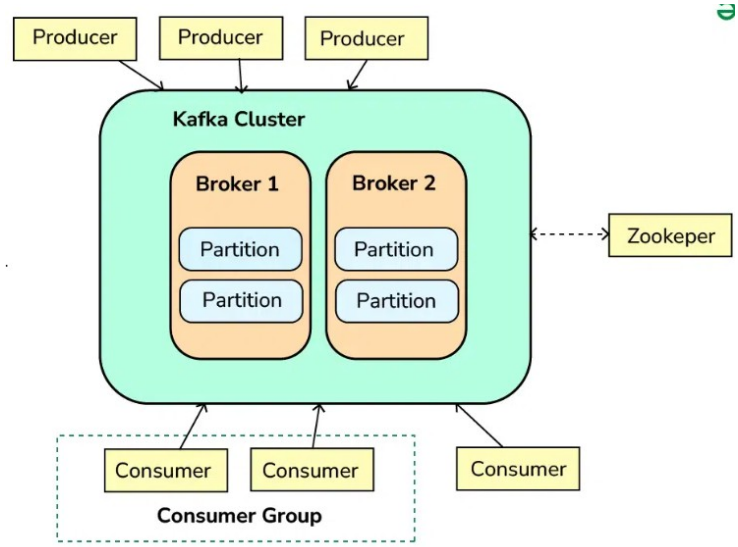


Figure Core components of Kafka's Architecture

1. Producers

Producers are responsible for generating or collecting and sending data to the Kafka topics. Kafka producer does not send message directly to consumers rather it pushes messages to Kafka

 doi:10.20944/preprints202603.1269.v1

server or broker.They publish messages independently to one or more topics this ensures that there is real-time data flow. In a smart traffic system, producers could include GPS devices, Vehicle sensors and roadside IoT units. Producers ensure that there is continuous stream of data, which is very important for real-time applications like navigation and traffic monitoring.

2. Topics

A topic in Kafka is a channel or category where messages are published and classified. Topics help to organize data efficiently by combining similar messages together. In kafka data is stored in topics where producers write their data to topics , and consumers read the data from these topics(*Apache Kafka Architecture and Its Components -the A-Z Guide*, n.d.). In a smart traffic system, different subjects can be created for traffic congestion updates, which are data about slow-running traffic or accidents. Road accidents which is for information about crash, construction areas, or lane closure. Weather conditions which include updates on temperature, rain, fog, and other factors affecting driving. Each topic can have multiple partitions, allowing scalability and parallel processing of large data streams.

3. Brokers

Brokers are very important components in a Kafka system, which act as a server for storing and managing message distribution. Each broker handles a portion or part of the Kafka workload, which ensures both scalability and fault tolerance within the system. When a manufacturer sends a message, brokers receive these messages and temporarily store until consumers reconstruct them. In a distributed Kafka cluster, many brokers collaborate to balance the load, effectively prevent any one point of failure. For example, if a broker fails, another broker can basically play its role, ensuring that the data is accessible to users.

4. Consumers

Consumers are applications that subscribe to Kafka subjects and process data. They consume messages of subjects at their speed, which ensures flexibility in data processing. In terms of traffic management, consumers can include traffic monitoring dashboards that imagine real -time traffic flows, navigation apps that provide optimal routes based on crowd data, and emergency response systems that detect accidents and inform the authorities. Additionally, consumers can be arranged in consumer groups, allowing many consumers to work efficiently in processing data from a subject.

5. ZooKeeper

Apache Zookeeper is an important component that manages the metadata and operation of the Kafka cluster. This leader plays several important roles, including elections, where it ensures that a broker is named as a leader to handle partition coordination. Zookeeper helps with configuration management by storing cluster settings and synchronizing changes in brokers.

Additionally, it monitors the broker health, monitors active brokers and detects any failure. Without zookeepers, Kafka groups will face significant challenges in broker coordination and handling failures effectively managing.

 doi:10.20944/preprints202603.1269.v1

Table Simplified explanation on Kafka components

Kafka Component	Description
Producers	Generate and send data to Kafka topics but do not send messages directly to consumers. They ensure a continuous real-time data stream. Examples in smart traffic systems include GPS devices and IoT sensors.
Topics	Channels where messages are categorized and published. Producers write data to topics, and consumers read from them. Topics can have multiple partitions for scalability. Examples: traffic congestion, accidents, weather conditions.
Brokers	Servers that store and distribute messages. They ensure scalability and fault tolerance by handling workload distribution. Multiple brokers work together to prevent failures in a Kafka cluster.
Consumers	Applications that subscribe to topics and process data at their own speed. Used in traffic monitoring dashboards, navigation apps, and emergency response systems. Can be grouped to enhance efficiency.
ZooKeeper	Manages Kafka cluster metadata, leader election, configuration, and broker health monitoring. Ensures proper coordination between brokers and prevents system failures.

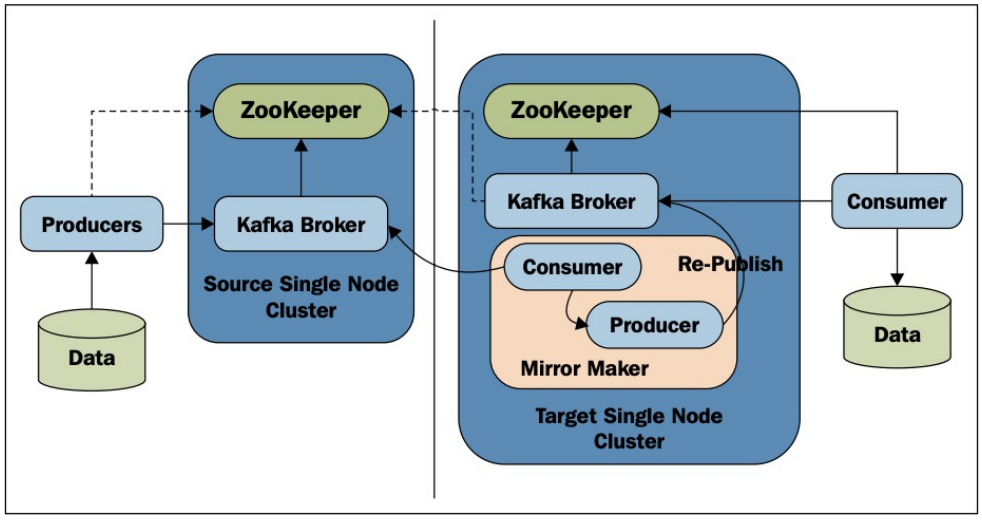


Figure showing an architecture of Kafka (Kafka in Action.Simon & Schuster,.2022)

Justification for Using Kafka

- **Real-Time Data Processing**

Kafka's main benefit lies in its ability to handle streaming data, which allows it to process information as it is being generated or collected. This capacity or ability is crucially important in traffic management, where the timely updates are important because they can prevent congestion, accidents or disabled routes plan. Kafka uses a published-co-conscious model, this is where manufacturers, such as GPS devices, vehicle sensors and IOT units, send continuously data, while

 doi:10.20944/preprints202603.1269.v1

the consumers, such as the navigation apps and traffic dashboards, process immediately and respond to this information.

Example Scenario:

For example, when a traffic sensor detects a sudden recession on the highway due to an accident, the sensor (the manufacturer) sends this information to the subject of "traffic congestion" in Kafka. A navigation app (consumer) then retrieves data and re -drives vehicles in real time, effectively reduces congestion and improves traffic flows. By ensuring low-lonely data streaming, Kafka enables authorities and systems to respond to road conditions within seconds, which increases public safety and efficiency.

- Scalability

In large metropolitan regions, transport networks generate large-scale data such as millions of connected devices, such as GPS trackers, smart traffic lights and ride-sharing applications. Kafka is designed for scalability, which means that this performance can handle high data throughput without a decline. This scalability is obtained through the subject division and a distributed broker architecture. Kafka subjects are divided into several partitions, allowing messages to many servers (brokers). When data production increases, such as during rush hours, Kafka may distribute loads in more brokers to maintain performance.

Example Scenario:

For example, during peak traffic hours, thousands of vehicles, weather stations and road sensors should be processed simultaneously. Kafka efficiently distributes this charge, ensuring smooth operation without bottlenecks. This scalability makes Kafka ideal for transport network that continuously expands, allowing cities to integrate new techniques such as autonomous vehicles and smart traffic lights without re -organizing the entire system.

- Fault Tolerance

Fault tolerance is an important aspect of traffic management systems because reliability is necessary; Failures can give rise to disruptions that affect public safety and emergency response time. Kafka is designed to be defective-tolerant, which means that it can continue to work, even if some servers, known as brokers, fails.

This fault is obtained through a process called tolerance replication. Each subject division consists of replicas stored on many brokers. If a broker fails, Kafka automatically turns into a replica to prevent data loss.

Example scenario:

For example, consider a city traffic control system that depends on Kafka to collect data from thousands of sources. If a broker who is storing the accident report becomes offline, another broker immediately handles. This ensures that the system continues to receive an accident alert and processing without any dissolution.

The high availability and flexibility of Kafka makes it an ideal solution for mission-mating applications, such as emergency response, autonomous vehicle coordination and real-time traffic management.

1. Apache Flink

The Apache Flink is an open-source stream processing framework that is designed at a scale for real-time and batch data processing. It enables high-throughput, low-latency data processing in a distributed and fault-tolerant manner, making it ideal for handling continuous data streams. Flink supports statistical computation which allows it to track and process events over time, which is

important for applications such as fraud detection and recommended systems. It basically integrates with large data technologies such as Apache Kafka, Hadoop and Cassandra. Along with its event-time processing and checkpointing mechanism, flink also ensure accuracy and flexibility in front of failures. Overall, it is a powerful tool for organizations that require real-time insight from their data.

 doi:10.20944/preprints202603.1269.v1

Figure Apache Flink Logo

Flink Architecture and Components

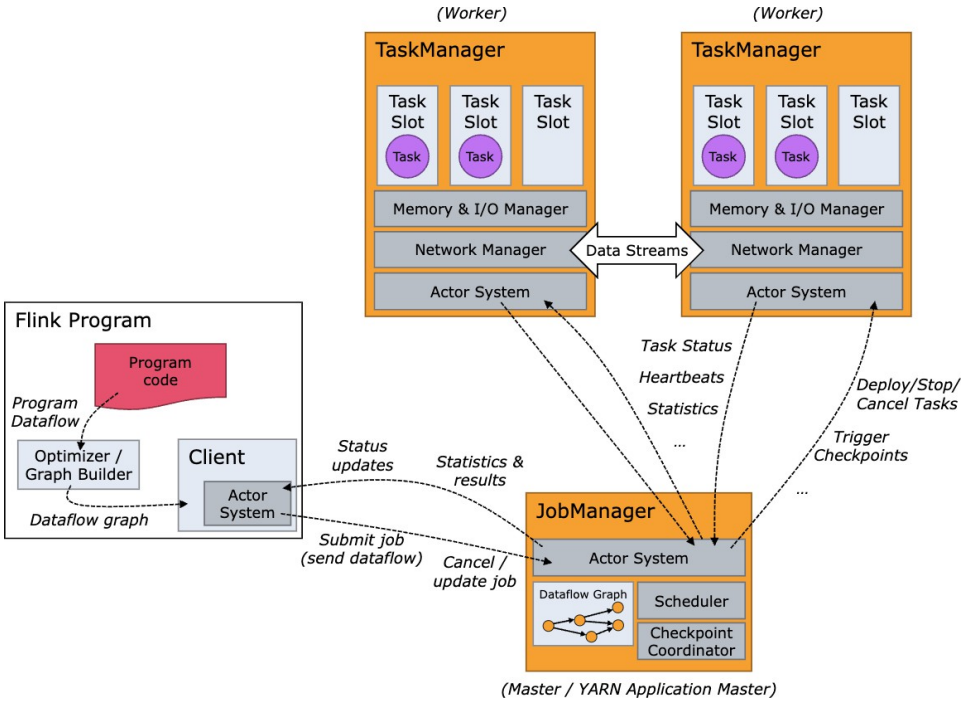


Figure showing fink architecture including its components (*Flink Architecture*, 2025)

1. Job Manager

The job manager acts as the central control unit within the flink, its job is for overseeing the execu@on of flink applica@ons, which are also known as jobs. It provides tasks to individual processing units, called the task managers, and op@mizes the execu@on plans to ensure efficient resource use. The Job Manager monitors the progress of the job, detects failures, and if necessary the rescheduling of failed tasks is done. To increase availability, a standby job manager can be deployed to take into case of failure, preven@ng any down@me.

2. Task Managers

The Task manager, also known as workers, is responsible for execu@ng the actual data processing works specified by the job manager. Each task manager has several task slots, which enables the flink to execute different tasks in parallel, for the func@oning of several CPU core and machines. Task managers store intermediate processing results and facilitates efficient data exchange between various tasks within a flink job. They communicate with each other to support distributed compu@ng and maintain state data when needed.

 doi:10.20944/preprints202603.1269.v1

3. Data Streams

Data streams represent the con@nuous flow of real -@me data in flink . These streams can arise from diverse sources, including IOT devices, transporta@on sensors, financial transac@ons, and log files. Flink supports both bounded (finite) and unbounded (infinite) data stream, which makes it suitable for both batch and streaming applica@ons. Framework processes data in real @me with low-latency analy@cs, which enables businesses and smart systems to make @mely decisions based on the latest informa@on.

4. Stateful Computation

Stateful computa@on in the flink allows applica@ons to maintain data in many processing stages. Unlike Stateless Stream Processing, where each event is handled independently, the stateful applica@ons recalls previous events, which enables complex event processing. Flink provides checkpoin@ng and savepoints to ensure the recovery of the state in case of failures, which is essen@al for applica@ons such as fraud detec@on, session tracking and predic@ve analy@cs. The state can be done to manage the data efficiently in the distributed nodes (divided by a unique iden@fier).

5. CEP (Complex Event Processing)

The complex event processing (CEP) in the flink allows to detect patterns, discrepancies and trends within the data stream. CEP can correct several data phenomena over time and identify sequences, such as abnormal vehicle movements that indicate a possible accident. This capacity is widely used in traffic monitoring, fraud detection, predictive maintenance and security surveillance [44–48]. By defining rules and patterns, flinks can trigger real -time alerts for abnormal events, empowering the decision makers to take immediate action.

Table simplified explanation on Flink components

Flink Component	Description
Job Manager	Central control unit overseeing Flink job execution. Assigns tasks to Task Managers, optimizes execution plans, monitors progress, and reschedules failed tasks. A standby Job Manager ensures high availability.
Task Managers	Execute data processing tasks as assigned by the Job Manager. Contain multiple task slots for parallel execution. Store intermediate results and support distributed computing by maintaining state data.
Data Streams	Represent real-time data flow from various sources (IoT devices, sensors, transactions, logs). Support both bounded (batch) and unbounded (streaming) data for real-time analytics and low-latency processing.
Stateful Computation	Maintains data across processing stages for complex event handling. Uses checkpointing and savepoints for failure recovery. Essential for fraud detection, session tracking, and predictive analytics.
CEP (Complex Event Processing)	Detects patterns, anomalies, and trends in data streams. Used in traffic monitoring, fraud detection, predictive maintenance, and security surveillance. Enables real-time alerts for immediate decision-making.

 doi:10.20944/preprints202603.1269.v1

Justification for Using Flink

- **Supports real-time responses to traffic incidents and congestion through event-driven processing**

Event-driven processing is a major feature of flink that enables traffic events and congestion to have real-time responses. This approach means that the system dynamically reacts to the data coming in real time rather than processing data in batches. In the traffic system, the flink can immediately process and analyze the traffic feed, which can enable immediate reactions to incidents such as accidents, road closure, or congestion. For example, if an accident is detected through the IoT sensor or camera feed, the flink may trigger an automatic response, such as redirecting vehicles or alerting emergency services.

This flink approach plays a very important role in improving traffic management by reducing delay in response time. This helps the authorities to take active measures to prevent further congestion and improve commuter experience by dynamic adjusting traffic signals and suggests alternative routes in real time [49,50].

- Ensures quick reactions in automated transport systems through low latency

Low latency is an important aspect for the performance of the flink this ensures that there is quick reactions in automatic transport systems. This means minimal delays between data input (example, an event detecting sensor) and system reaction (example, adjusting traffic signals or rebuilding vehicles). In smart transport, it is important to avoid low delay collisions and safely navigate for autonomous vehicles. Smart traffic lights can immediately accommodate signals depending on real -time traffic density, which can prevent unnecessary congestion. Emergency vehicles can achieve priority route by reducing delay in significant conditions.

Less delays increase road safety by ensuring rapid decision making, improves efficiency in automatic and connected vehicle systems, reduces travel time, and reduces fuel consumption and emissions by stopping stop-end-go traffic.

- Handles increasing traffic data with growing urban populations.

Scalability is a major feature of the flink that ensures that it can handle the increase of traffic data with an increasing urban population. Scalability refers to the capacity of a system to handle the increasing amount of data and workloads without the decline of performance. As cities grow, there is an increase in the number of vehicles that generate traffic data, pedestrians and sensors. Flink distributed architecture ensures that it can process millions of data points per second, and maintain smooth operation even during peak hours. Traffic officials can integrate new data sources (example, additional sensors, AI-based traffic cameras) without overloading the system.

This scalability is important because it is a future proof transport infrastructure against increasing urbanization. It ensures frequent performance, even data volume increases, and supports the smart city initiative, making transport more efficient and durable.

4. Ethical and Privacy Considerations in Smart Transport Systems

The rapid progress of Big Data Technologies has revolutionized the transport industry, which has improved clever systems, traffic flows and increased road safety. However, comprehensive use of large data in transport increases significant moral and privacy concerns. It is important to address these issues to ensure that individual privacy rights and moral standards are felt the benefits of data-operated transport systems.

1. Data Privacy

Data privacy refers to control a person's right to control, how their personal information is collected, used and shared. In transportation, it includes location, travel patterns and data related to personal details. Increasing use of GPS-competent devices, monitoring cameras and other data collection technologies has made it easier to track individuals, which increases concerns about unauthorized monitoring, profiling and discrimination.

To reduce privacy risks, transport systems must apply strong anonymous obvious policies controlling data collection, storage and use, and users should provide access, correction, or deletion

 doi:10.20944/preprints202603.1269.v1

of their information. The consent system should be transparent and clear. The article "Privacy and security challenges in smart and sustainable mobility" discusses the importance of data privacy in transportation, including the need for strong anonymization policies and transparent consent systems [52].

2. Data security

Ensuring safety of transport data is important to protect sensitive information from violations and cyber-attacks. The collection of large volumes of data poses risks of cyberattacks and data breaches [53]. The large dataset is often stored in a centralized database or cloud environment, giving them attractive goals for hackers. Common threats include data violations, cyber-attacks and supply chain weaknesses. To combat these dangers, organizations must adopt strong encryption, access control mechanism, regular safety audit, and up-to-date safety patch and employee cyber security training. An article on Artificial Intelligence & Transportation which addresses safety, privacy and ethical challenges highlights the importance of ensuring the safety of transport data to protect sensitive information from violations and cyber-attacks [54].

3. Regulatory compliance

Compliance with data protection laws ensures moral data usage and safety of individuals. Major rules include GDPR (General Data Protection Regulation) and CCPA (California Consumer Privacy Act) adhering to these regional data protection laws is crucial because they dictate how user data should be handled and protected [55]. By implementing clear data governance policies to the smart transportation system, ensuring the consent of the user and adopting the privacy-by-design principles should be aligning with these rules.

4. Real world ethical considerations

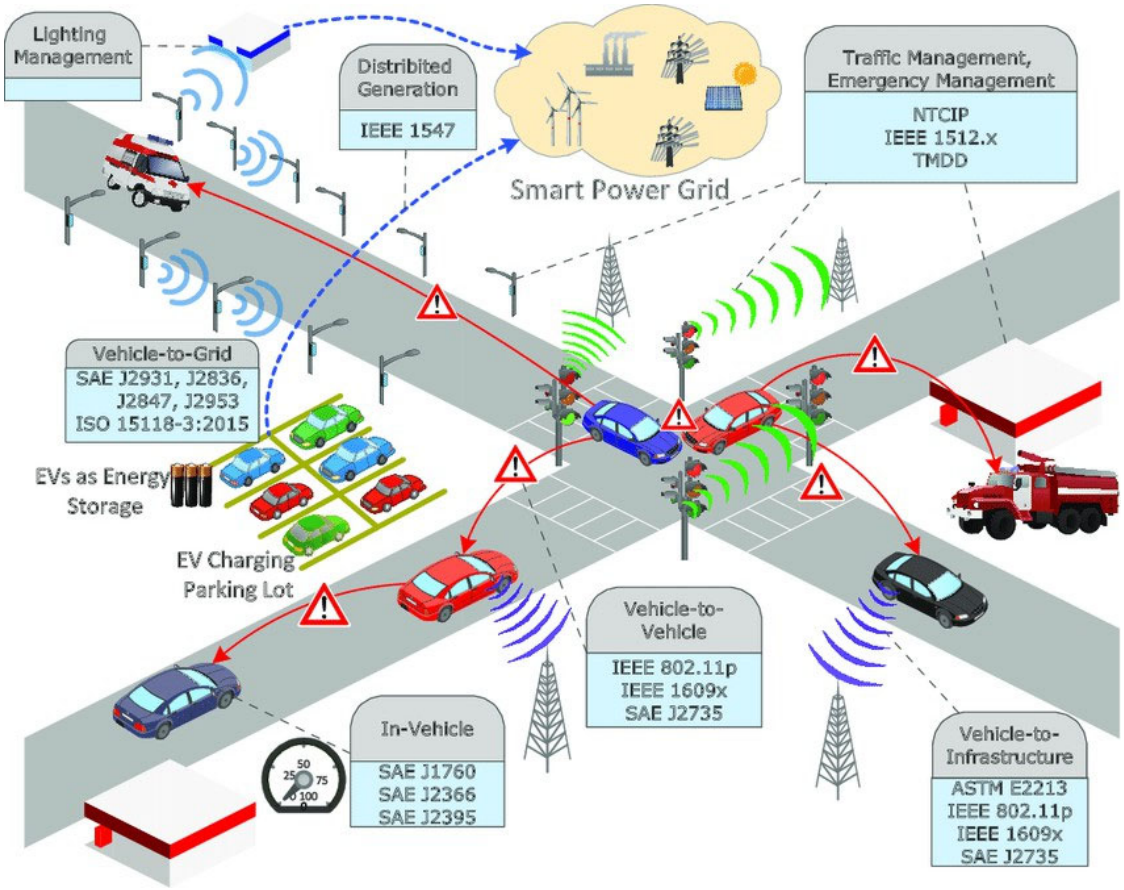
The city using GPS data to track vehicles for traffic management reflects both the benefits and moral risk of large data in transport. While such data can adapt to traffic flows and infrastructure plans, it also shows concerns about personal privacy. To balance innovation with moral responsibility, cities should ensure transparency, obtain consent, anonymity, strengthen security measures and follow rules. A new set of ethical issues are raised due to the use of artificial intelligence (AI) in transportation, especially in the form of autonomous vehicles (AVs). While there are many advantages to these self-driving cars, there are also serious concerns about public trust, safety, and decision-making algorithms because these autonomous vehicles make decisions on their own, lack human judgement and so many others all of which might affect the company's reputation and financial success (Bezrukov, 2024).

Overview of the Selected Study

Integrating intelligent and sustainable transportation systems in Jeddah: A multidimensional approach for urban mobility enhancement [56]. The article talked about the transport system in Jeddah, Saudi Arabia, it especially focuses on integrating intelligent and durable practices to solve its current and emerging challenges. The city has heavy dependence and reliance on central private vehicles an underdeveloped public transport networks, and on top of that unique climatic conditions that cause additional complications. The article adopts a multidimensional approach as it analyzes various factors, including technological progress in transport, durable and environmentally friendly practices, and cultural and policy transport systems shaped landscapes.

This research or article highlighted the immediate need for the city Jeddah to shift to a more efficient and environmentally friendly transportation options. This change is crucial to reduce current transport problems, ensure Sustainable urban development, and improvement in quality of life to the individuals or the people living in Jeddah. According to the research the application of Intelligent Transportation System (ITSS), developing strong public transport networks, encouraging non - neutralized mode of transport, and to incorporating permanent urban design principles is what is needed for a better transformation. The Intelligent Transport system is the application that has been used in the study to improve the transportation system in Jeddah, apart from Jeddah it has been applied in other case areas such as Dubai, Abu Dhabi, Doha , Kuwait , Riyadh and Muscat.

 doi:10.20944/preprints202603.1269.v1



Intelligent Transportation System.

An intelligent transport system (ITS) is an advanced application that integrates technology, communication and data analytics to improve transportation efficiency, safety and stability. It includes a wide range of systems that improves or enhances traffic management, public transport, road safety, and vehicle-to-vehicle (V2V) or vehicle-to-infrastructures (V2I) communication. The target is to optimize transport network, reduce congestion, improve security and reduce environmental impact through real -time monitoring and automation.

According to (Super User, 2023) based on big data analysis technology, the sustainable travel prediction of the intelligent transportation system (ITS) tracks the road network congestion in real time and predicts future transportation demand, hence adjusting the traffic signals in a targeted manner which then leads to optimizing route planning, reducing traffic congestion, and improving road capacity.

ITS manages transport operations using advanced information, communication and sensing technologies. With the acceleration of urbanization and the continuous increase in the ownership of the car, many problems have become increasingly severe, such as traffic congestion, environmental pollution and energy consumption. Its purpose is to solve these problems and achieve permanent transport development. Based on big data analysis technology, sustainable travel prediction of ITS monitors road network congestion in real time and predicts future transport demand, thus adjusting traffic signals in a targeted manner, optimizing route planning, reducing traffic congestion, and improving road capacity. By analyzing transport data, it provides appropriate travel suggestions for governments and enterprises, and effectively encourages passengers to choose public transport means, thus reduces road pressure, reduces the frequency of private car use, and receives appropriate resource allocation. At the same time, it helps reduce traffic congestion and driving time, thus reduces energy consumption and exhaust emissions, and improves urban environmental quality.

 doi:10.20944/preprints202603.1269.v1

5. The System Architecture and Workflow

The study's architecture for the intelligent transportation system is based on a number of different important elements that when combined form a unified and effective system. Each of these elements is essential to maintaining the efficient and effortless functioning of the transportation system. The following are the several components discussed in the study:

1. Data Collection Layer

The system collects data from different sources such as IoT sensors or devices, social media data, and public transport data. These different data sources generate data and provide or offer up-to-date details on the passenger demand, the location of vehicles, traffic conditions and even weather conditions that affect the transportation system.

- GPS Devices:

ITS relies on mobile and GPS-based data sources as they provide real-time location, speed, and travel patterns.Different mobile and GPS-based data sources play an important role in the providing of real-time information about vehicle movements and travel

patterns which contributes significantly to the development of smart transport systems. These data sources take advantage of embedded location-tracking capabilities in various devices and vehicles.

GPS data from vehicles and smartphones offers real -time space, speed and travel patterns. This data is necessary for traffic monitoring, root planning and congestion management. Connected vehicles, equipped with V2X communication technology, infrastructure (V2I), other vehicles (V2V), and data with the network (V2N). This communication enables advanced security facilities, such as conflict warnings and adaptive cruise control, and facilitates cooperative driving strategies.

- IoT Sensors and devices:

Internet of Things devices or sensors are widely used in the study as a data source to provide real-time data that is crucial for traffic management and incident detection. In the study it states that IoT provides drivers in a smart city with many benefits, including traffic management, improved logistics, efficient parking systems, and enhanced safety measures. The usage of CCTV streams and road sensory data from sources such as radar and LiDAR sensors, Automatic Number Plate Recognition (ANPR), and weigh-in-motion sensors has been discussed by several writers in various publications. Chen et al. proposed a wireless sensor network for Intelligent Transportation System (WITS), a prototype for intelligent transportation, as another IoT application. The WITS system is used to collect and transfer the data produced by various vehicles and other sources. Additionally, Deep Laxmi et al. developed an innovative smart transport system called the Smart Vehicle Assistance and Monitoring System (SVAMS) using the Internet of Things.

- Communication Networks:

Communication networks such as cellular networks like 4G, 5G , Wi-Fi and other short-range communications that enable data exchange between vehicles, infrastructure and control centers were also presented in the study. It is noted that 5G technology aims to connect different vehicles through Cooperative Intelligent Transport Systems (CITS) (Oladimeji et al.,2023).5G helps enhance safety and efficiency of automated transportation hence improving the transportation system. The Jeddah-based study demonstrates that ITS depends on a network of communication systems, including vehicle-to-everything (VTE) communications technologies, that provide data transfer between sensors, control centres, vehicles, and road user. The use of communication protocols (Wi-Fi, 4G/5G, TCP/IP) researchers obtain an advanced and improved transportation systems.

- Infrastructure Data sources:

Infrastructure data sources provide important information about roads, traffic lights and transport facilities, contributing significantly to the development of smart transport systems. These data sources provide insight or understanding into the physical infrastructure and its performance, which enables better traffic management and adaptation.

Traffic signals and controller light cycles, pedestrian crossings and data on adaptive signal control provide data to the intelligent transportation systems. This information is very crucial for the

 doi:10.20944/preprints202603.1269.v1

analyzing of traffic flow patterns and optimizing the signal timings so as to reduce the congestion and improve traffic efficiency. Roadside units (RSU) act as communication hubs which collect. and transmit transport data between infrastructures and vehicles. RSUs collect data on the traffic amount, speed and vehicle types, which then provides real -time insight into the road conditions.

Public transit systems also generate valuable data including schedule, passenger load and delay. This information enables adaptation of routes, schedules and resource allocation to ensure efficient and reliable public transport services.

- Crowdsourced & Social Media Data:

User-related information provides valuable insight into the real-time traffic conditions which allows for a more accurate and responsible traffic management.

Crowdsource traffic apps such as Google Maps, Wazes, and Maps rely on real -time traffic reports from the users. These apps collect data from different users' smartphones and provide information about the traffic flows, congestion and potential delays. Social media platforms such as Twitter and Facebook also play an important role as valuable sources of information. Users often share updates about accidents, obstacles and delays in public transit, which provide a real - time view of the transport landscape.

Civil reporting apps empower users to actively contribute to traffic management. These apps allow users to report road conditions, pits or accidents, which directly ensure quick reactions to relevant officers, and improve road safety.

- Weather and Environmental Data:

Weather and environmental data are very important to understand the impact of external factors on the transport systems. Weather conditions such as rain, fog and temperature, can greatly affect the road conditions as they can cause delays, accidents and disruption.

Weather stations and satellite data provide real -time information on these weather conditions which enables active traffic management and route planning. Air quality sensors check or monitor pollution levels hence assessing the environmental impact of traffic and provide valuable data for urban planning and emission reduction strategies. This data helps identify areas with high pollution levels, enabling targeted intervention to improve air quality and promote permanent transport.

- Administrative and Historical data :

Administrative and historical data provide valuable insights on long -term traffic patterns, safety trends and road use which has been maintained by government agencies and transport officials.

This data is required for informed decision making, infrastructure scheme and optimization of transport operations.

Traffic volume and flow data, collected in the extended period, reveals the historical pattern of the vehicle movement. This information helps identify recurring congested areas, enabling target intervention to improve traffic flows and optimize the ability of infrastructure. Accident and event reports provide significant data on high -risk areas, allowing authorities to implement road design to implement safety measures and reduce accidents. In the article it stated the main method of data collection will involve reviewing secondary sources such as government reports, urban planning documents, policy papers, and scholarly which will provide a strong base of existing knowledge and official perspectives on transportation.

2. Data Transmission Layer

The communication layer in an intelligent transport system (ITS) act as a bridge between data collection layer (perception layer) and data processing layer. Once collected from data sensors, vehicles, road infrastructure and public transport systems, it should be transmitted to central control units, cloud servers, or edge computing devices for further analysis and decision making.

This data transmission is to ensure real-time reactions and efficiency through various wireless communication technologies, cloud-based platforms and edge computing. The few data transmission layers that enable data transmission are as follows:

- Distributed Computing:

 doi:10.20944/preprints202603.1269.v1

Distributed computing is a computing model that involves the dividing of tasks among multiple computers or nodes that work together to achieve a common goal. Applications for smart transportation have been supported and presented using centralized computing such as cloud computing but this has brought out different challenges in transporting and processing transportation-related data, such as CCTV streams or road sensory data, due to the constantly growing amount or number of linked vehicles. Due to the challenges, currently many applications require distributed data processing so as to reduce latency and have the ability to manage large volume of transportation data.

In order to escape traffic congestion, the drivers in urban areas frequently need to make immediate decisions about changing lanes or routes in order to reach their destinations. Different applications must collect and analyze the necessary information in real time, including location, speed, traffic flow, and accidents, in order to assist the drivers in making these judgements.

However, quickly gathering and analyzing massive amounts of data is very difficult for the cloud infrastructures. This can be reduced while still achieving the necessary efficiency with a distributed data processing system.

Chen et al.'s Wireless Sensor Network for Intelligent Transportation System (WITS) is an illustration of a smart transportation system. Three different kinds of nodes vehicle units, roadside units, and intersection units are used in this system to collect and transmit data. The roadside unit receives the parameters that are gathered by the vehicle unit. Based on predetermined objectives, the intersection unit evaluates this data to identify the best course of action.

Hull et al.'s mobile distributed sensor computing system, CarTel, serves as another illustration. CarTel is made up of embedded mobile computers that are linked to a number of sensors. Data is processed locally by each node before being sent to a central gateway for storage and additional analysis. With a focus on traffic data, CarTel improves the gathering, processing, and display of various information from mobile devices.

Within distributed computing there is grid computing that also allows the transmission of data in intelligent transportation systems.

- Grid Computing:

Grid computing is a distributed computing model that connects multiple computer systems and different resources across different locations so as to work together on complex or difficult tasks. As per the article grid computing uses numerous computer resources to work together and it is loosely coupled to solve a specific problem. In order to use resources that are available effectively, grid computing divides a big task into numerous work stations. The Shanghai Transportation Information Service Application Grid (STISAG) is one example of the use of grid computing that is pointed out in the paper (Oladimeji et al.,2023). A wide range of real-time traffic and travel information services are offered to end customers by the Shanghai Transportation Information Service Application Grid (STISAG), which focusses on the problem of traffic congestion in Shanghai.

- Edge Computing:

Edge computing can be defined as the use of different technologies that allow data processing to happen right at the edge of the network, close to where the data is generated. The exploration of edge-cloud computing is being done by established cloud providers with the aim of improving their services so as to meet the growing demand of users. Compared to cloud computing, edge computing lowers latency by enabling data processing at the network edge. Roadside infrastructure and road users get control signals from Intelligent Transportation Systems (ITS), which use remote sensors to assess road conditions in real time. Vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication will be necessary for future ITS in order to exchange driving plans, alert the drivers to potential dangers, and to improve traffic flow. In order to get alerts or open up lanes for emergency vehicles, automobiles can also communicate with roadside infrastructure.

The operating costs of mobile networks and the remote data centers cause latency problems for traditional cloud computing. Cloud providers have responded to this through the introduction of edge-cloud computing, which has enabled computation to take place in closer range to the user. Edge

 doi:10.20944/preprints202603.1269.v1

computing improves real-time traffic management and collision prevention while reducing reliance on the centralized data centers. In order to ensure the effective and safer transport networks, edge computing is essential in allowing faster, more dependable, and flexible ITS solutions as user demands or expectations change towards wireless and mobile networks.

Edge computing is crucial for intelligent transportation system as it reduces latency by ensuring that there is real-time decision making. It also enhances or improves reliability whereby the traffic systems remain operational even when connectivity is lost.

- Wireless Communication Technologies

Wireless communication is very important for intelligent transportation systems (ITS) as it enables spontaneous interaction between the different components of the transportation system. The 5G/4G LTE cellular networks provide high speed, low latency data transfer for real-time vehicle communication and transport management. It enables vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) to communicate, supports self-driving cars, helps transport agencies to monitor road conditions, and allow emergency response teams to receive immediate alerts about accidents. 5G ultra-fast speed and ultra-low latency are important for autonomous driving and smart traffic systems. While Bluetooth and RFID are used for short range communication in applications such as toll collections, parking systems and traffic monitoring. These technologies help reduce manual intervention, speed up toll payments and improve traffic flows. By the use of communication protocols (Wi-Fi, 4G/5G, TCP/IP) researchers obtain an advanced and improved transportation systems.

1. Data Analysis Layer:

The analytics layer is the layer that is responsible for processing and interpreting transportation data collected from different sources such as from GPS devices, IoT sensors, transport data and others. This component enables data collected or gathered to be efficiently managed and

analyzed by using different advanced technologies. The processing layer transforms raw data into actionable insights hence enhancing the decision-making processes. This is the layer that transforms raw data into meaningful insights that help improve the traffic flow, enhances road safety, and optimize the public transport operations. The following are the different functions of the layer that were presented in the study:

Functions of the Processing Layer

1. **Incident Detection and Prevention**

Analytics layer is responsible for filtering and checking the accuracy of data so as to improve security and prevent misfortunes and incidents in transportation networks or even detect them. A critical application of the ITS is incident management, which enables quick detection and response to road accidents or incidents, reducing their impact on traffic flow and enhancing road safety. According to the study, V2V, V2X, V2I, and V2P have applications in collaborative driving assistance, decentralized probe vehicles, and information communications. With the help of this technology, cars can notify other cars to avoid collisions when changing lanes. Singh et al proposed there being a Wi-Fi equipped highways with dashboard systems that record real-time accident footage, hence providing analytics or data for accident prevention. According to (Alamoudi et al., 2024), the ITS's integration of advanced technologies like dynamic message signs, traffic monitoring cameras, incident detection devices, and a widespread fiber-optic network for communication has improved incident monitoring by 63%, slashed emergency response times by 30%, and reduced travel times in Dubai by 20%.

2. Real-Time Data Analysis

In the study this is discussed through the linked traffic cloud that collects and analyzes real-time data from connected cars, infrastructures, and devices to help with operational decision making, improved navigation, fuel consumption, and time resource optimization and so much more. The study shows or demonstrates how real-time data analytics is very important for the evaluation of large amounts of transportation data, which can then be applied to enhance public transportation

 doi:10.20944/preprints202603.1269.v1

services, enhance traffic prediction , and optimize routes. According to Alamoudi et al. (2024), this technology can be very helpful in Jeddah for controlling the traffic patterns, particularly during rush hours and major occasions like the Hajj or Ramadan.

This shows how data is gathered and processed to get different valuable insights from the transportation system which is in alignment to the analytics layer.

3. Public Transportation optimization

Collecting, processing and analyzing historical and live traffic data is a function of analytics layer that helps in the improvement of public transportation system. In the article the VBA route planning algorithm introduced by Chang et al was discussed which uses real-time traffic data from vehicular ad-hoc network (VANETs) and Google Maps to determine optimal travel routes hence improve public transportation. The report on the integration of Intelligent and Sustainable Transportation Systems in Jeddah claims that ITS is very crucial to improve the effectiveness of public transportation since it makes fleet management easier through route optimisation, through maintenance scheduling, without forgetting vehicle location tracking. The report shows that iorder to improve the efficiency and consumer satisfaction, some systems need to apply or use data analytics to modify public transport services in response to real-time demand. As a component of ITS, automated and intelligent ticketing systems make it simple to pay for and access public transport, hence promoting its use (Alamoudi et al., 2024).

4. Traffic flow optimization

Intelligent Transportation System (ITS) use AI and data-operated approach to dynamically optimize traffic flows, which adjust traffic signal timing in real-time based on congestion levels,

analyze road network correlation to suggest alternative routes, and rely on decision support systems to make decisions.

As per the article (Alamoudi et al., 2024) Applications of ITSs in traffic management are varied. One such application is traffic flow optimization, where traffic data is utilized to adjust signal timings and manage traffic dynamically to minimize congestion.

Key Technologies in the Processing Layer

1. **Big Data Analytics**

Big data analytics include large scale collection and processing of traffic related data to identify patterns and trends. This data is used to predict the level of traffic, forecast travel demand and explore incidents such as accidents and obstacles. For example, Google Maps analyzes real-time GPS data from mobile phones to estimate traffic situations, while traffic control centers use video feeds for monitoring crowds. By analyzing historical and real-time data, Big Data Analytics helps the city planners to estimate the peak time and optimize the public transport program.

2. Artificial Intelligence and Machine Learning

Artificial intelligence (AI) and machine learning (ML) use algorithms to analyze data, learning from patterns and to improve decision making over time. In this, AI and ML are used to adapt to traffic signals by adjusting signal timing dynamically based on traffic flows, predicting potential accidents hotspots based on drivers' behavior and previous events, and assist autonomous vehicles by processing the sensor data for real-time decisions. For example, AI-operated smart traffic lights reduce the waiting time on the intersections, and the deep learning models analyze the camera feed to detect Jaywalking pedestrians.

3. Graph- Based Models

Graph-based models, such as graph convolutional network (GCN) and graph Attention network (GAT), analyze road networks and traffic relations using mathematical graph structures. The GCNs help to understand the complex road network by capturing the spatial dependence between the connected road segment, while the Gats improve demand prediction by focusing on significant traffic routes. These models are used to analyze traffic flows, identify crowded routes, make optimal detours and optimize public transport by predicting the urban population density and traveling on

 doi:10.20944/preprints202603.1269.v1

commuting patterns. For example, a city can use a GCN model to improve the metro station placement by analyzing the passenger movement pattern.

4. Decision Support Systems (DSS)

The decision support system (DSS) provides recommendations to transport officials for effective traffic management. DSS is used to help traffic operators respond to congestion, accidents and weather disruptions in traffic management centers (TMCs), real -time passenger helps to prioritize emergency vehicles by adjusting bus and train time based on passenger demand and dynamically adjusting traffic signals.

1. User Interface Layer (Application Layer)

Application layer, also known as a user interaction layer, serves as an interface among ITS and users including drivers, pedestrians, public transport users, traffic operators and emergency respondents. This layer ensures that the processing layer has been presented in a meaningful way to improve real -time data, transport efficiency, road safety and overall user experience.

Application of Application Layer

1. Traffic management systems

These systems provide real -time monitoring and control of road networks. Traffic Control Center (TCCS) allow officers to monitor and manage urban traffic flows. AI-based systems can

detect accidents or crowds and can take corrective action through automatic event management, such as re-starting traffic. Intelligent toll collection systems use RFID and GPS for seamless toll payments, reducing the congestion in toll plazas. For example, the emergency response time has decreased by 30% as a result of the Dubai Smart Traffic Management System (Alamoudi et al., 2024).

2. Smart Traffic Lights and Adaptive signal control

This function improves the traffic efficiency by adjusting the signal time continuously depending on the real-time traffic conditions. Adaptive Traffic Signal Control (ATSC) uses AI and IoT to analyze congestion levels and adjust signals accordingly. Traffic signals can prioritize emergency vehicles, turn into green color to go near ambulances, fire trucks and police vehicles. Smart crossings detect pedestrians and expand the duration of green light as required for safety.

3. Navigation and guidance app

This function offers the optimal route recommendations and real -time traffic updates to users. GPS-based apps such as Google Maps, Ways, and Apple Maps use real-time data from IOT sensors and connected vehicles to suggest the fastest routes. AI-based traffic predictions in apps predict congestion levels and suggest alternative routes based on historical data. Vehicle-to-Navigation (V2N) enables communication vehicles to communicate directly with navigation systems for better routing.

4. Public transport optimization

This increases public transport operations by providing real-time transit information and optimizing scheduling. Live bus and train tracking apps display exact arrival time and departure programs. UBER, LYFT and other mobility services can be used directly through the public transport app through ride-sharing integration. The AI fleet tracks and optimizes the deployment of buses, trains and metro services based on the demand pattern through management.

Data Flow in the System

Data flow in a system refers to how data or information moves through different parts of the system. In smart transportation data flow consists of data collection, data transmission, data processing, decision making and user interface. Intelligent Transportation System (ITS) rely on a strong data flow process to effectively manage, optimize the routes and increase safety. This process includes several major stages:

- **Data Collec-on**: Its systems collect real-@me data from various sources, including vehicle-to-vehicle (V2V), vehicle-to-infrastructure (V2I), and vehicle-to-everything (V2X) communica@on networks. IoT sensors and GPS devices provide insights such as the traffic volume, vehicle

 doi:10.20944/preprints202603.1269.v1

movement, conges@on levels and environmental condi@ons. Also, 3D cameras and RFID systems monitor the presence of vehicles, availability of parking and road obstacles. Apart from that roadside units (RSU) collect data from traffic signals and smart highways, which contribute to the widespread understanding of the transport network.

- **Data Transmission**: The collected data is transmi]ed to processing centers using wireless networks such as 5G, Wi-Fi, and DSRC which ensure that there is real-@me data exchange. In addi@on to that cloud-based systems and edge compu@ng manage large-scale data processing with minimal delays, able to make efficient analysis and decision making.
- **Data processing and analysis**: The system appoints AI and machine learning algorithms in order to analyze traffic pa]erns, predict conges@on levels and customize routes. The decision support system processes the data to generate ac@onable insights, such as the op@mal route planning for drivers , accident detec@on and preven@on, and smart traffic light control based on real -@me data.
- **Decisions and execu-on**: AI-driven decisions lead to dynamic traffic signal adjustment to reduce conges@on, naviga@on guidance and rerou@ng for drivers, public transport schedule adapta@on based on traffic demand, and emergency response ac@va@on in case of accidents.

User interac@on and feedback loop: Processed informa@on is displayed on road signals, on mobile apps, in public transport systems and on smart parking applica@ons hence providing the drivers and passengers with real -@me updates. The system con@nuously learns from feedback and the future model updates thus increasing the future decision making and ensures more efficient and responsible transport systems.

6. Dataset Introduction and Preprocessing

The dataset used in this study arises from an experimental application of edge computing in the intelligent transportation system (ITS). The primary goal of the dataset is to analyze the effect of edge computing on traffic monitoring, congestion mitigation and accident prevention. The data was collected from five separate highways under different circumstances, including various light landscapes, weather conditions and traffic density.

Dataset consists of several major matrix:

Emergency detection rates: data comparing the number of increased emergency situations with traditional ITS and and ITS which incorporates edge.

Response Times: Time taken by monitoring system to detect and react to discrepancies in traditional ITS vs. edge computing-based ITS.

Accident rate:The impact of Edge computing on reducing accident rates in various highways. **Traffic congestion levels**: ability to monitor and reduce congestion during various periods such as rush hours, holidays and free flow periods.

Congestion duration : Comparative data shows how the edge computing reduces the duration of the congestion.

The dataset enables a comprehensive analysis of how the edge computing optimizes this by improving real -time monitoring, reducing delays in reactions, and increasing the overall traffic safety and efficiency.

 doi:10.20944/preprints202603.1269.v1

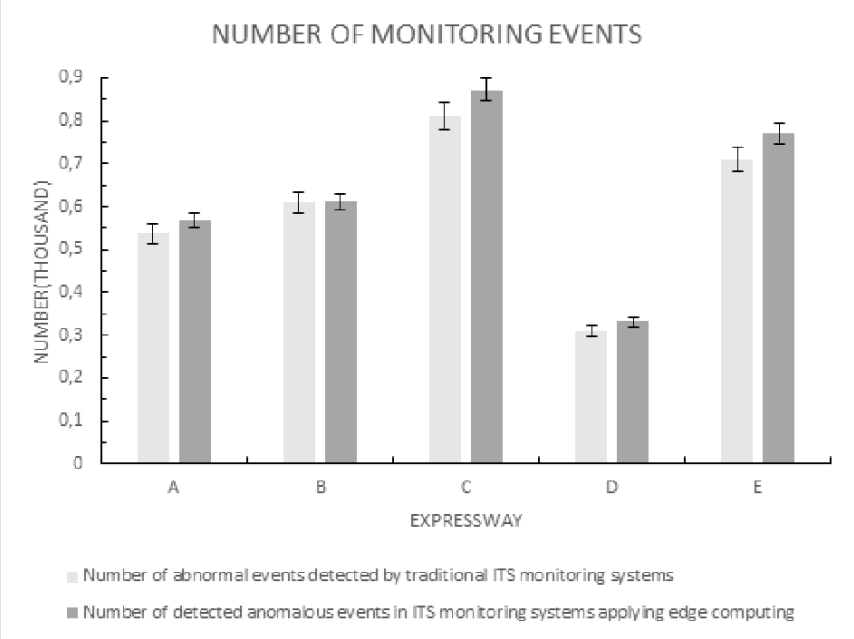


Figure on detection capability improvement of ITS monitoring system by edge computing

Both the traditional monitoring system and the monitoring system that follows the application of edge computing are tested in the above figure. After the implementation of edge computing on the different five highways, it was then discovered that the number of emergencies that were identified by the system's monitoring equipment had increased. While Highways C and D are chosen for times when visibility is poor, such as during rain, snow, or dense fog, Highways A and B are chosen for times when there is adequate lighting and excellent lighting conditions.

When there is poor lighting, Highway E chooses the tunnel section for a thorough comparison.

Shang and Wang (2024) According to the experimental findings, the edge computing monitoring system's average index of emergency monitoring is 5.7% higher than the traditional monitoring system's.

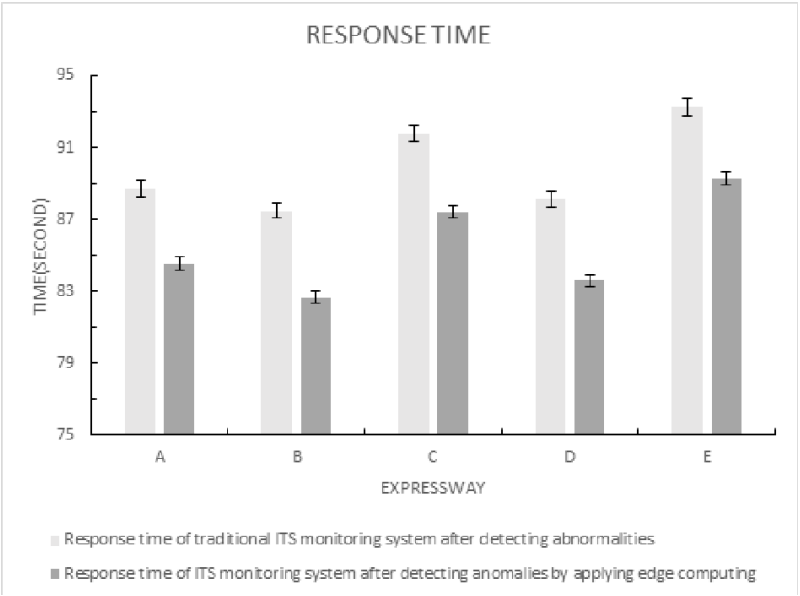


Figure on the enhancement of edge computing on the surveillance system's response time after identification of anomalies (Wang & Shang, 2024)

 doi:10.20944/preprints202603.1269.v1

The response times of the traditional monitoring system and the monitoring system following the implementation of edge computing are tested in the above figure. On the other hand, it has been discovered that edge computing technology can increase the monitoring apparatus's response time.

Among these, Highways A and B are chosen for every day operations, Highways C and D for nighttime operations, and Highway E for permitted holidays. According to the study's (Wang & Shang, 2024)findings, using edge computing technology increases the monitoring system's response time by 4.9%.

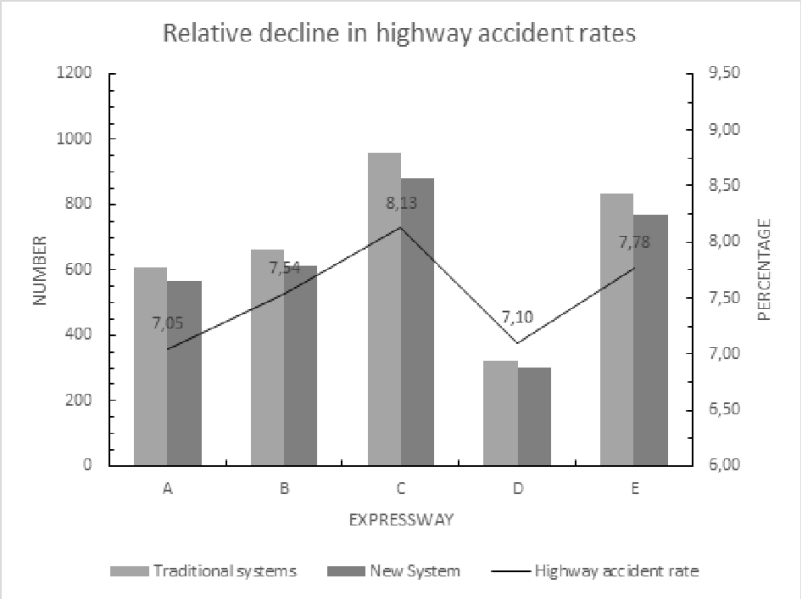


Figure on Edge computing's positive effect in reducing the amount of express-ways accidents

The statistics on the accident rate of the chosen roads are displayed in the above figure, and overall, the road with the edge computing-based monitoring equipment has a lower traffic accident rate. Highway E has a generally moderate traffic flow, Highways C and D have a comparatively high traffic flow, while Highways A and B have a relatively low traffic flow. The use of edge computing technologies has resulted in a 7.63% decrease in the overall rate of motorway accidents (Wang & Shang, 2024). It is clear that the monitoring system's capabilities have been greatly enhanced by edge computing. It guarantees travel safety in addition to increasing emergency detections, decreasing accident rates, and speeding up response times.

 doi:10.20944/preprints202603.1269.v1

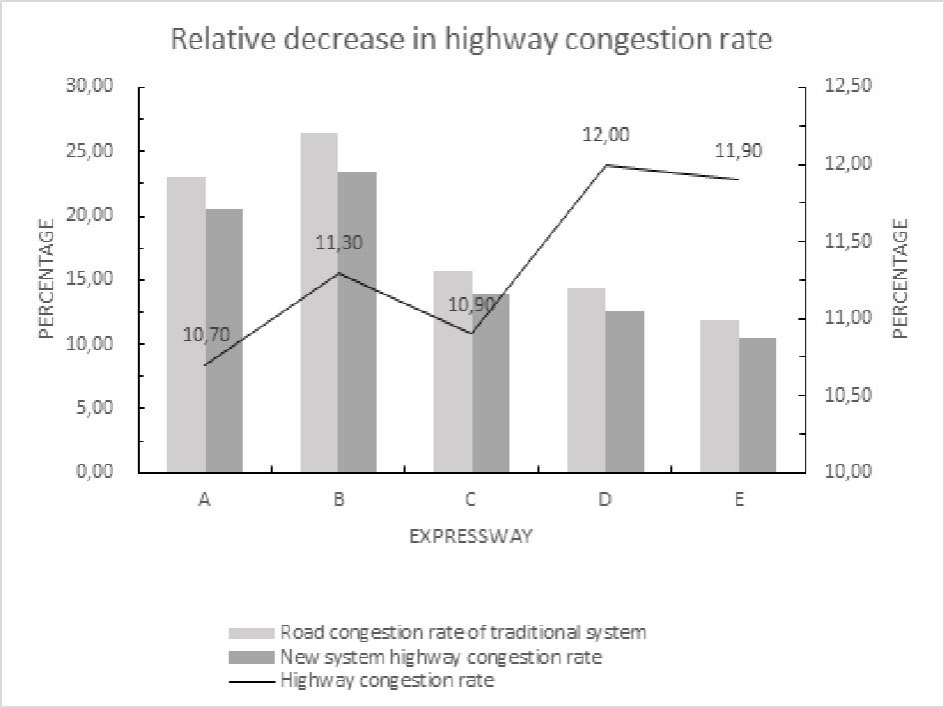


Figure showing the Alleviation of highway congestion rate by edge computing (Wang & Shang, 2024)

Roads C and D are chosen for the free time periods on highways during holidays, whereas roads A and B are chosen for the morning and evening peak commute times, according to the above figure. This study chose a time period during which Highway E's traffic volume is lower than the former's and conducts a thorough comparison. According to the study, edge computing technology reduces ITS traffic congestion by 11.27% (Wang & Shang, 2024).

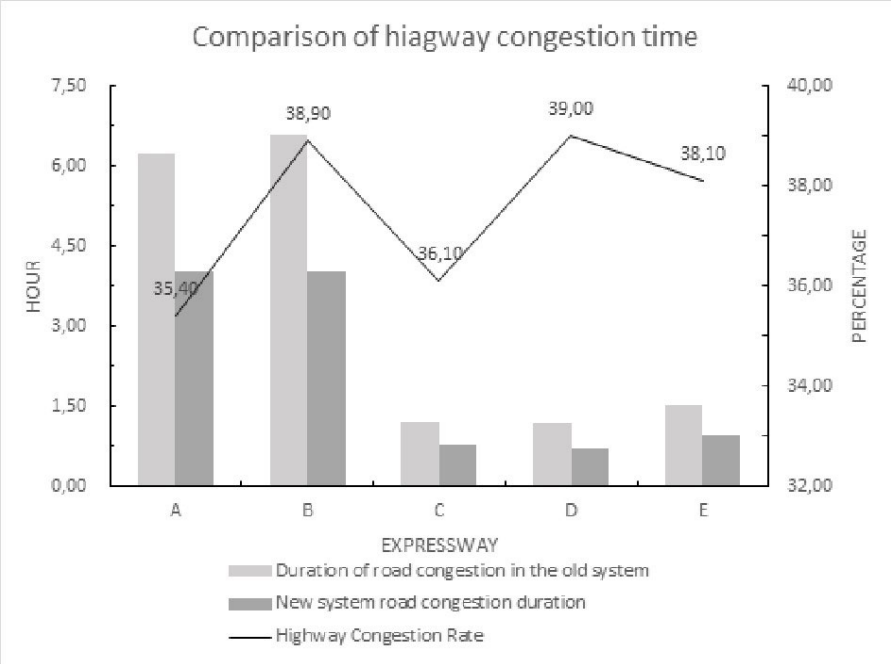


Figure Demonstrating how edge computing can help reduce high traffic congestion time (Wang & Shang, 2024)

 doi:10.20944/preprints202603.1269.v1

Roads C and D are chosen for periods of high morning and evening commuter traffic, roads E are chosen for periods of comparatively low traffic flow in comparison to the former, and roads A and B are chosen for free highway usage during holidays. According to the data results, using edge computing technology reduces ITS congestion duration by 37.3% (Wang & Shang, 2024). It is evident that edge computing helps the intelligent transportation system make the right choices. In addition to keeping an eye on the amount of traffic on the highways, it also reduces congestion and guarantees passenger comfort.

Preprocessing Steps

Data Cleaning

The process of finding and eliminating incorrect, inadequate, redundant, or unnecessary data from a dataset in order to guarantee its reliability and accuracy even before analysis is known as data cleaning.

Preprocessing includes data cleaning in the first step, which ensures dataset's accuracy and reliability which involves the identifying and removing of duplicate records, eliminating irrelevant data points and handling missing values. Missing values are addressed using techniques such as mean/mode replacement for numeric variables and classified data copying, ensuring that the data is complete and consistent.

Data Transformation

The process of changing a data's format, structure, or scale so that it can be used for analysis, machine learning, or storage is known as data transformation. Data standardization, structure, and ease of interpretation are ensured by this stage.

In the study data transformation are necessary to prepare data for analysis and modeling, this involves converting time-series data into structured formats, facilitating trend analysis.

Additionally, generalization of emergency detection rates, response time and congestion levels is important to ensure comparability in various highways. This allows for meaningful comparison and analysis of trends.

Feature Engineering

The process of developing new features or changing current ones in order to enhance the performance of machine learning and data analysis models is known as feature engineering. In the study of intelligent transportation systems feature engineering focuses on creating new, informative features from existing data. This includes percentage improvement in emergency detection, accident rates, and traffic congestion reform indices which are derived metrics. Rush hour indicators increase the predictive modeling. Also capturing temporary patterns in data, including, seasonal variations and day/night classification enables effective prediction.

Data Aggregation and Grouping

The process of summarizing and combining data to provide valuable knowledge is known as data aggregation. Just before using aggregation methods, grouping involves organizing data into categories according to shared characteristics.

In the research on Intelligent transportation systems data aggregation and conversion involves summarizing the data gathered based on specific criteria which includes groups based on highway types, traffic density levels and time patterns. Classifying the records based on environmental factors, such as weather conditions (fog, rain, snow) and road classification (urban highway, rural highway, tunnel), provides valuable insight into affecting various factors.

 doi:10.20944/preprints202603.1269.v1

Analysis of Exploratory Data (EDA)

The process of analyzing datasets in order to identify their significant characteristics, discover different trends, detect anomalies, and obtain different valuable insights is known as exploratory data analysis, or EDA.

In intelligent transportation system study, understanding the data and identifying any patterns and links requires the use of exploratory data analysis. To do this, visualizations of trends such as the emergency detection, the response time, and congestion reduction must be made or created.

Finding the connections between the use of edge computing and increases in traffic safety is made easier by performing correlation analysis.

7. Findings, Limitations, and Recommendations

Key Findings From the study

1. Better traffic movement

Improvement in traffic management that is brought from the use of intelligent traffic management technologies is one of the most important results of the study. Traditional traffic control systems, such as fixed-time traffic signals, often do not adjust the current traffic status, resulting in unnecessary delays and stops. On the other hand, the suggested method makes the use of real -time traffic monitoring through machine learning models and Internet of Things Sensors to guide efficiently cars, improve road use and continuously modify traffic signals.

For example, to make timely decisions of indications in real time, the machine learning-based adaptive traffic signal management system assesses traffic data coming from several sources,

including cameras, GPS-equipped cars and roadside sensors. These adaptive systems can reduce stop-end-go traffic patterns, resulting in selecting specific lanes during peak hours or by changing the signal duration based on the number of vehicles (Zhang et al., 2023), the congestion and fuel waste are specified.

In addition, automatic traffic routing systems are required to increase its transfer nature of vehicle movement. By offering the recommendations of the real -time route depending on the condition of the roads, these systems reduce deviation in travel time and enable more consistent distribution of cars in the road system. According to studies, when properly applied, smart routing systems can delay the delay in overall traffic by more than 30% (chain et al., 2022).

2. Reduced Congestion

In cities, traffic congestion is still a major problem resulting in travel delays, such as high emissions of greenhouse gases such as carbon dioxide and financial losses. By continuously controlling the amount of traffic in various routes and distributing vehicle loads, the study shows how effective congestion management strategies greatly reduce the issue.

Preventive congestion control, which uses both historical and current traffic data to identify the first crowd areas, is one of the primary approaches used in the study. Machine learning models can predict crowd locations and actively modify traffic signs or recommend other routes to avoid jams using traffic flow patterns, street occupancy level and vehicle speed data (Singh et al., 2023).

The study also sees how a vehicle-to-infrastructure (V2I) can help reduce connectivity traffic. Organized reactions to traffic problems are made possible by V2I technology, which enables direct communication between automobiles and control centers, road sensors and traffic lights. For example, neighboring traffic signs can be coordinated to allow a long green light if there is a significant traffic accumulation at a certain intersection, which will reduce the delay and increase the efficiency overall.

Demand-based traffic control, which modifies traffic lights and road access restrictions according to the current demand pattern, is another essential component of congestion deficiency. According to the studies, demand-based solutions reduce the average of 25–40% peak-hour traffic in cities (Ahmed et al., 2021).

3. Optimal Allocation of Resources

 doi:10.20944/preprints202603.1269.v1

In addition, studies show how smart transport system improves resource allocation, especially in the plan of routes and traffic signal control. Traditional traffic management strategies often follow the scheduled timetable that does not take into account changing traffic demands.

However, more efficient use of data-operated adaptation techniques makes it possible to allocate resources such as fuel consumption, traffic signal time and parking sites.

Smart Traffic Signal Coordination, which ensures that traffic lights at many junctions collaborate to provide smooth vehicle movement, is an important part of this improvement. Traffic lights can be adjusted to reduce passive time on junctions using intensive teaching algorithms analyzing traffic patterns in real time. This results in low travel time and low fuel use (Huang and Zhao, 2022).

The study additionally highlights whether the route adaptation technique can reduce unnecessary stops. Navigation systems can suggest the most effective routes by analyzing GPS data from thousands of cars. It reduces the total amount of traffic on major roads and distributes traffic burden between alternative routes. In high -density urban settings, transport networks are shown to reduce travel time by 20% (liu et al., 2023).

4. Effective Data Utilization

The ability to use the huge amounts of data from many sources to improve decision making is a significant advancement in smart transport systems. The study shows how to get better traffic predictions and more efficient urban dynamics scheme through real -time sensor data, GPS tracking and integration of historical traffic records.

Use of future stating analysis is one of the most common uses of data, where machine learning models analyze real -time input and historical traffic data to predict potential delays, accidents and crowds. Methods of preventive management, including demand traffic management, and signal pre-adjustment, have been made possible by this. According to research, the overall road efficiency of cities adopting forecast traffic analytics increased in 35% (Zoo et al., 2022).

In addition, public traffic data is important from automobiles and smartphone apps (example Waze and Google maps) to increase traffic predictions. By combining the data of thousands of users, these platforms are able to identify accidents, road closure and slow traffic areas in real time. Traffic control centers can reduce delays by including these data in the transport system and implementing immediate changes.

Additionally, authorities can carry out specific tasks such as temporary lane changes or suggestions for public transit using methods of AI-operated detection to identify unusual traffic patterns, such as roadwork or special events (Kim & Park, 2023) sudden increase in congestion. Research findings highlight that modern traffic control methods have the ability to change urban transport. Studies show how the city can get better resource usage, less traffic and smoother traffic flow by combining real-time data collecting, AI-driven predictive analytics, and automated congestion control measures.

5. Smart Parking

Intelligent transportation systems (ITS) are changing or transforming how to navigate our cities, and smart parking solutions are an important part of this development. These solutions aim at enhancing the parking space usage and offer real-time parking availability information and facilities such as mobile app-based payments. The benefits of smart parking are beyond the facility, as they can reduce congestion in commercial and city areas. By providing drivers with real -time information about available parking sites, smart parking systems help them avoid wasting time for parking, leading to a greater efficient use of valuable urban space. The market for smart parking solutions in Saudi Arabia (Alamoudi et al., 2024), especially with high number of vehicles in cities such as Jeddah, is ready for significant growth, which is inspired by the increasing adoption of ITS and continuous increase in vehicle ownership.

6. Efficient Public Transport Expansion to reduce emission and congestion

Developing a broad and efficient public transport system is important to reduce the dependence on private vehicles, reduce traffic congestion and reduce emissions of greenhouse gases. This requires a multimodal approach that includes buses, metro and trams which have played an important role

 doi:10.20944/preprints202603.1269.v1

in providing a comfortable and accessible travel option. It is necessary to expand the public transit network in a city, Jeddah with increasing population and increasing traffic volume (Alamoudi et al., 2024). By providing reliable and frequent services, the city can encourage more inhabitants to choose public transport as they are preferred to travel. This change towards public transport will not only reduce the congestion on the roads, but will also contribute to a cleaner and more sustainable urban environment. By investing in a strong and user -friendly public transit system, Jeddah can create a more efficient and living city for its people.

7. Integrated public transport and active mobility

Integrating public transport services with other methods of transport, such as bike-sharing programs, can further increase their effectiveness and promote more sustainable urban environment. There is a need to encourage the use of environmentally friendly methods of transport, such as cycling and walking, safe and accessible infrastructure. This involves applying measures to ensure the presence of bike lanes, pedestrians' zones and applying safety measures to ensure comfortable and safe experience for users. In Jeddah, with its warm climate, developing ways to shaded walking and cycling paths can make these options more viable and attractive to the people living in Jeddah (Alamoudi et al., 2024). Transit hubs, residential areas and major sites will improve the strategic placement access to these paths and encourage more people to choose active transport. By creating a network of interconnected transport options, Jeddah can create a more durable and efficient urban environment, reduce dependence on private vehicles and promote a healthy lifestyle for its inhabitants.

Limitations of the Approach

1. Data collection and integration challenge

Construction of options and system efficiency can be affected by various difficulties associated with data collection and integration. The existence of various sources of information, such as social media, GPS devices, Internet of Things sensors and data from public transport, is a significant concern. It can be quite difficult to combine these various inputs in real time. The sensor or communication network may result in incorrect or partial data as a result of wrong conclusion, so data quality and accuracy are also important. Scalability also becomes an issue when the transport network expands as the increasing amount of data can lead to inefficiencies.

Additionally, the noise in sensor data may have negative effects on the system performance as accuracy and the whole.

2. Real Time Processing Challenges

In transport systems, the processing layer has several limitations that can reduce its efficiency. There is a significant problem, which is latency, centralized computing like cloud computing can cause delays that are inappropriate for accident prevention and real -time traffic control.

Distributed computing architecture has a high processing burden and demands a lot of computational resources, even if they are useful for processing large -scale versions of transport data. In addition, while the edge computing reduces delay, it often lacks processing capacity of cloud-based systems, making it difficult to efficiently handle complex AI and Big Data Analytics activities. These restrictions suggest that to improve the performance of transport systems, processing architecture must take a balanced approach

3. Scalability Issues

The proposed solution may face scalability challenges, particularly when applied to large metropolitan areas with complex and dense traffic networks. Systems that perform adequately in smaller or less complex environments might struggle under the increased data volume and processing demands of a larger city. Scalability issues can lead to delays in data processing and decision-making, reducing the system's overall efficiency. To address this, adopting decentralized approaches, such as federated learning combined with meta-learning, can enhance scalability by distributing the computational load and enabling the system to adapt to various urban scales.

 doi:10.20944/preprints202603.1269.v1

4. Security Concerns

There are cyber security risks. The spread of IOT devices within transport systems introduces several entry points to the potential cyber-attack. Recent incidents have demonstrated weaknesses in connected vehicle systems, where security flaws have allowed unauthorized access to vehicle control and data (Greenberg, 2024). Such violations are underlined by the need for strong cyber security measures to protect both infrastructure and its users. The report (Oladimeji et al.,2023) also highlights that privacy and security continue to be major obstacles to the widespread adoption of connected automobile technologies. A fully automated system, such as an autonomous car, would be appealing to terrorist organizations, hackers, selfish people, and dissatisfied workers. In the worst situation, these vehicles might be used for terrorist activities without a driver.

The comprehensive data collection contained in smart transport systems also increases significant privacy concerns that have been released data privacy. Modern vehicles equipped with advanced sensors and connectivity facilities can collect detailed personal information including history and behavior patterns (Jackson, 2024). How this data is stored, used, and shared, a debate about it is going on, with special concern over potential misuse by manufacturers or third party

5. Infrastructure and Cost Limitations

Another limitation is high implementation cost. Establishing a fully functional smart transportation requires adequate investment in new technologies, including IOT sensors, advanced communication networks and computing infrastructure. The financial burden of deploying these techniques or technologies can be prohibitive, especially for limited budget areas.

Beyond the initial deployment, the ongoing maintenance and periodic upgrading of these systems is required to ensure their reliability and relevance. This continuous requirement for resources can stress the public budget and require careful planning and allocation. Also due to cost as stated in the article different communities and families aren't able to access different smart transportation technologies such as autonomous vehicles. It is stated (Oladimeji et al.,2023) that it will take some time before autonomous vehicles become a norm in middle-class families.

In addition, integrating new techniques with heritage transport systems can be challenging. Older infrastructure may not be compatible with modern smart technologies, requiring extensive and expensive amendments to adjust new systems.

6. Legal and Regulatory Challenges

The rapid development of smart transport technologies has led the surpass of universal standards and protocols. This lack of standardization can lead to issues of compatibility and an obstacle to spontaneous integration of various systems and equipment.

Data privacy laws and other regulatory structures can limit how transport data is collected, stored, and used. To navigate these rules, a delicate balance is required between taking advantage of data for system optimization and respecting personal privacy rights.

7. Algorithmic and AI Challenges

Since artificial intelligence and machine learning models are used to control traffic flow and predict the needs of transport, algorithm and AI are increasing in importance in transport systems. Biasness is an important problem in these models; If they are programmed on wrong or bias data, the results may be deformed, resulting in uneven services for various population.

Another further difficulty is the complexity of real -time analytics. Research and development is always focused on creating algorithms that can process large -scale versions of data generated by the smart transport system by providing yet excellent performance. To be fair and successful, these issues must be resolved.

8. Communication and Network Issues

The reliability of 4G and 5G networks is important for the development of cooperative intelligent transport systems (CITs), yet inconsistent availability of high-speed networks presents a significant challenge. For example in different areas, 5G rollouts have faced obstacles due to infrastructure boundaries and policy decisions, resulting in sub form connectivity in both urban and rural areas.

 doi:10.20944/preprints202603.1269.v1

This contradiction can obstruct the effectiveness of the CIT, which depends on the stable and rapid data exchange.

Additionally, urban environments often deal with high data traffic volumes, causing network congestion that causes delay in data transfer. These delays can adversely affect the processes of important real -time decision making processes for traffic management and autonomous vehicle operations. While the integration of 5G is aimed at addressing these issues by offering high bandwidth and reducing the delay, its success is accidental on the development of widespread and justified infrastructure (Gohar & Nencioni, 2021).

Recommendations for Improvement

1. Implementing deep learning techniques to enhance predictive accuracy:

Advanced machine learning models, especially deep learning techniques, can greatly increase future accuracy in various applications. Using a deep neural network, which is made up of many layers, these models can extract a complex pattern from large datasets. For example, the conversion neural networks (CNNs) are effective for analyzing spatial data, while recurrent neural networks (RNNs) excel in processing temporary data.

In terms of smart transportation systems, deep education can be implemented to predict traffic congestion, optimize the route plan and improve vehicle dispatch. A CNN can analyze traffic camera images to detect events in real time, while an RNN can use historical and current data to predict the traffic flows. This progress does not only improve efficiency, but it also contributes to safe and more effective transport.

2. Utilizing edge computing to process traffic data closer to the source:

Using edge computing to process traffic data close to the source provides significant benefits in terms of efficiency and accountability. Edge computing involves processing data near its origin, such as inside vehicles or in roadside units, rather than sending it to a centralized cloud server. This approach reduces delays, it increases the response time, and also reduces the amount of data sent. In practical applications, the edge devices can analyze different sensor data from vehicles and infrastructure in real time, enabling immediate tasks such as adjusting traffic signals, detection of dangers and facilitating vehicle-to vehicle (V2V) communication. These capabilities can improve the overall traffic management and increased security on the roads.

3. Intergrating data from multiple sources, such as weather conditions and social media, to improve decision-making:

Integrating or combining different data from many sources, such as weather conditions and social media data, can significantly increase the efficiency of decision making in transport systems. Data fusion techniques such as Kalman filtering and Bayesian networks, allow for uninterrupted combination of information from various sources, providing a more comprehensive approach to the transport environment.

For example, incorporating weather data enables the system to estimate adverse conditions, while social media may offer events and real -time updates when the road closure. This rich dataset improves decision making for more accurate predictions and traffic management and emergency response, eventually enhancing overall security and efficiency on the roads.

4. Employing advanced encryption and authentication mechanisms to protect data integrity:

To protect data integrity in smart transportation systems, it is very necessary to use or apply advanced encryption and certification techniques. Applying strong encryption algorithms ensures that the data is safely transmitted and stored, protecting sensitive information from unauthorized access and cyber-attack such as hacking. This level of encryption is important in maintaining the confidentiality and integrity of the data collected.

In addition to encryption, strong certification measures such as multi-factor authentication (MFA) and biometric verification play an important role in preventing unauthorized access to the

 doi:10.20944/preprints202603.1269.v1

system. Regular safety audit and updates can further increase data security by identifying the weaknesses and can ensure that the latest security protocols are in place.

These measures are particularly important for the protection of data collected from smart transport systems, including personal information, communication between vehicle telemetry and equipment. By integrating these safety protocols, organizations can significantly reduce the risk of data violations and ensure safety of the information managed by them.

5. Increasing sensor deployment and utilizing crowdsourced data for a more comprehensive dataset:

Increase in deployment of sensors and using crowdsourced data can significantly increase the understanding of transport dataset.

Extending the network of sensors such as cameras, lidar and GPS across the transport infrastructure ensures better coverage and more detailed data collection. These sensors are capable of monitoring traffic flows, detection of events and collecting valuable environmental data, which is necessary for effective traffic management.

Additionally, taking advantage and using crowdsourced data from smartphones and connected vehicles provides further knowledge and insights. For example, applications that allow users to report traffic conditions or events can complement the data collected from traditional sensors. This real -time information adds another layer of expansion to the dataset.

In a rich dataset, by combining sensor data with crowdsourced information which enables more accurate traffic modeling. This integration supports better decisions for traffic management, route adaptation and emergency response, ultimately leading to security and improving efficiency on the roads.

6. Zoning law for sustainable transportation:

It is important to adopt zoning rules that encourage mixed-use development and transit-oriented development (TOD) around public transportation hubs. These rules should prioritize the need to reduce long -distance travel requirement and promote the use of permanent transport.

Additionally, zoning laws should integrate green space, parks and community gardens in urban planning to increase environmental stability and have pleasant walking and cycling (Alamoudi et al., 2024).

7. Public-private participation or participation for an enhanced transportation system**:**

It is necessary to accelerate the development and implementation of permanent transport projects through public-private partnership. Private institutions can bring innovation and financial assistance to carry forward the transport goals of the city. Cooperation with government agencies and private sector stakeholders is important to develop a holistic transition for permanent transport, ensure a coordinated approach and take advantage of resources and expertise from different regions.

Conclusions

Smart transportation systems generate a very large amount of data from sensors, vehicles, cameras, and mobile devices. Managing and analyzing this data is important for improving traffic flow, reducing congestion, and increasing road safety. Traditional systems often struggle to handle such large and continuous data streams.

This paper discussed a distributed big data architecture for smart transportation using technologies such as Hadoop, Apache Kafka, and Apache Flink. Hadoop provides scalable storage and supports large-scale data processing, while Kafka enables reliable real-time data streaming from multiple sources. Apache Flink allows real-time analysis of transportation data and supports quick responses to traffic events and incidents.

The study also showed how intelligent transportation systems can benefit from data analytics, machine learning, and edge computing. These technologies help detect traffic incidents faster, improve congestion management, and support better decision-making for transport authorities.

 doi:10.20944/preprints202603.1269.v1

References

1. Ruseruka, C., Mwakalonge, J., Comert, G., Siuhi, S., & Perkins, J. (2023). Road condition monitoring using vehicle built-in cameras and gps sensors: a deep learning approach. *Vehicles*, *5*(3), 931-948.
2. Misra, A., Gooze, A., Watkins, K., Asad, M., & Le Dantec, C. A. (2014). Crowdsourcing and its application to transportation data collection and management. *Transportation Research Record*, *2414*(1), 1-8.
3. Kim, S., Lewis, M. E., & White, C. C. (2005). Optimal vehicle routing with real-time traffic information. *IEEE Transactions on Intelligent Transportation Systems*, *6*(2), 178-188.
4. Yokoya, Y. (2004). Dynamics of traffic flow with real-time traffic information. *Physical review E*, *69*(1), 016121.
5. Bora, P. S., Sharma, S., Batra, I., Malik, A., & Ashfaq, F. (2024, July). Identification and classification of rare medicinal plants. In *2024 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC)* (pp. 1-6). IEEE.
6. Sujatha, R., Aarthy, S. L., & NZ, J. (2021). A machine learning way to classify autism spectrum disorder. International Journal of Emerging Technologies in Learning (iJET), 16(6), 182-200.
7. Cohn, N. (2009). Real-time traffic information and navigation: An operational system. *Transportation research record*, *2129*(1), 129-135.
8. Faizal, A., & Aisyah, N. (2023). Innovative Approaches to Enterprise Database Performance: Leveraging Advanced Optimization Techniques for Scalability, Reliability, and High Efficiency in Large-Scale Systems. *Reliability, and High Efficiency in Large-Scale Systems*.
9. Boppiniti, S. T. (2020). Big data meets machine learning: Strategies for efficient data processing and analysis in large datasets. *International Journal of Creative Research In Computer Technology and Design*, *2*(2).
10. Ray, S. K., Pawlikowski, K., & Sirisena, H. (2008, October). A fast MAC-layer handover for an IEEE 802.16 e-based WMAN. In International Conference on Access Networks (pp. 102-117). Berlin, Heidelberg: Springer Berlin Heidelberg.
11. Lee, S., Abdullah, A., & Jhanjhi, N. Z. (2020). A review on honeypot-based botnet detection models for smart factory. International Journal of Advanced Computer Science and Applications, 11(6).
12. Li, Q., Zhou, W., & Zheng, X. (2024). Distributed learning in intelligent transportation systems: A survey. *Information*, *15*(9), 550.
13. Wei, W., Danman, W. U., Qiuwei, W. U., Shafie-Khah, M., & Catalão, J. P. (2019). Interdependence between transportation system and power distribution system: A comprehensive review on models and applications. *Journal of Modern Power Systems and Clean Energy*, *7*(3), 433-448.
14. Samaras, V., Daskapan, S., Ahmad, R., & Ray, S. K. (2014, November). An enterprise security architecture for accessing SaaS cloud services with BYOD. In 2014 Australasian Telecommunication Networks and Applications Conference (ATNAC) (pp. 129-134). IEEE.
15. Hannan, S. A. (2016). An overview on big data and hadoop. *International Journal of Computer Applications*, *154*(10).
16. Vora, M. N. (2011, December). Hadoop-HBase for large-scale data. In *Proceedings of 2011 International Conference on Computer Science and Network Technology* (Vol. 1, pp. 601-605). IEEE.
17. Thusoo, A., Sarma, J. S., Jain, N., Shao, Z., Chakka, P., Zhang, N., ... & Murthy, R. (2010, March). Hive-a petabyte scale data warehouse using hadoop. In *2010 IEEE 26th international conference on data engineering (ICDE 2010)* (pp. 996-1005). IEEE.
18. Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107-113.
19. Maitrey, S., & Jha, C. K. (2015). MapReduce: simplified data analysis of big data. *Procedia Computer Science*, *57*, 563-571.
20. Dey, K., Ray, S., Bhattacharyya, P. K., Gangopadhyay, A., Bhasin, K. K., & Verma, R. D. (1985). Salicyladehyde 4-methoxybenzoylhydrazone and diacetylbis (4-methoxybenzoylhydrazone) as ligands for tin, lead and zirconium. J. Indian Chem. Soc.;(India), 62(11).
21. Jin, H., Ibrahim, S., Qi, L., Cao, H., Wu, S., & Shi, X. (2011). The mapreduce programming model and implementations. *Cloud computing: principles and paradigms*, 373-390.

 doi:10.20944/preprints202603.1269.v1

22. Misra, A., Gooze, A., Watkins, K., Asad, M., & Le Dantec, C. A. (2014). Crowdsourcing and its application to transportation data collection and management. *Transportation Research Record*, *2414*(1), 1-8.
23. Anwesa Chaudhuri, A. C., & Sanjib Ray, S. R. (2015). Antiproliferative activity of phytochemicals present in aerial parts aqueous extract of Ampelocissus latifolia (Roxb.) Planch. on apical meristem cells.
24. Roorda, M. J., Shalaby, A., & Saneinejad, S. (2011). Comprehensive transportation data collection: case study in the Greater Golden Horseshoe, Canada. *Journal of urban planning and development*, *137*(2), 193-203.
25. Torre-Bastida, A. I., Del Ser, J., Laña, I., Ilardia, M., Bilbao, M. N., & Campos-Cordobés, S. (2018). Big Data for transportation and mobility: recent advances, trends and challenges. *IET Intelligent Transport Systems*, *12*(8), 742-755.
26. Cheng, N., Lyu, F., Chen, J., Xu, W., Zhou, H., Zhang, S., & Shen, X. (2018). Big data driven vehicular networks. *Ieee Network*, *32*(6), 160-167.
27. Muzafar, S., & Jhanjhi, N. Z. (2020). Success stories of ICT implementation in Saudi Arabia. In Employing Recent Technologies for Improved Digital Governance (pp. 151-163). IGI Global Scientific Publishing.
28. Xu, W., Zhou, H., Cheng, N., Lyu, F., Shi, W., Chen, J., & Shen, X. (2017). Internet of vehicles in big data era. *IEEE/CAA Journal of Automatica Sinica*, *5*(1), 19-35.
29. Ray, S. K., Sirisena, H., & Deka, D. (2013, October). LTE-Advanced handover: An orientation matching-based fast and reliable approach. In 38th annual IEEE conference on local computer networks (pp. 280-283). IEEE.
30. Muzammal, S. M., Murugesan, R. K., Jhanjhi, N. Z., & Jung, L. T. (2020, October). SMTrust: Proposing trust-based secure routing protocol for RPL attacks for IoT applications. In 2020 International Conference on Computational Intelligence (ICCI) (pp. 305-310). IEEE.
31. Xu, W., Zhou, H., Cheng, N., Lyu, F., Shi, W., Chen, J., & Shen, X. (2017). Internet of vehicles in big data era. *IEEE/CAA Journal of Automatica Sinica*, *5*(1), 19-35. Gerla, M., & Kleinrock, L. (2011). Vehicular networks and the future of the mobile internet. *Computer Networks*, *55*(2), 457-469.
32. Ahmad, A., Paul, A., & Rathore, M. M. (2016). An efficient divide-and-conquer approach for big data analytics in machine-to-machine communication. *Neurocomputing*, *174*, 439-453.
33. Hauck, M., Machhamer, R., Czenkusch, L., Gollmer, K. U., & Dartmann, G. (2019). Node and block-based development tools for distributed systems with AI applications. *IEEE Access*, *7*, 143109-143119.
34. Jabeen, T., Jabeen, I., Ashraf, H., Ullah, A., Jhanjhi, N. Z., Ghoniem, R. M., & Ray, S. K. (2023). Smart wireless sensor technology for healthcare monitoring system using cognitive radio networks. Sensors, 23(13), 6104.
35. March, S. T., & Rho, S. (1995). Allocating data and operations to nodes in distributed database design. *IEEE Transactions on Knowledge and data engineering*, *7*(2), 305-317.
36. Azzedin, F. (2013, May). Towards a scalable HDFS architecture. In *2013 international conference on collaboration technologies and systems (CTS)* (pp. 155-161). IEEE.
37. Ismail, M., Niazi, S., Ronström, M., Haridi, S., & Dowling, J. (2017, May). Scaling HDFS to more than 1 million operations per second with HopsFS. In *2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)* (pp. 683-688). IEEE.
38. Yang, X., Yin, Y., Jin, H., & Sun, X. H. (2014, September). SCALER: Scalable parallel file write in HDFS. In *2014 IEEE International Conference on Cluster Computing (CLUSTER)* (pp. 203-211). IEEE.
39. Shah, I. A., Jhanjhi, N. Z., Amsaad, F., & Razaque, A. (2022). The role of cutting-edge technologies in Industry 4.0. In Cyber Security Applications for Industry 4.0 (pp. 97-109). Chapman and Hall/CRC.
40. Kaur, N., Verma, S., Jhanjhi, N. Z., Singh, S., Ghoniem, R. M., & Ray, S. K. (2023). Enhanced QoS-aware routing protocol for delay sensitive data in Wireless Body Area Networks. IEEE Access, 11, 106000-106012.
41. Dhulavvagol, P. M., & Totad, S. G. (2023). Performance enhancement of distributed system using HDFS federation and sharding. *Procedia Computer Science*, *218*, 2830-2841.
42. Javed, D., Jhanjhi, N. Z., Ashfaq, F., Khan, N. A., Das, S. R., & Singh, S. (2024, July). Student performance analysis to identify the students at risk of failure. In *2024 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC)* (pp. 1-6). IEEE.
43. Ruseruka, C., Mwakalonge, J., Comert, G., Siuhi, S., & Perkins, J. (2023). Road condition monitoring using vehicle built-in cameras and gps sensors: a deep learning approach. *Vehicles*, *5*(3), 931-948.

 doi:10.20944/preprints202603.1269.v1

44. Arebey, M., Hannan, M. A., Basri, H., Begum, R. A., & Abdullah, H. (2010, June). Solid waste monitoring system integration based on RFID, GPS and camera. In *2010 international conference on intelligent and advanced systems* (pp. 1-5). IEEE.
45. Patire, A. D., Wright, M., Prodhomme, B., & Bayen, A. M. (2015). How much GPS data do we need?. *Transportation Research Part C: Emerging Technologies*, *58*, 325-342. Andrews, L. J. B., Sarathkumar, D., & Raj, R. A. (2023, February). IOT Based Surveillance Camera with GPS Module. In *2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)* (pp. 1-3). IEEE.
46. Brohi, S. N., Jhanjhi, N. Z., Brohi, N. N., & Brohi, M. N. (2023). Key applications of state-of-the-art technologies to mitigate and eliminate COVID-19. Authorea Preprints.
47. Maitrey, S., & Jha, C. K. (2015). MapReduce: simplified data analysis of big data. *Procedia Computer Science*, *57*, 563-571.
48. Dittrich, J., & Quiané-Ruiz, J. A. (2012). Efficient big data processing in Hadoop MapReduce. *Proceedings of the VLDB Endowment*, *5*(12), 2014-2015.
49. Lee, K. H., Lee, Y. J., Choi, H., Chung, Y. D., & Moon, B. (2012). Parallel data processing with MapReduce: a survey. *AcM sIGMoD record*, *40*(4), 11-20.
50. Lin, X., Meng, Z., Xu, C., & Wang, M. (2012, September). A practical performance model for hadoop mapreduce. In *2012 IEEE International Conference on Cluster Computing Workshops* (pp. 231-239). IEEE.
51. Jain, N. K., Saini, R. K., & Mittal, P. (2018). A review on traffic monitoring system techniques. *Soft computing: Theories and applications: Proceedings of SoCTA 2017*, 569-577.
52. Biersack, E., Callegari, C., & Matijasevic, M. (2013). *Data traffic monitoring and analysis*. Heidelberg, Germany: Springer Berlin Heidelberg.
53. Nadeem, T., Dashtinezhad, S., Liao, C., & Iftode, L. (2004, January). Trafficview: A scalable traffic monitoring system. In *IEEE International Conference on Mobile Data Management, 2004. Proceedings. 2004* (pp. 13-26). IEEE.
54. Biswas, S. P., Roy, P., Patra, N., Mukherjee, A., & Dey, N. (2015, September). Intelligent traffic monitoring system. In *Proceedings of the Second International Conference on Computer and Communication Technologies: IC3T 2015, Volume 2* (pp. 535-545). New Delhi: Springer India.
55. Erbacher, R. F. (2001, July). Visual traffic monitoring and evaluation. In *Internet Performance and Control of Network Systems II* (Vol. 4523, pp. 153-160). SPIE.
56. Alamoudi, M., Imam, A., Majrashi, A., Osra, O., & Hegazy, I. (2024). Integrating intelligent and sustainable transportation systems in Jeddah: a multidimensional approach for urban mobility enhancement. *International Journal of Low-Carbon Technologies*, *19*, 1301-1314.