International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 04 April-2026 | Impact Factor: 3.5

Big Data for Smart City Traffic Management

Nitin Kumar
SCSE, *Galgotias University,*
Greater Noida,India
nitinthakur8950@gmail.com

Bhavya verma
SCSE, *Galgotias University,*
Greater Noida,India
vermabhavya911@gmail.com

Shrddha sagar
SCSE, Galgotias University,
Greater Noida, India

How to Cite this Article:
sagar, S., verma, B. & Kumar, N. (2026). Big Data for Smart City Traffic Management. International Journal of Creative and Open Research in Engineering and Management, 02(04). https://doi.org/10.55041/ijcope.v2i4.667

https://doi.org/ 10.55041/ijcope.v2i4.667

Abstract - Urban areas around the world are experiencing rapid urbanisation and steady increases in their population which have lead to large increases in the number of vehicles on the roads due to the increasing demand for services because of urbanisation. This increase in vehicle traffic has created a number of challenges, such as increased congestion on the roads, increased fuel consumption, increased carbon emissions into the atmosphere, increased amount of time travelling to a destination and an increasing concern regarding the safety of our roadways and the high number of accidents occurring as a result of vehicle traffic. Currently, the traditional traffic management approaches (using fixed traffic signals, manual monitoring of traffic conditions, and separate traffic control methods) used to manage vehicle traffic in cities are unable to keep pace with the rapid changes taking place within modern urban traffic systems. Traffic management approaches such as those currently being used are unable to respond to the ever-changing needs of urban traffic management systems in real time, nor do they function well in the event of unplanned occurrences such as congestion caused by rush hour traffic, accidents or road blockages and/or weather conditions.

In order to address these issues, this document presents a comprehensive Smart City Traffic Management System powered by Big Data.
Through the use of cutting-edge technology, this system leverages data for smarter traffic management by integrating numerous data sources including: an Internet of Things (IoT) based traffic sensor network, video surveillance systems, globally positioned satellite (GPS) vehicle systems, and mobile applications for capturing real-time information as well as historical information traffic data. Additionally, this system uses technologies such as the Apache Hadoop and Apache Spark Framework for both storing and processing vast amounts of data very quickly and performing analytical functions on the data in real-time (as it is received through many different channels).

Keywords— Smart City, Big Data Analytics, Traffic Management, IoT, Machine Learning, Hadoop, Spark

1. Introduction

The rapid growth in urban populations and increased automobile use, coupled with the upsurge of economic activity throughout the world, has placed substantial demand on current transportation systems in the developing world. Previously designed to meet only moderate traffic flows, these road systems must support increasing volumes of traffic with far more complicated patterns of movement. Traditional traffic management techniques consist predominantly of signal timing, manual supervision and separate controls for each location.

In contrast, these methods lack the ability and sophisticated adaptability necessary for responding adequately to changing dynamic events such as traffic congestion resulting from vehicular accidents or road closures due to adverse weather conditions, the announcement of rush hour traffic and other forms of significant public activity in urban communities. As a result, Urban Regions continue to suffer from significant negative impacts including excessive traffic congestion, long travel times, high levels of fuel usage, greenhouse gas emissions, and a reduction in the overall level of Road Safety. Nevertheless, Smart City Concepts are emerging as a new paradigm for addressing the myriad challenges facing cities and urban planners today. The primary objective of a Smart City Concept is to improve the overall effectiveness, efficiency, sustainability, and livability of Urban Areas by taking advantage of New Digital Technology. Specifically, Smart City Technologies will include the technologies of the Internet of Things (IoT), Cloud Computing and Big Data Analytics. The Technologies that make up the Smart City Solution Stack will include a variety of IoT devices. These devices will include Road Spotters, Video Surveillance Cameras (CCTV) and GPS enabled Vehicles, which will generate literally billions upon billions of a large quantity of Real-Time Traffic Data, from every imaginable type of roadway, in every conceivable location throughout the world on a continual basis. Due to the volume, pace of production, and diversity of such information it will involve high-technology data storage, data processing and data analysis methods in order to correctly process, store, and eventually,

have meaningful analytical knowledge of the Traffic Data that will subsequently emerge as an outcome of the Smart Cities Technologies.

The suggested smart city traffic management system that is created on the basis of the big data analysis would allow monitoring the traffic in real-time, predictive analysis and adaptive control. It is proposed that a scalable big data architecture be created using a wide variety of data sources including: IoT sensors, traffic cameras, GPS devices, and mobile applications, to allow for the most accurate and useful data analysis. The method by which the transportation operational processes will be optimized on the fly using distributed processing frame work and machine learning algorithms, to allow for intelligent comprehensive analysis. This study provides evidence that utilizing big data analysis in the transportation industry can greatly improve the effectiveness of transportation systems in cities, and therefore increase urban transportation efficiency, fuel efficiency, and enhance/assist in the development of more sustainable and resilient smart city transportation systems.

2. Literature Review

Several studies have demonstrated the potential for Big Data and Internet of Things (IoT) technologies to tackle urban mobility issues, particularly in the context of smart city traffic management. The first study focused on applying IoT-based sensing systems like inductive loop detectors, RFID sensors, traffic surveillance cameras, and GPS-equipped vehicles to obtain real-time traffic data. Gubbi et al believe that IoT is a basic technological innovation in smart city applications which depends on on-going sensing, communication and surveillance of urban systems. The definition of Smart cities by Batty et al. (2015) refers to a complicated, data-driven ecosystem, which need ongoing real-time data-processing and analytics to aid sound operational and strategic decision-making procedures.

As the amount and rate of traffic data continues to increase, more studies have been undertaken on scalable Big Data processing models with the ability to deal with large-scale urban data sets. Scalability, fault-tolerance, and efficiency of operations have made the Apache Hadoop and Apache Spark types of distributed computing platforms the most popular among traffic data analytics. Application to Hadoop MapReduce paradigm of long-term traffic trend analysis and batch processing have been witnessed whereas Apache Spark uses in-memory computing to assist with low-latency real-time traffic analytics. These structures allow managing various data streams brought about by sensors, cameras, social media feed and mobile apps, which are applicable when it comes to smart city traffic management systems.

Combining machine learning with Big Data frameworks have played a crucial role in improving the performance of traffic prediction and congestion detecting methods. Some of the models that have been utilized in the traffic state classification and incident detection through these include Support Vector Machines (SVM) and Random Forest classifiers because of their strength and interpretation ability. Through the building of artificial neural networks (ANN), we have been able to achieve nonlinear models of complex behavior that will evolve over time. Long short-term memory (LSTM) techniques have also been gaining increased popularity for modelling the time based nature of congestion forecasting, because they are able to catch temporal dependencies between the different components of sequentially measured traffic data. Predictive models enable municipal governments, and all the agencies involved, to make vital and timely decisions regarding the movement of vehicles through their city.

Smart traffic management solutions are still facing challenges in terms of data scalability, interoperability with diverse data sources, data privacy and security, and real-time system responsiveness, which remain significant but not fully resolved. Additionally, several approaches now focus on separate elements like prediction or signal management, instead of offering a complete end-to–end system.

Based on previous studies, this research proposes a unified system for urban traffic control, based on big data technology, that combines large-volume data processing with machine-learning based traffic forecasting and real-time optimization of adaptive traffic signals. The intended benefit of this system for smart city environments is that it provides scalability and real time responsiveness, so that we can improve urban mobility, reduce congestion, and promote sustainable transportation.

3. System Architecture and Methodology

3.1 System Architecture

A Big Data architecture that has multiple layers and is modular/scalable for flexibility has been used to build a new system for collecting real-time data from IoT devices, GPS, traffic cameras, mobile applications, etc.

The Data Collection Layer gathers the collected Real-Time Data from the different sources mentioned above.

The Data Ingestion Layer will use Apache Kafka and Apache Flume to stream this real-time data at a very high velocity.

The Storage Layer manages scalable amounts of collected Real-Time Data in two different formats, HDFS and Cloud Store.

The Process Layer primarily uses Apache Spark and Hadoop MapReduce to perform real-time analytics using Streaming Data and Batch Analytics on collected data that can be processed using the process layer.

The use of the Analytics Layer is based on the usage of the Machine Learning Models to forecast the future state of traffic as well as congestion detection in the cities.

3.2 Methodology

The systematic approach for implementing a Smart City Traffic Management System allows for the collection and analysis of data, through smart data collection methods and intelligent processes, to provide adaptive Traffic Management. The methodology incorporates a diverse range and quantity of large scale traffic data to produce a responsive and scalable methodology in real-time. The general workflow consists of multiple step processes including data input(s), data output(s), pre-processing of the data, analytics on the data, and data forecasting.

3.2.1 Data Acquisition

The Urban Analysis methodology for traffic monitoring starts with the continued collection of data from various locations

situated in the urban environment. Data from traffic sensors at intersections, installed on IoT platforms will allow for real-time collection of vehicle numbers, speed, lane occupancy and density. Queue lengths and conditions of traffic flow will be captured using data obtained from installed traffic surveillance cameras. Vehicles that are equipped with GPS will provide Spatiotemporal trajectory data. Mobile applications and crowdsourcing platforms will provide additional traffic updates and create enhanced situational awareness for Traffic Analysis by integrating multiple data sources.

3.2.2 Data Ingestion and Streaming

Traffic information is collected and subsequently sent to a high-throughput processing pipeline that connects to the central processing unit. Apache Kafka provides the capability for immediate streaming of high-volume real-time data streams. In addition, reliable collection of data from sensor networks and various logging sources is done via Apache Flume.

3.2.3 Data Preprocessing

Data preprocessing ensures consistent and high-quality data from incoming traffic data. Raw data cannot consist of sensors' direct collection, nor from an external source containing noise, missing values, or inconsistency due to sensor failure or delays in communication. In order to address these issues, various data pre-processing methods are implemented such as noise filtering, normalizing data, imputing missing values, and synchronizing timestamps to achieve clean, standardized data for use with a variety of machine learning models. In addition to the above methods described, the identification of features from data is also accomplished using different ways to derive meaningful features such as average vehicle speed, traffic density, congestion indices etc., that serve as inputs into the machine learning models.

3.2.4 Big Data Processing

Distributed frameworks for large-scale data management are intended to support both real-time analytic (or on-line analysis) and batch analytic (or offline analysis) applications using clean, parsed datasets for both types of analysis. Apache Spark enables the ability to generate real-time analytics on live streaming data streams using its in-memory processing capability enabling immediate insights from the most recent streaming data. Conversely, Hadoop MapReduce utilizes a batch mode to generate Reports from historical traffic data (such as Monthly and Daily Reports) to highlight long-term traffic trends and patterns. By combining the immediate benefit of real-time insights, with the analysis of historical traffic trends and patterns, organisations can make better quality, data-driven decisions.

3.2.5 Machine Learning–Based Prediction

Traffic pattern analysis and prediction of future traffic congestion are performed with the use of machine learning models. Time-series traffic predictions utilize Long Short-Term Memory (LSTM) networks because of their ability to identify temporal dependencies and understand the overall trend of traffic over long periods of time. Random Forest Classifiers are also utilized to classify traffic congestion levels. Historical data of traffic patterns is used to train the models, which are then validated against unseen test data to validate the generalization and dependability of the models. Performance metrics include accuracy, root mean square error (RMSE), precision, and recall. All of these metrics are used to assess the performance of the model.

3.2.6 Adaptive Traffic Signal Control

Using prediction data from Machine Learning (ML) techniques, Adaptive Traffic Signal Control method manages vehicle movement through signals using information specific to current roadway conditions and anticipated congestion levels. Adaptive Traffic Signal Control allows the City of London to continuously monitor the state of roadways and alter green light durations and minimize delays to commuters while improving overall traffic flow. Adaptive Traffic Signal Control successfully manages traffic congestion beforehand rather than after congestion has occurred.

4. Implementation and Results

The Smart City Traffic Management System proposed will use distributed big data technology combined with machine-learning platforms to allow for scalable, reliable, real-time traffic data processing. The backbone of the data processing infrastructure will be based on the use of Apache Hadoop for distributed file storage and batch analysis, along with Apache Spark for streaming data in real-time and having a low-latency for analysis of that data. Apache Kafka will be utilized as a messaging broker to facilitate the robust collection of fast-paced traffic source data. Python-based libraries, including TensorFlow, Scikit-learn, and NumPy, will be leveraged to create machine-learning models.

Traffic data will emerge from domestic and international sources that use diverse technologies, including simulated IoT-based sensors configured for collecting traffic, GPS trajectory data that show each movement of vehicles across areas of the country and feeds from traffic cameras that help determine traffic density and queue lengths. The various sources of traffic data will generate streaming data that includes a mix of structured and unstructured formats. All of the streaming traffic data will be directed to HDFS for analysis by using Kafka topics and will contain inherent variations of the formats in which the various data-types were collected.Before using the traffic data to create machine-learning models, several techniques will be applied for data preprocessing purposes to enhance the integrity and reliability of the traffic data for increasing model accuracy. Examples of data preprocessing techniques will include noise filtering, missing value handling, normalization, and feature extraction.

4.1 Machine Learning Analysis

To predict congestion levels on a time series basis utilizing historical and Real-Time traffic data, various machine-learning models have been used to analyze traffic Patterns. LSTM networks have been deployed as part of a time-series based approach to forecast traffic due to their ability to capture temporal dependencies on time-series data and Long-Term Transport Continuous Flow. An LSTM model forecasts short term traffic flows and levels of congestion at intersections and optimizes by being proactive in optimizing traffic signals for time-sensitive events.

Additionally, Random Forest Classifiers were used for the purpose of classifying traffic types of different congestion levels. Random Forest models provide a robust computational approach to identify normal traffic patterns despite large amounts of noisy data. Moreover, Random Forest can manage very high-dimensional datasets. The Traffic data set was split

into two data sets for training (70%) and testing (30%) so that testing data set remains completely unbiased when evaluating model performance. Evaluation metrics for the model include standard benchmark metrics of accuracy, Root Mean Square Error (RMSE), Precision, Recall and F1-Score. Evaluation results indicate that LSTM models consistently predict high accuracy rates for short-term traffic forecasts, and Random Forest classifiers detect congestion levels that are generally reliable over varying traffic flows.

4.2 Performance Evaluation

To evaluate the proposed system, a comparison was made with a traditional fixed-time traffic management system under similar traffic conditions. The evaluation used the following key performance indicators (KPIs): average time spent waiting for vehicles at intersections, traffic flow efficiency, fuel consumption, carbon dioxide (CO_2) emissions. The results of the evaluation show that the proposed system is effective.

Metric	**Traditional System**	**Proposed System**
Average Time	High	Significantly Reduced
Traffic Efficiency	Moderate	High
Fuel Consumption	High	Reduced
CO_2 Emissions	High	Reduced

Real-time traffic density and predicted congestion planning determine the length of green lights through Adaptive Signal Control Technologies (ASCTs). By improving upon outdated and inefficient forms of traffic management, the use of intelligent transportation systems has improved the performance of traffic operations at intersections, allowing for the flow of traffic to be uninterrupted, thereby reducing the need for "stop" and "start" movements; reducing unnecessary delay, and providing for improved traveller efficiencies. In addition, fewer hours of idling have reduced fuel consumption and carbon emissions, thereby contributing to smart cities' environmental sustainability goals through the integration of predictive analytics powered by big data and machine learning with traditional traffic management systems. Based on the results of our testing, we believe that the integration of both of these technologies will provide urban cities with an opportunity to improve their efficiency, reduce congestion, and enhance the overall quality of the transportation system in urban environments.

5.RESULT & FUTURE SCOPES

In this paper, we will discuss an Intelligent Transportation System utilizing Big Data, Smart Cities and the Internet of Things (IoT) to tackle the major problems of Urban Transportation. This new Smart Traffic System uses multiple data sources which enables both real time monitoring and statistical analyses of traffic conditions on a large scale. The processing of large volumes of data will be accomplished through scalable big data technologies in order to make accurate traffic predictions and model traffic patterns, using Machine Learning.

Experiments conducted by Researchers have established that this system has proven to be superior to traditional methods of managing traffic with fixed timetables. The combination of Predictive Analysis and Adaptive Signal Control would ultimately result in reduced wait times for vehicles, improved vehicles' ability to travel through an intersection with less fuel used and thus less CO2 emitted into the atmosphere. The use of Intelligent Transportation Systems that are data driven will contribute positively to urban mobility, create an increased level of safety for drivers and passengers, and support sustainable development goals. In addition, because this system is sufficiently reasonable, it can be implemented in multiple environments implementing different traffic solutions. Therefore, the system is especially suitable in today's Building Smart Cities and similar environments.

Although the initial findings indicate promise, many additional avenues exist for enhancement and implementation in an actual environment. In addition to this initial phase, future research will focus on the integration of the proposed system with existing urban traffic systems via the addition of live urban traffic infrastructure (e.g., real-time traffic signals) and centralized urban Traffic Control Centers to measure and validate the system's performance under actual conditions and environments. By applying advanced sensing technologies (e.g., LiDAR), developing systems for connected vehicle technology, and introducing high-quality camera technologies, researchers can provide additional methods of accurately assessing traffic conditions. The integration of advanced deep learning techniques (e.g., Convolutional Neural Networks [CNNs]) to detect traffic flow based on traffic images as well as advanced hybrid deep learning architectures can provide an additional mechanism for predicting traffic congestion and forecasting traffic flow.
More research will be needed to determine how to apply edge computing to minimize latency and provide quicker, local decision-making at intersections, so the system has the ability to respond more quickly. In addition, extending the architecture to include a multi-city and large-scale deployment capability will allow for coordinated traffic management in multi-city urban areas. Additionally, integration with autonomous and connected vehicle networks is an excellent opportunity to allow for cooperative traffic control and next-generation intelligent transportation systems. Addressing future research will allow this system to help facilitate the development of effective, resilient, and sustainable solutions for urban transportation in smart cities.

7.REFERENCES

[1] M. Batty, K. Axhausen, F. Giannotti, A. Pozdnoukhov, A. Bazzani, M. Wachowicz, G. Ouzounis, and Y. Portugali, "Smart cities of the future," *European Physical Journal Special Topics*, vol. 214, no. 1, pp. 481–518, 2012.

[2] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1645–1660, 2013.

[3] R. Kitchin, "The real-time city? Big data and smart urbanism," *GeoJournal*, vol. 79, no. 1, pp. 1–14, 2014.

[4] Apache Software Foundation, "Apache Spark: Lightning-fast unified analytics engine," 2024. [Online]. Available: https://spark.apache.org

[5] Apache Hadoop Project, "HDFS architecture guide," 2024. [Online]. Available: https://hadoop.apache.org