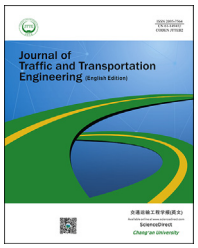
Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.keaipublishing.com/jtte



Review Article

An overview of Hadoop applications in transportation big data

Changxi Ma [a,b], *Mingxi Zhao* [a,*], *Yongpeng Zhao* [c]

[a] *School of Traffic and Transportation, Lanzhou Jiaotong University, Lanzhou 730070, China*
[b] *Key Laboratory of Railway Industry on Plateau Railway Transportation Intelligent Management and Control, Lanzhou Jiaotong University, Lanzhou 730070, China*
[c] *Gansu Highway Traffic Construction Group Co., Ltd., Lanzhou 730000, China*

HIGHLIGHTS

- The results related to the application of Hadoop to transportation big data since 2012 are summarized.
- The 8 major application scenarios of Hadoop in transportation big data are summarized and refined.
- The results of Hadoop computational model optimization and Hadoop combined with Spark are summarized.
- Existing issues and future work on the development of Hadoop and transportation big data integration are identified.

ARTICLE INFO

Article history:
Received 7 January 2023
Received in revised form
3 May 2023
Accepted 5 May 2023
Available online 30 September 2023

Keywords:
Information technology
Transportation big data
Hadoop
Intelligent transportation
Cloud computing

ABSTRACT

As an open-source cloud computing platform, Hadoop is extensively employed in a variety of sectors because of its high dependability, high scalability, and considerable benefits in processing and analyzing massive amounts of data. Consequently, to derive valuable insights from transportation big data, it is essential to leverage the Hadoop big data platform for analysis and mining. To summarize the latest research progress on the application of Hadoop to transportation big data, we conducted a comprehensive review of 98 relevant articles published from 2012 to the present. Firstly, a bibliometric analysis was performed using VOSviewer software to identify the evolution trend of keywords. Secondly, we introduced the core components of Hadoop. Subsequently, we systematically reviewed the 98 articles, identified the latest research progress, and classified the main application scenarios of Hadoop and its optimization framework. Based on our analysis, we identified the research gaps and future work in this area. Our review of the available research highlights that Hadoop has played a significant role in transportation big data research over the past decade. Specifically, the focus has been on transportation infrastructure monitoring, taxi operation management, travel feature analysis, traffic flow prediction, transportation big data analysis platform, traffic event monitoring and status discrimination, license plate recognition, and the shortest path. Additionally, the optimization framework of Hadoop has been studied in two main areas: the optimization of the computational model of Hadoop and the optimization of Hadoop combined with Spark. Several research results have been achieved in the field of transportation big data. However, there is less systematic research on the core technology of Hadoop, and the breadth

* *Corresponding author.*
E-mail addresses: machangxi@mail.lzjtu.cn (C. Ma), 12221046@stu.lzjtu.edu.cn (M. Zhao), sun-belt@163.com (Y. Zhao).
Peer review under responsibility of Periodical Offices of Chang'an University.
https://doi.org/10.1016/j.jtte.2023.05.003

and depth of the integration development of Hadoop and transportation big data are not sufficient. In the future, it is suggested that Hadoop may be combined with other big data frameworks such as Storm and Flink that process real-time data sources to improve the real-time processing and analysis of transportation big data. Simultaneously, the research on multi-source heterogeneous transportation big data is still a key focus. Improving existing big data technology to enable the analysis and even data compression of transportation big data can lead to new breakthroughs for intelligent transportation.

1. Introduction

In the field of transportation, data collection methods have evolved from traditional fixed detection methods such as induction coils to new detection methods such as satellite navigation systems and smartphones. This has aided in the formation and development of transportation big data. The main difference between transportation big data and traditional transportation data is reflected in their characteristics. Transportation big data not only possesses the 6V characteristics, but also has unique 3C characteristics, as shown in Table 1 (Lu et al., 2015).

Traditional methods and models for analyzing transportation data are inadequate for the demands of big data analysis and processing. New theories and methods must be researched and new models established to adapt to the big data context, providing decision support for intelligent transportation management and control (Lyu et al., 2021). The Hadoop cloud computing platform is a powerful tool for efficiently processing big data sets, and its big data tools facilitate the collection, transmission, processing, analysis, storage, and publication of transportation big data. By integrating the Hadoop platform into the traffic research methodology system, information features can be extracted, operational laws can be discovered, and various applications such as travel feature analysis and traffic flow prediction can be enabled.

Given the importance of Hadoop in transportation big data applications, much research has been conducted. Recently, innovative application scenarios and data-driven solutions have emerged due to the increase in computing power and the development of information transmission technologies. Therefore, this study provides a systematic analysis of the important literature in the past 10 years, an in-depth summary of the current state of Hadoop applications in transportation big data, and a discussion of future trends. The contributions of this study can be summarized as follows. First, by reviewing and quantitatively analyzing the literature in the past 10 years, we obtain a co-occurrence visualization network graph of keywords and clearly summarize 8 application directions and research progress of Hadoop in transportation big data. Second, we summarize the research results and applications of optimization to Hadoop from two aspects. Third, we summarize the changes brought by the Hadoop cloud computing platform to transportation big data and discuss future research.

The remainder of this paper is organized as follows. The literature data collection and the results of the bibliometric analysis are presented in Section 2. A detailed description of the Hadoop big data platform is provided in Section 3. Section 4 classifies the existing literature and summarizes the eight Hadoop application scenarios and research progress in transportation big data. Section 5 summarizes the main research results on optimization for Hadoop, and Section 6 discusses the research shortcomings and future research directions. The last section summarizes the results and findings of this study. This study can greatly help researchers understand various application scenarios evolving from the integration of cloud computing technology

Table 1 – Characteristics of transportation big data.

Features	Details	Meaning
6V	Volume	Wide range of data sources and large volumes
	Velocity	Fast data processing is necessary for time-varying and time-sensitive situations
	Variety	Rich types and multi-state characteristics
	Veracity	Missing data, errors, redundancy and other anomalies
	Value	With spatio-temporal multidimensional characteristics, it is the basis of multi-service
	Visualization	Strong visual presentation
3C	Complexity	Huge amount of data and multiple sources
	Continuing	Consistency in time and space
	Connnecting	Interconnectedness and dependency between multiple sources of data

and transportation big data in the context of big data. It also provides support for traffic staff to seek more rational and efficient means of traffic management and control.

2. Data collection and review methods

To conduct a comprehensive and objective review of literature on the application of Hadoop in transportation big data, we conducted a search of both English and Chinese language databases. The search covered various areas, such as taxi operation management, travel feature analysis, traffic flow prediction, traffic event monitoring, transportation big data analysis platform, license plate recognition, shortest path, and transportation infrastructure monitoring. The English literature was screened using at least one of the EI, SCI, and SSCI search types, while the Chinese literature was screened using at least one of the Chinese Science Citation Database (CSCD), Chinese Core Periodicals Catalogue (Peking University Core), Chinese Social Science Citation Index (CSSCI), and EI search types. Using the search strategies mentioned above, we screened a total of 98 Chinese and English literature from 2012 to 2023 related to the application of Hadoop in transportation big data.

Furthermore, we used VOSviewer software, which is developed by the Science and Technology Research Center of Leiden University, to conduct bibliometric analysis by exporting citation information of these literatures. VOSviewer is a widely used software tool for constructing and visualizing bibliometric networks. With its advanced features, such as network layouts and clustering algorithms, the software enabled us to systematically analyze citation information, thus improving the reliability and quality of the literature review.

To analyze the evolution of researchers' interests over time, we extracted all the keywords from the 98 documents and conducted a visual analysis. In the figure, the node size indicates the number of occurrences of each keyword, and the color reflects the average publication year of the literature to which the keywords belong, with blue representing early years and red representing recent years.

As shown in Fig. 1, prior to 2016, researchers focused primarily on building distributed models using MapReduce, data pre-processing, intelligent transportation systems, and taxi operations. From 2016 to 2018, there was a shift towards Hadoop, big data processing and analysis, traffic flow prediction, public transportation, and shortest path-related problems, which enriched and deepened the application of Hadoop in transportation big data. Currently, researchers are placing greater emphasis on integrated traffic management through enhanced computing power. The latest hot topics for researchers include transportation big data platforms, travel feature analysis, traffic infrastructure and event monitoring, and license plate recognition.

3. Hadoop big data platform

The Nutch subproject of Apache Lucene created the Hadoop distributed computing framework, an open-source solution written in Java. This framework enables the creation of a dependable, fault-tolerant, scalable, and extendable

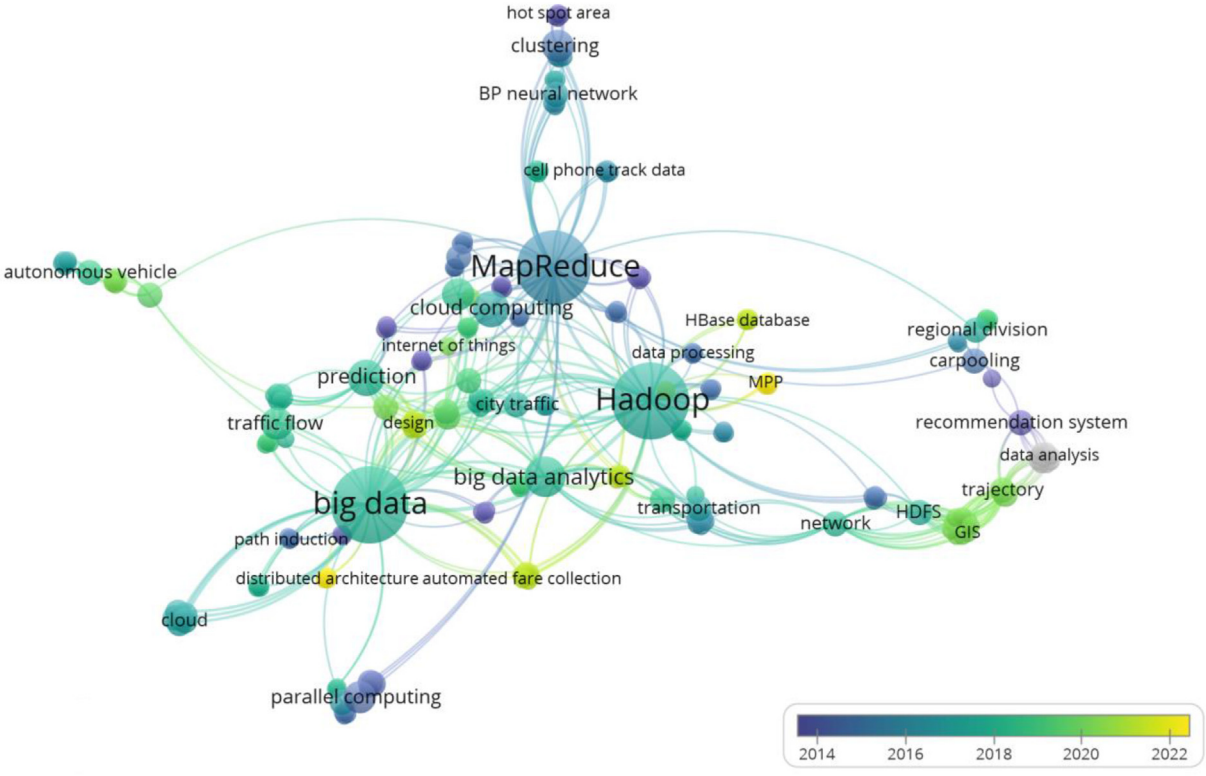


Fig. 1 – Keywords co-occurrence visual network diagram.

architecture for big data batch processing, utilizing the computational and storage capabilities of clusters. The design of Hadoop enables distributed storage and parallel processing of massive amounts of data, making it an ideal solution for handling large-scale data processing needs.

Hadoop's significant advantages have made it a mainstream technology for big data analytics. Many companies and organizations have adopted Hadoop for a variety of use cases. For example, Facebook utilizes a Hadoop cluster for machine learning and data analysis, while Yahoo! uses it for advertising and web search research. Taobao processes and stores its e-commerce transaction data using a Hadoop system, and Baidu employs Hadoop for web data mining and search log analysis. Additionally, the China Mobile Research Institute uses a Hadoop "big cloud" system for data analysis and to provide computing services to the general public.

Fig. 2 illustrates the Hadoop ecosystem, consisting of HDFS and MapReduce, along with several big data tools such as Hadoop YARN, Chukwa, HBase, Hive, Mahout, Pig, Spark, and ZooKeeper (Yang et al., 2017).

3.1. *Distributed file system*

Hadoop distributed file system (HDFS) is an open-source implementation of Google file system (GFS). It's designed to provide high-throughput data access and is well-suited for storing and processing parallel data on a large scale. The fundamental structure of HDFS is illustrated in Fig. 3. It uses a master-slave architecture, with the HDFS cluster consisting of a single metadata node called the NameNode, multiple data nodes called DataNodes, and a secondary node called the secondary Namenode (Kim et al., 2015).

In HDFS, the NameNode serves as the master node that oversees the operation of each slave node. Specifically, it's responsible for keeping track of changes to the NameSpace and managing the NameSpace itself. HDFS data operations follow a "write once, read many times" model. When a file is stored in HDFS, it's typically divided into multiple 64 MB data blocks, each of which is stored on a separate DataNode. The NameNode plays a critical role in this process by managing and mapping the data blocks to the corresponding DataNodes.

When a client wants to read or write a file, it communicates with the appropriate DataNode(s) to perform the requested operation. The DataNode processes the client's requests and carries out actions such as creating or deleting data blocks based on the instructions provided by the NameNode. To access a file in HDFS, a client must first retrieve the location of each data block within the file from the NameNode. It can then retrieve the corresponding data from the appropriate DataNodes. This two-step process ensures that data is retrieved efficiently and accurately.

Overall, HDFS's distributed architecture and data handling mechanisms make it a powerful tool for managing large volumes of data in parallel and distributed environments.

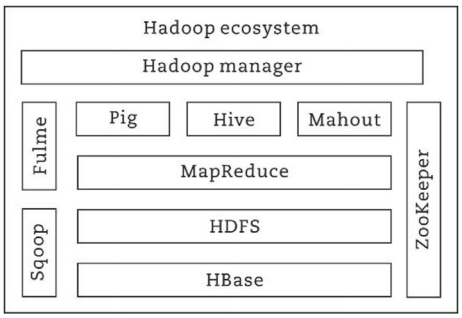


Fig. 2 – Diagram of the Hadoop ecosystem.

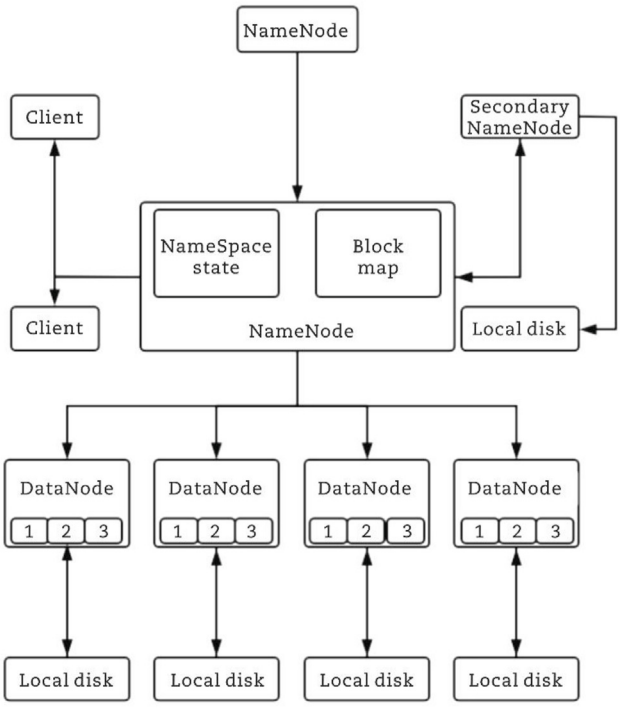


Fig. 3 – HDFS basic structure.

3.2. *Distributed computing framework*

The MapReduce framework, initially proposed by Google in 2004, is a programming model for distributed parallel computing. It enables the processing of vast amounts of data, overcoming the inefficiencies of traditional computing methods. The MapReduce program consists of two main phases: the Map phase and the Reduce phase.

The Map function only accepts input in the <key, value> format, and Hadoop utilizes the InputFormat() method to automatically generate input data as <key, value> pairs for processing by the Map function. The key value denotes the byte offset of each data record in the data slice, and the value represents the content of each row. Similarly, the Reduce function also has input and output in the form of <key, value>. It takes the output of the Map function as input and operates on it. The core idea is to divide tasks into smaller, more manageable portions using the Map process, and then combining the results through the Reduce process. This approach enhances the efficiency of processing large-scale data by leveraging the Hadoop data platform servers to process vast amounts of data in parallel through distributed

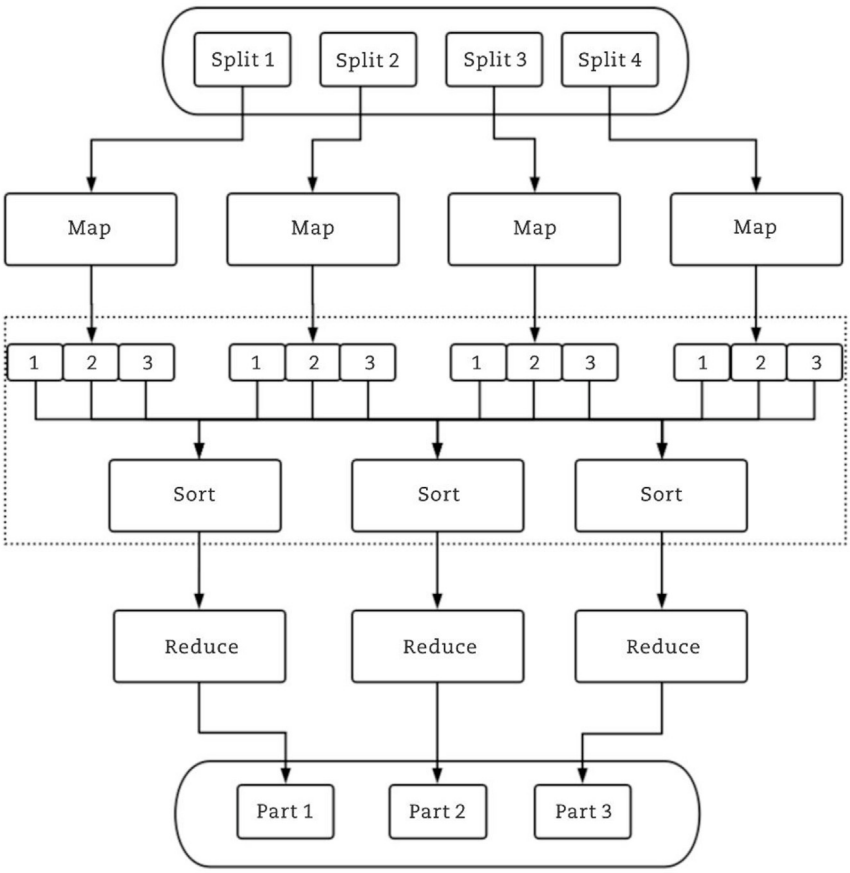


Fig. 4 – MapReduce workflow.

computing. Fig. 4 depicts the typical execution process of a MapReduce program (Kang et al., 2012).

The MapReduce execution process includes the following parts.

(1) Dividing the massive input data into smaller portions and distributing them to different machines for processing.
(2) The Map task worker parses the input data into <key, value> pairs, and the user-defined Map function transforms them into intermediate <key, value> pair.
(3) Sorting and aggregating the intermediate <key, value> pairs based on their key values.
(4) Distributing different key values and their corresponding value sets to various machines to execute the Reduce operation.
(5) Generating the output of the Reduce operation.

4. Application scenarios of Hadoop in transportation big data

The primary sources of transportation big data include fixed detection data (such as geomagnetic coils), mobile detection data (such as floating automotive data), GPS detection data, and smartphone detection data. By utilizing Hadoop big data technology, these data could be mined and analyzed, promoting the research and application of Hadoop big data technology in the transportation industry. Upon reviewing the literature, we have identified eight primary application scenarios that current research in this field has mainly focused on. These scenarios include transportation infrastructure monitoring, taxi operation management, travel feature analysis, traffic flow prediction, traffic event monitoring and status discrimination, transportation big data analysis platforms, license plate recognition, and shortest path.

4.1. *Transportation infrastructure monitoring*

Transportation infrastructure plays a critical role in the transportation network, and any deterioration in it can significantly impact transportation operations. Consequently, some scholars have attempted to create big data platforms based on monitoring and detection data from transportation infrastructure.

The monitoring data for city tunnels consist of various forms, including traffic lane instantaneity, speed, occupancy rate, wind speed, temperature, humidity, light intensity, and CO concentration. Zhong et al. (2014) proposed an improved *k*-means based parallel clustering algorithm for processing data

in cloud computing using MapReduce. This algorithm not only has the advantage of being able to process large amounts of data but is also more efficient. The clustering results from the tunnel data are analyzed to identify abnormal data, ensuring tunnel safety and improving efficiency. To address the challenges of data collection, storage, and analysis, as well as business management and data feedback application, caused by the complexity of shield and tunneling engineering production, Sun et al. (2020) utilized a distributed HBase column storage database and a high-performance Spark in-memory computing engine. They created an industry Hadoop cluster eco-architecture and established a platform that integrates intelligent monitoring, comprehensive analysis, collaborative management, and big data application. This platform provides a comprehensive collaborative service system for tunnel boring machine equipment, construction, risk, and geology, laying the foundation for empirical data analysis and mining. For the characteristics of highway tunnel construction in Karst areas, Li et al. (2018) proposed a Hadoop-based platform for monitoring and prediction of Karst landscape construction. This platform takes monitoring and early warning of landscape features as the starting point to effectively warn and analyze the measurability data as well as risk indicators of mountainous highway tunnel construction in Karst landscapes. This visualizes all unknown factors and risk coefficients in the pre-construction stage. Lastly, to solve the tunnel safety management problem, Wang and Xue (2022) proposed an association rule method Apriori algorithm based on Hadoop platform and applied it to the task of mining historical tunnel safety data. The experimental results demonstrate that the method can use a large amount of historical data for tunnel safety management and provide effective decision support for securing tunnel safety.

The primary objective of bridge health monitoring is to identify safety hazards and evaluate bridge health conditions in a timely and scientific manner. However, current research on bridge health monitoring fails to make full use of the information contained in the monitoring data at various time scales, which makes it challenging to achieve an efficient long-term monitoring mechanism. To address this issue, Yan et al. (2019) proposed a Hadoop platform-based approach to bridge health monitoring data using a time series method for processing bridge health monitoring data. The method achieves the goal of timely monitoring and accurate bridge health status by using a parallelized and improved time series algorithm to build a monitoring model for the intermediate data sets. The simulation results show that the processing efficiency and prediction accuracy of monitoring data using Hadoop platform are significantly higher than the traditional methods. Similar to bridge health monitoring, bridge service condition assessment is mainly based on structural dynamic response, such as natural frequency, vibration mode, stiffness, and damping. While many infrastructure buildings and bridges have a structural health monitoring (SHM) system installed, most health monitoring systems can only analyze historical data and cannot process real-time monitoring data. To address this issue, Wang et al. (2020a) developed a multifunctional Hadoop-Spark big data platform for bridge monitoring and assessment services by combining Hadoop and Spark computing frameworks. The Hadoop-Spark big data platform uses an online real-time data processing module that not only accomplishes real-time data pre-processing but also provides structural safety warning analysis. Finally, important structural response indicators are monitored through data experimental analysis, including temperature, wind speed and direction, humidity, strain, deformation, and train speed. The established big data platform has advantages in offline computing performance, real-time online performance, scalability and fault tolerance, which can meet the actual operational requirements and realize the real-time analysis and offline analysis health monitoring system of high-speed railroad bridges. On this topic, another interesting study showed that Zhao et al. (2022) established an information data chain of crossing vehicles, a strain response, a cracking state, and a vehicle restriction based on the weighing in motion (WIM) system and SHM system data of a highway prestressed concrete box girder bridge. They proposed a set of structural cracking and heavy vehicle overload warning methods oriented to the physical state of the vehicle-bridge system. This method can support the digital operation and maintenance of existing bridges and make up for the technical shortcomings of the existing vehicle overload restriction measures, which do not consider the current safety state of the bridge structure.

4.2. *Taxi operation management*

Taxis play a vital role in urban transportation. Analyzing taxi trajectory data can offer suggestions for managing taxi operations, assess the overall traffic conditions in the city, and ultimately enhance the quality of service provided to taxi passengers.

Wang et al. (2015) aimed to solve the problem of difficult taxi rides by utilizing the Hadoop platform to handle, analyze, and compute large-scale GPS track data. They established a model of taxi probability and waiting time based on empirical distribution at waiting feature locations and time points, which allowed them to forecast the likelihood of a taxi at a specific place and time for users. The feasibility of the scheme was proven, and the model demonstrated strong scalability. Jing and Hu (2016) approached the problem from the passengers' standpoint and employed Hadoop as a data storage and computing platform. They proposed a road network storage structure based on map rasterization to enhance the speed of searching maps. Additionally, they refined a map matching algorithm based on computational geometry to boost the accuracy of matching. Based on experimental results, the empty car probability recommendation algorithm exhibited a correct rate of approximately 87%, while the waiting time recommendation algorithm had a correct rate of 88.4%. Similarly, Lyu et al. (2016) introduced a recommendation system for cabs that implemented region segmentation. The MapReduce computational model is an effective means of collecting and analyzing route information and passage times from specific area markers. This application aids passengers in rapidly connecting with available taxis during inclement weather or periods of heavy traffic. Moreover, the

model enables the instantaneous pairing of two pairs of passengers with similar routes, providing a carpooling service that encourages ride-sharing and helps to reduce traffic congestion. Compared to Zhang et al. (2014) CallCab system, system of Lyu et al. (2016) reduced the total mileage for transporting passengers by 30%. This not only significantly reduces vehicle emissions but also proves to be more efficient in terms of time consumption, which is of greater importance to customers. Additionally, Jiang et al. (2016) proposed an intelligent recommendation method for taxi parking sites based on the MapReduce framework. This method suggests suitable locations for taxi drivers and passengers to facilitate passenger pick-up or to find a taxi easily.

The management of taxi operations is reliant on recommendation algorithms. To enhance the efficiency of these algorithms, Hadoop can serve as a valuable tool for processing data, enabling faster problem-solving. However, there exists a contradiction between the efficiency and accuracy of the algorithm. As such, there is a need for further research to study the balance between accuracy and rationality of the algorithm.

4.3. *Travel feature analysis*

Analyzing multidimensional data resources related to people, vehicles, and roads allows us to monitor the activities and travel patterns of urban residents, providing scientific and precise support for addressing urban traffic problems. This is a critical component for developing macroscopic urban traffic development strategies and construction planning, as well as implementing microscopic road traffic management and control.

From the perspective of analyzing people, Yang et al. (2014a) presented a clustering approach for distributed parallel intelligence analysis (DPIA) that utilizes both MapReduce and ant colony optimization (ACO) algorithms. This approach can be applied in a wide area network (WAN) environment, and can handle the challenge of mining group behavior patterns in large spatio-temporal trajectory data. The method is designed to be implemented in a parallel and incremental manner, allowing for efficient processing of the data. To evaluate the effectiveness of the method, they compared it with the current parallel ACO method that relies on MapReduce, using vehicle trajectory data collected from the road traffic monitoring system in Jiangsu Province. The results indicate that the DPIA method has superior clustering characteristics, making it a promising solution for addressing the challenges of mining group behavior patterns in large spatio-temporal trajectory data in a distributed storage environment. Researchers have also explored the behavior of groups of passengers in civil aviation. Feng et al. (2015) analyzed the conventional recency, frequency, monetary (RFM) model and proposed a new parallel algorithm based on the time, cost, service, destination, and group size (TCSDG) model to segment civil aviation passengers and capture their behavioral preferences. To handle the large volume of booking data, the algorithm was combined with the Hadoop parallelized computing platform. This enabled efficient processing of the data and allowed for more accurate analysis of passenger behavior. Kong et al. (2016) proposed the movement features-based judging urban population flow (MF JUPF) algorithm to address the urban population mobility problem. The algorithm analyzes the data of a user's cell phone base station records using a MapReduce framework to generate user activity trajectories. The algorithm then extracts trajectory features related to in-and-out city behaviors and utilizes a classification model to determine whether users have engaged in such behavior. Experimental results indicate that the method can accurately determine whether users have entered or left the city with an accuracy and recall rate of over 80%.

When evaluating transportation modes, it's important to consider each mode's unique passenger travel characteristics. Gui et al. (2012) proposed a MapReduce-based distributed parallel approach for extracting traffic hotspot regions from taxi trajectories. This approach categorizes taxi trajectory data from different time periods and runs the clustering algorithm in parallel across multiple nodes, compensating for the lack of efficiency of single-computer centralized data processing methods. Experimental results show that the method is more efficient in processing large data volumes and more accurate in extracting hotspot regions. Similarly, Wu et al. (2018) developed a MapReduce-based parallel origin and destination (OD) approach to analyze large public transportation passenger data. The method integrates continuous travel chain and residents' travel characteristics to derive more precise passenger boarding and alighting stations, accurately and comprehensively reflecting passenger flow information. The proposed method was validated using public transport data from Xiamen City. Additionally, Siangsuebchart et al. (2021) developed a taxi trip extraction technique using Hadoop Hive to handle vast amounts of data and determine the OD of each trip. Overall, these studies demonstrate the effectiveness of utilizing distributed computing methods to analyze transportation data and extract valuable insights.

Numerous meaningful studies have investigated travel characteristics from a road network perspective. Statistical analysis of daily vehicle travel patterns is crucial for making informed traffic control decisions and allocating road network resources. Gao and Niu (2018) addressed this issue by combining the Apriori mining algorithm with the Hadoop distributed processing framework to develop a parallelized Apriori algorithm for mining the similarity of travel patterns. The authors of the study also developed a parallelized Hadoop-based *k*-means clustering algorithm for statistical mining of morning and evening fluctuations in road network traffic. This allowed them to establish a daily travel model of urban road vehicles. Similarly, Cao et al. (2019) proposed a parallel Apriori algorithm model based on the fuzzy *C*-means clustering algorithm to identify the route pattern of goods transportation and improve logistics distribution. Additionally, frequent pattern mining has been widely used in the analysis of transportation data. Tang and Chen (2015) proposed a distributed approach to frequent pattern mining to analyze the traffic data flow in the intelligent transportation system in real-time and derive the current state information of the traffic system. Overall, these studies demonstrate the potential of utilizing distributed computing

methods for analyzing traffic data and gaining valuable insights.

4.4. *Traffic flow prediction*

Traffic flow refers to the continuous movement of vehicles on a road, which creates a stream of traffic. However, this term encompasses not only automobiles but also other vehicles and pedestrians. Traffic flow is dynamic, changing constantly in both time and location. It is influenced by numerous factors such as vehicles, pedestrians, and other disruptions, which make it highly unpredictable and uncertain. As a result, predicting traffic flow has become an important aspect of transportation big data processing. By providing accurate information on future traffic patterns, traffic flow prediction can assist traffic managers in making informed decisions about traffic management. Additionally, it can also help drivers to choose less congested routes, thereby avoiding or reducing the impact of traffic congestion.

The *k*-nearest neighbor (KNN) algorithm is a type of supervised learning algorithm that finds the *k* closest training samples in the training set based on a distance metric when given a test sample. It then uses the mean of the output markers of these *k* training samples as the prediction result. Researchers have explored the use of a parallelized *k*-nearest neighbor algorithm in a MapReduce environment to improve the efficiency and accuracy of short-term traffic flow prediction with massive data. For instance, Wang et al. (2014) resolved the efficiency and accuracy issues of short-term traffic flow prediction with massive GPS data by utilizing a parallelized *k*-nearest neighbor algorithm in a MapReduce environment. They improved the state vector and distance vector of the *k*-nearest neighbor short-term traffic flow prediction algorithm. Xia et al. (2016b) introduced a novel MapReduce-based nearest neighbor (NN) approach for predicting traffic flow. The authors developed a real-time forecasting system consisting of two crucial modules: offline distributed training (ODT) and online parallel prediction (OPP). This system was designed to enable faster and more accurate traffic flow forecasting on the Hadoop platform. Liang et al. (2015) increased the search speed of the *k*-nearest neighbor technique by introducing the MapReduce framework for parallel computing. They also used the genetic algorithm to improve the setting of important parameters in the data pre-processing stage. Additionally, they utilized MapReduce to speed up the parameter optimization process to address the issue of the genetic algorithm's lengthy iterative computation time. They presented a non-parametric regression short-term traffic flow prediction technique based on MapReduce and genetic algorithm, which considerably increased the prediction speed and scalability while maintaining the accuracy of traffic flow prediction. Lastly, Wang and Ding (2019) were able to accomplish efficient site-wide traffic flow prediction using a MapReduce-based KNN regression prediction method for high-speed road network large data.

The BP neural network is a multilayer feedforward network that uses the error backpropagation algorithm for training. The learning rule of this algorithm is to modify the network's weights and thresholds through backpropagation, using the quickest descent technique to minimize the network's sum of squared errors. Zhao et al. (2016c) proposed an integrated BP prediction model that utilizes MapReduce. The integrated model combines multiple BP neural network models that are implemented using MapReduce. Tests conducted on this model showed that it has high prediction accuracy and reasonable real-time performance. It can handle a large amount of traffic flow prediction data, and provides a dependable method for predicting the magnitude of vehicle diversion flow at intersections. Xie et al. (2017) developed a segmentation prediction model of bus arrival time based on the combination of MapReduce-based clustering and BP neural network. They built a MapReduce-based parallelization framework for this segmentation model. Experimental results showed that the segmentation model outperforms the conventional BP neural network prediction model in terms of both prediction accuracy and prediction speed.

Currently, the majority of parallelization techniques used to improve traffic flow prediction focus on the parallelized *k*-nearest neighbor and parallelized BP neural network algorithms. Moreover, Chen et al. (2013) proposed a MapReduce-based expectation maximization (EM) algorithm to effectively perform model parameter learning for data-driven traffic flow prediction systems. This is an efficient method for handling large amounts of data. Another approach is the parallelization of the random forest technique using the MapReduce computing model, as demonstrated by Yang et al. (2019), which achieved accurate and real-time predictions of urban short-term traffic flow states. Furthermore, some works in deep learning have employed the long short term memory (LSTM) algorithm and MapReduce to predict short-term traffic flow, as in the study conducted by Xia et al. (2020), which indicates that future traffic conditions can be correctly forecasted using big data, machine learning, and other methods.

Despite these advancements, the basic elements of the traffic system, such as its complexity and chaos, limit the short-term predictability of traffic flow to a narrow range. Additionally, insufficient research has been conducted on the stability of prediction models, and the resilience of relevant models requires further verification. With the development of cloud computing and artificial intelligence technology, it is possible to converge all traffic data directly to the cloud and uniformly process the data using high-performance servers in the cloud. This enables real-time reflection of traffic conditions and the management and control of all aspects of traffic.

4.5. *Transportation big data analytics platform*

In recent years, transportation big data platforms with real-time features have become essential for integrated traffic management in China's major cities. Dong et al. (2014) developed the Guanlan traffic data processing platform, which can run batch and real-time processing tasks simultaneously by combining Apache Hadoop and the S4 open-source architecture. Daniel et al. (2017) proposed an efficient real-time big data analytics architecture for autonomous vehicles that includes a distributed data storage mechanism for real-time analytic stream processing

based on Hadoop, an in-vehicle cloud server tool for batch processing of offline data, and a workflow model for processing stream data in real-time. Babar and Arif (2018) proposed an intelligent transportation system architecture based on big data analysis to achieve real-time big data processing and interaction in an internet of things (IoT-based) intelligent transportation environment. The architecture consists of three phases: big data organization and administration, real-time big data processing, and business management. Data processing is performed on Hadoop using Apache Spark, and the efficacy of the architecture is confirmed by evaluating various datasets. Jan et al. (2019) developed a model to evaluate traffic data using Hadoop and Spark for real-time processing of traffic data, and presented a system capable of handling real-time traffic data processing.

ITS combines various modern technologies, including electronic sensor technology, data transmission technology, and intelligent control technology, to create transportation systems. As a result, ITS generates large amounts of data that can significantly impact its design and implementation. To enhance the processing capability of historical data and obtain dynamic information about goods in logistics centers, Zhao and Huang (2015) proposed an electronic product cod (EPC) IoT data processing algorithm. This approach uses sensing nodes deployed in different logistics center areas and Hadoop technology to process logistics center goods information. Zhao et al. (2016a) proposed an extended MapReduce-based perceptual data processing model that takes into account the diverse characteristics of massive traffic-aware data processing requirements. Subsequently, they developed an integrated traffic-aware data processing platform based on the proposed model. The platform's design and implementation aimed to effectively process and manage large-scale traffic-aware datasets. Li et al. (2020) utilized the fusion of data mining and distributed parallel Hadoop technology to create an intelligent transportation system. They acquired data from the perception layer using IoT and performed operational status analysis using mining algorithms. To describe customer needs and interests from multiple dimensions, Liu et al. (2023) combined the logistics service quality evaluation model with a big data system. They designed a recommendation model for logistics and distribution services with customers as the recommendation core and developed algorithms to provide suitable logistics and distribution services recommendations. Finally, they implemented a Hadoop-based recommendation system for logistics and distribution services. Hadoop technology has been widely utilized in the construction of big data platforms for transportation. Specifically, Table 2 provides a summary of current research that highlights the crucial role played by Hadoop in this regard.

Data collection and processing, based on the development of information technology, is an inevitable path for empowering traditional transportation industries and promoting the development of big data platforms for transportation. The transportation big data platform brings together a large volume of data from many sources and of mixed types. Service users are diverse, including traffic decision-makers, managers, traffic planners, system developers, and others, each

Table 2 — Research on the application of some transportation big data platforms.

Reference	Name of the platform	Design of the platform	Platform features
Huang et al. (2015)	Map matching system	Applying serial transversal map matching algorithms in a MapReduce-based cloud computing environment.	Enables GPS map matching and protection of sensitive GPS data in traffic data centers.
Hua et al. (2016)	Traffic security big data system	HBase stores traffic records.	Realize the analysis of vehicle access to sensitive locations and the analysis of suspect vehicles at crime scenes, etc.
Pan et al. (2020)	Civil aviation air traffic control big data processing platform	MapReduce takes structured, semi-structured, and unstructured data from HDFS and processes it into a structured format.	Provide support for air traffic control (ATC) operations, such as traffic forecasting, conflict detection, intrusion alarms, etc.
Wang et al. (2020b)	Highway intelligent transportation cloud platform	HDFS and HBase are used for big data storage and high throughput data access.	Realize agile, intelligent, and integrated collaboration in decision-making and early warning for highway traffic.
Zhang and Chen (2021)	License plate search platform	HDFS stores videos and Hadoop extracts video keyframes.	Store the video data from highway monitoring to achieve license plate recognition for vehicles.
Shang et al. (2022)	Intelligent networked vehicle big data platform	Real-time vehicle dynamics information is processed in parallel using MapReduce and stored in HDFS.	Real-time collection and processing of vehicle dynamic information, and real-time transmission of optimal path to networked vehicles.
Zhu (2022)	Urban rail transit data center	Process all types of data using Hadoop in combination with MPP architecture.	Perform monitoring, statistical analysis, and other tasks to manage and optimize various systems within the rail network.
Alexakis et al. (2023)	Distributed big data analytics (DBDA) platform	Hadoop for DBDA platform big data management.	The system processes vast amounts of data generated by vehicles and road facilities to assess traffic conditions accurately and enhance its efficiency.

with unique requirements. When designing the data architecture for a platform, it is crucial to consider the specific characteristics of the data and analyze the varied requirements, objectives, and methods of the different users. This process entails the consolidation and classification of transportation big data application scenarios to establish an effective data architecture for the traffic application context. The primary objective of this approach is to address the challenges associated with data aggregation, current state analysis, and decision support. By adopting a structured approach to data architecture design, the platform can effectively manage and utilize the data, thereby achieving optimal outcomes in transportation data analysis and decision-making processes.

4.6. *Traffic event monitoring and status discrimination*

Traffic event detection and status identification involve analyzing collected data to determine the likelihood or severity of traffic events. This process is crucial for ensuring the safety and smooth flow of traffic on roads while reducing the number of fatalities and property damage. There has been a great deal of study and application in this area, particularly regarding the Hadoop big data platform.

In the field of road transportation, Xia et al. (2016a) developed an HBase-based transportation big data processing framework to assess data from the intelligent monitoring and recording system (IMRS), including passing cars' time, position, direction, and license plate number. Experiments on vehicle track tracking confirm that the HBase technique provides superior performance to Oracle, with an average query speed eight times faster than that of Oracle. Massobrio et al. (2018) devised and implemented a distributed computing model that utilized MapReduce on a Hadoop framework to analyze massive amounts of historical bus location data and smart card ticketing data from the intelligent transportation system of Montevideo, Uruguay. The goal was to assess the service quality and passenger mobility of the public transportation system. To tackle the issue of monitoring and analyzing large amounts of traffic video data and to realize the analysis and alarm of abnormal traffic events such as vehicle parking, reversing traffic, congestion, and collisions, Asadianfam et al. (2021b) devised and built the TVD-MRDL system based on MapReduce technology and deep learning. This system is capable of using distributed architecture to analyze data from traffic control centers and detect driver violations (unsafe behavior) to avoid traffic accidents. In a similar study, Asadianfam et al. (2021a) proposed a MapReduce-based convolutional neural network and long and short-term memory network traffic image labeling and classification system to assist traffic management. The system segments the input image and extracts factors such as vehicles, locations, and traffic signs to classify and prevent driver violations in this process.

In the field of train transportation, Wang et al. (2017) devised and implemented a system for the storage and analysis of train signal data based on Hadoop technology. The system aids in monitoring the condition of railroad signal equipment, uncovering hidden signal equipment faults, and enhancing the administration of signal equipment. Jiang et al. (2021) proposed a train wheel pair fault diagnosis method based on big data analysis, which utilizes a recurrent neural network algorithm for feature extraction and MapReduce fast computation to achieve accurate and real-time diagnosis of train wheel pair faults. Similarly, Tan and Cui (2021) developed a MapReduce-based screening tool software for PHM ground system faults of the CRH2 type rolling stock network control system, which is also used for fault detection. Furthermore, the vehicle operation quality trackside dynamic monitoring system (TPDS) is a big data analysis system deployed in the railroad integrated information network internal service network for centralized deployment. The system utilizes B/S mode to provide three-level networking application services, including truck fault diagnosis and prediction, truck operation, and maintenance auxiliary decision-making support. Shi et al. (2022) created a Hadoop-based TPDS to conduct in-depth excavation and regular analysis of TPDS data, providing more effective decision support for truck fault diagnosis and prediction, as well as truck operation and maintenance. Overall, these studies demonstrate the importance of big data analysis in addressing critical issues related to train operations and maintenance, including fault diagnosis, prediction, and decision support. By leveraging advanced technologies such as recurrent neural networks, MapReduce, these studies provide valuable insights into the performance of rolling stock and associated ground systems, which can inform more effective decision-making processes and improve the overall safety and efficiency of train operations.

In the context of water shipping, Zhang (2016) uses Hadoop to store and analyze monitoring data from the quay mooring system. This includes information on flow speed and direction, wind speed and direction, transverse and longitudinal movement, cable tension, wave height, fender pressure, and more. Zhang's work also involves examining and improving the original cloud platform architecture, database, and other monitoring system functionalities for dock moorings. Li (2016) proposed a parallel computing model for defect detection that is based on the fuzzy C-mean clustering technique. This model provides technological assistance for ship diagnostic systems. By leveraging the power of Hadoop, Wu et al. (2020) have demonstrated the potential of big data technologies in addressing key challenges related to ship health monitoring, including the management and processing of large amounts of data in real-time. Overall, this study highlights the significant impact that big data technologies can have in the marine transportation industry, providing valuable insights into ship health and enabling more effective management of critical systems for safer and more efficient marine transportation.

In air transportation, the quick access recorder (QAR) is a device in the aircraft record recording system that plays a vital role in flight quality monitoring, engine condition detection, aircraft system failure diagnosis, and 3D animation analysis. However, the lack of current data warehouses for QAR data prompted Feng et al. (2017) to suggest a Hadoop Hive-based QAR data warehouse. By analyzing the characteristics of Hadoop Hive and QAR data structures, they designed the

general architecture and storage structure of a QAR data warehouse based on Hadoop Hive. In further research, Feng and Liu (2017) utilized the efficient distributed programming and operation framework provided by the MapReduce model to parse data from the aircraft operation monitoring system (ADS) based on GPS location and ground/air data chain communication. The parsed data was then stored in the Hive-based ADS data warehouse. The study aimed to evaluate the trajectory information of aircraft and more accurately predict their arrival time, thereby increasing the percentage of aircraft that arrive on time.

The use of Hadoop in combination with various algorithms has proven to be an effective solution for processing and analyzing large volumes of traffic event data in real-time. By leveraging the power of cloud computing and sensor detection, it is now possible to enhance the accuracy and effectiveness of traffic event monitoring systems. Additionally, the emergence of computer vision technologies has opened up new possibilities for improving the precision and scope of event monitoring.

4.7. *License plate recognition*

License plate recognition technology has become an increasingly popular means of collecting data on urban traffic vehicles in recent years. It works by extracting license plate information from images captured by cameras installed on urban highways, and generating license plate recognition data that includes vehicle identity, as well as time and location information (Zhao et al., 2016b). This technology has the potential to enhance traffic safety, improve urban security, reduce traffic congestion, and enable automated traffic management.

Li et al. (2015) developed a distributed video car retrieval system based on Hadoop. The system extracts video frames in a distributed environment using MapReduce. The license plate numbers are then retrieved using a license plate recognition algorithm, and the time at which vehicles appear in the video is calculated. This approach enables efficient and fast analysis and processing of traffic surveillance data. Li and Liu (2016) proposed a parallel detection method called TP-finder for the detection of other people's license plates on vehicles. This method is based on historical license plate recognition data (ANPR) and employs a data chunking strategy based on integer division to handle data skewing problems in parallel processing of large-scale data. This method significantly improves the performance of detecting other people's license plates on vehicles.

Vehicle landmark recognition (VLR) is a challenging application in intelligent transportation systems due to the complexity of the geometries and surroundings. Convolutional neural networks (CNNs) have demonstrated impressive performance in various machine vision tasks, including VLR. Li and Hu (2018) developed a VLR distributed system framework based on Hadoop and deep learning, which was inspired by the excellent performance of CNNs. They proposed a MapReduce-based CNN, MRCNN, to train the network. This approach considerably improved the training speed, reduced computational cost, and enhanced the recognition accuracy.

Apart from recognizing vehicle signs and license plates, research has also been conducted on recognizing vehicle types to improve the accuracy of traffic volume surveys (Xu et al., 2016). In these recognition tasks, Hadoop is primarily utilized for two purposes: distributed storage of video or image data and parallelization of recognition algorithms. This enables models and algorithms to become more robust and scalable, thereby enhancing recognition accuracy and speed.

4.8. *The shortest route*

The shortest route problem is a crucial matter in both operations research and the transportation industry. As the transportation road network continues to expand outward, its complex scale is also rising, finding the shortest route in large-scale road networks has become a fundamental and essential field of study. In China, Baidu and Alibaba Group have effectively tackled this challenge by developing apps like Baidu Maps and Gaode Maps, which rely on cloud computing technologies.

The use of Hadoop in solving the shortest path problem aims to enhance classical operations research algorithms. In a series of studies conducted by Yang et al. (2013), the ant colony algorithm was parallelized using the MapReduce programming mode, and the simulated annealing algorithm was added to address the limitations of the ant colony algorithm in solving the shortest path problem in urban road networks. The upgraded ant colony algorithm was shown to effectively manage large data sets, making it efficient and practical for solving the shortest route problem. In a similar vein, Yang et al. (2014b) parallelized the genetic algorithm using the MapReduce programming model to improve its shortcomings in solving the shortest path problem in urban road networks. Other researchers have also used Hadoop to enhance the efficiency of algorithms for the shortest path problem. For example, Sadiq et al. (2016) used the Dijkstra method with the Hadoop framework to search for the shortest route in a road network while considering air pollution levels. Praveen and Raj (2023) optimized the traffic management system using Dijkstra's algorithm and Hadoop's powerful data processing capabilities, achieving a traffic control improvement of up to 96.23% compared to the current method. To address the computational complexity of large-scale road network path search algorithms, Zhang et al. (2018) developed a parallel search method based on subgraph partitioning and the MapReduce parallel programming model computing framework, enabling efficient shortest path search in ultra-large-scale real traffic road networks. In addition, Hong et al. (2023) proposed an innovative approach to path planning tasks by developing an improved A* algorithm based on Hadoop and the tile pyramid strategy. The objective was to enhance the speed of path planning tasks for each tile. The proposed algorithm, referred to as OC-RA-A*, uses open and closed lists featuring random access to achieve more efficient path planning. Experimental results showed that the OC-RA-A* algorithm is 3.59 times faster than the conventional A* algorithm in long-distance path planning tasks. The case study presented in Table 3 highlights the potential of Hadoop in solving the shortest route problem.

Table 3 – Case study of the shortest path.

Reference	Method used	Problems solved
Niu and Zhang (2014)	A MapReduce-based parallel algorithm combined with GIS simulation has been developed to solve the shortest path problem.	The shortest route issue in large-scale urban road networks for logistics and distribution.
Chang et al. (2014)	Parallel BP algorithm based on MapReduce model.	Provides bike-sharing riders with the fastest route to a nearby parking location.
Xu et al. (2015)	Dynamic shortest path parallel computing model based on MapReduce.	Dynamic shortest paths in large scale urban road networks.
Niu and Zhang (2015)	A color-coded-based parallelization paradigm for the MapReduce breadth-first algorithm.	Logistics distribution path optimization problem.
Tang et al. (2018)	Algorithm for dynamic vehicle route planning based on the Hadoop big data platform.	Dynamic vehicle route planning issue.
Xu and Guo (2018)	Parallelizing and solving genetic algorithms using MapReduce.	Using large-scale data analysis and processing technologies to optimize vehicle routing issues.

The study's findings show that by using Hadoop, the proposed method achieved better performance and accuracy than traditional methods. This underscores the potential of big data technologies in solving complex transportation problems and improving transportation system efficiency.

Currently, solving the shortest path between ODs using traditional models and algorithms is only applicable to simple travel chains. However, traffic systems are complex, and computing the shortest path between nodes in the entire road network from a system perspective is a significant challenge. To address these challenges, integrating big models and big data with the Hadoop platform has emerged as a popular approach.

In addition, Hadoop has been implemented in vehicle scheduling (Li and Ma, 2019; Lei et al., 2017), civil aviation operation and management (Cao et al., 2015, 2017), ship management and control (Wang, 2020; Yang and Yang, 2021), and other application domains.

5. Optimization of Hadoop

Hadoop is a data processing tool that handles data placement. However, its current method for data placement focuses primarily on balancing data distribution and doesn't take into account the relationships between different datasets. Consequently, all HDFS data is placed based on the workload needs of the Hadoop cluster. This can result in a large amount of data transfer when MapReduce computations are conducted, leading to higher I/O expenses. To improve processing efficiency, several optimization methodologies have emerged. One such approach is CoHadoop, an optimization mechanism developed by IBM that assigns data blocks based on the needs of the application. However, before submitting large data to HDFS, CoHadoop requires partitioning of the data according to the application's requirements, which incurs a significant processing cost. Another tool that has gained popularity for data processing is Spark, a general-purpose parallel computing framework similar to Hadoop MapReduce. Spark also uses the MapReduce algorithm, but with some notable differences. Unlike Hadoop MapReduce, Spark can store the intermediate output of a job in memory, eliminating the need to read and write to HDFS. As a result, Spark is particularly well-suited for data mining and machine learning algorithms that require iterative processing. Table 4 provides a comparison between Hadoop and Spark in terms of various features such as fault tolerance, scalability, language support, visualization, real-time analysis, machine learning, and SQL support. This comparison highlights the distinct application scopes of each tool.

Table 4 – Comparison table between Hadoop and Spark.

	Hadoop	Spark
Fault tolerance	Failure does not require restarting the application.	Recover lost work without additional code or configuration.
Expandability	Has strong scalability potential and has been used in production on tens of thousands of nodes.	Highly scalable with the ability to continuously add n nodes to the cluster.
Language support	Mainly supports Java, others are C, C++, Ruby, Groovy, Perl, Python.	Support for Java, Scala, Python and R.
Visualization	Data visualization is zoomdata's ability to connect directly to HDFS as well as SQL-on-Hadoop technology.	Through a web interface for job submission and execution, or integrated into Apache Zeppelin.
Real-time analysis	MapReduce cannot handle real-time data because it is designed to perform batch processing on large amounts of data.	Real-time data can be processed.
SQL support	Users are able to run SQL queries using Apache Hive.	Users are able to run SQL queries using Spark-SQL.
Machine learning	Machine learning tools like Apache Mahout are needed.	There is a set of machine learning MLlib.

There have been numerous studies published on Hadoop that focus on its computational model optimization and the optimization of Hadoop when used in conjunction with Spark. This section will delve into these two key issues.

5.1. *Computational model optimization*

Since transportation big data consists primarily of dynamic, massive small files and frequent I/O operations of massive small files will reduce Hadoop analysis efficiency. To optimize the efficiency of Hadoop analysis for transportation big data, Zhang et al. (2017) proposed an OpenMP-based Java shared memory multi-threaded parallel programming model. By incorporating this computational model optimization into the Hadoop computing architecture and establishing a distributed parallel processing system, the team achieved a fast data processing mechanism that combines coarse and fine granularity with distributed operation among nodes and multi-threaded parallel computation within nodes on the Hadoop platform. This optimization ensured more accurate prediction of traffic flow during subsequent analysis of transportation big data. Similarly, Xia et al. (2018) suggested a MapReduce-based parallel frequent mode growth (MR-PFP) method for analyzing the spatio-temporal features of taxi operations on the Hadoop platform. The team used large-scale taxi trajectories and large-scale small file processing techniques, creating three techniques including Hadoop archives (HAR), combine file input format (CFIF), and sequence files (SF) to overcome the inherent disadvantages of Hadoop. Findings showed that MR-PFP surpasses conventional parallel FP growth (PFP) algorithms in terms of efficiency and scalability. Overall, these optimization methods provide effective solutions to the challenges posed by dynamic, massive small files and frequent I/O operations in the analysis of transportation big data.

Wei (2013) developed a Hadoop-based big data processing strategy for analyzing e-commerce logistics data called ECLHadoop. By placing related data blocks on the same data nodes, the technique aims to reduce the MapReduce I/O cost, particularly during the shuffling phase. The simulation experiment found that the technique could effectively execute data-intensive analysis in e-commerce logistics services and improve the computational efficiency of e-commerce logistics big data. Similarly, Rathore et al. (2017) developed a system for detecting unlawful traffic conduct, such as illegal U-turns, by analyzing video surveillance data from urban traffic surveillance cameras. The system processed each frame in a parallel environment using Hadoop and GPU, outperforming previous MapReduce implementations powered by CPU. To achieve real-time processing, the authors proposed an algorithm for computing image processing parameters equivalent to the MapReduce mechanism. They also partitioned the image/frame into fixed-size blocks, which enabled the identification of illegal traffic behaviors such as illegal U-turns, drunk driving, and speeding. Zhou et al. (2020) addressed the limitations of current spatio-temporal data indexing techniques for accessing ultra-high volume spatio-temporal trajectory data (HMSTD) in transportation, IoT, or other domains. They proposed path-divided Hadoop distributed file system (HDFS) data blocking (PDDB) based on the Apache Impala (PDDB-Impala) method to optimize the efficient access manner of HMSTD to enhance the efficiency of hyper data sharing. The experimental results showed that PDDB-Impala could enhance the efficiency of retrieving enormous amounts of data compared to Impala. PDDB-Impala also outperformed other high-performance data access platforms, such as MongoDB and HBase. In summary, there are several ways of processing large amounts of data, each with unique issues and bottlenecks that restrict their computational efficiency, especially for certain applications. However, the aforementioned studies demonstrate that leveraging Hadoop and other techniques, such as GPU and PDDB-Impala, can effectively process and analyze big data in various domains.

5.2. *Combining Hadoop and Spark*

By analyzing the demand characteristics of urban taxis and proposing an enhanced parallelized density-based spatial clustering of applications with noise (DBSCAN) method, Zhang et al. (2016) were able to lower the empty rate of cabs by constructing a Hadoop and Spark platform. This allowed them to combine the strengths of Hadoop and Spark to mine and analyze taxi trajectory data. The experiment demonstrated that the algorithm performed well and could advise taxi drivers on cruising directions and routes. Li et al. (2019) extracted a vast amount of taxi trajectory data based on regional density division. They evaluated the spatial distribution features of cabs in a city by computing the hotspot areas divided. Similarly, Wang et al. (2021) utilized the big data processing platform of Hadoop and Spark to store and mine the historical trajectories of cabs in Beijing. They combined real-time information of cabs with historical trajectory data to propose a recommendation strategy of passenger-carrying hotspots. This provided cab drivers with an optimal solution for locating passengers. Overall, these studies demonstrate the potential of using big data analytics to improve the efficiency and effectiveness of urban taxis. By combining various data processing and analysis techniques, researchers have been able to identify patterns and trends that can inform decision-making and improve the overall performance of taxi services.

Zhang et al. (2015b) implemented the batch integrated fast processing (BiF) architecture on the Storm and Hadoop platforms to effectively process continuous traffic surveillance video data. The architecture achieved seamless integration of real-time analysis, batch processing, distributed storage, and cloud services to meet the needs of video data processing and management. The performance, scalability, and fault tolerance of the architecture and video cloud were assessed and found to be effective. Using Hadoop and Spark, Yang et al. (2017) created a cloud-based system to assess the condition of urban traffic. The system provides real-time transit location and traffic status, particularly local real-time traffic status, using open data, cloud computing, bidding data technology, clustering algorithms, and an irregular moving average. Yoo et al. (2020) designed a sensor-based big data processing system for automatic driving automobiles in a C-ITS environment based on the

Hadoop ecosystem (Hadoop, Spark, and Kafka). The system enhanced the precision of road condition perception and traffic information. Rathore et al. (2021) proposed a traffic control model for continuous monitoring and mining of vehicle data in large cities using information physical systems (CPS) and sensor technologies. They constructed an intelligent transportation system based on big data and graph generation techniques, and their parallel processing. The automobile traffic data from Denmark, Spain, and Germany were saved in graph form using the Hadoop distributed file system (HDFS). The graphs were analyzed using GraphX, and the data was processed in real-time using the Hadoop ecosystem and Spark. The study found that combining Spark GraphX with the Hadoop environment significantly enhanced the overall system's efficacy. In summary, these studies demonstrate the potential of using Spark and the Hadoop ecosystem to process and manage continuous surveillance video data, assess the condition of urban traffic, enhance the precision of road condition perception and traffic information, and construct intelligent traffic systems. By combining various data processing and analysis techniques, researchers have been able to identify patterns and trends that can inform decision-making and improve the overall performance of traffic systems.

6. Discussions

The transportation sector is increasingly utilizing big data, as vast amounts of data can go to waste if not utilized effectively. Big data technology can be used to mine knowledge from transportation data and identify general laws governing human traffic actions. This can help improve traffic conditions and address various traffic issues. By examining relevant literature, the following interesting topics can be shown for a deeper understanding of research on using Hadoop to analyze massive traffic data.

(1) Several application studies have detailed the procedure for resolving traffic issues using the Hadoop big data platform. The procedure involves collecting both static and dynamic data, such as road environment, vehicle information, and implementation data like vehicle speed using GPS and network signals. The collected data is then processed using MapReduce, which facilitates data interchange across data exchange centers. The data is stored on Hadoop's HDFS for efficient integration and processing. The Hadoop-based control center visualizes these massive datasets to aid in resolving complex traffic issues. Hadoop outperforms other systems due to its maturity, reliability, and ability to efficiently process data in a distributed and parallelized manner, providing a robust architecture for big data analysis. Hence, using Hadoop to construct an intelligent transportation platform enhances the value and increases the usefulness of traffic data. However, research on structured and unstructured traffic data is unequal. Therefore, analyzing and processing unstructured data using Hadoop could enhance the exploitation of traffic data and expand Hadoop's application scenarios in transportation big data analysis.

(2) Hadoop is a widely used framework for big data analysis that allows for the storage and analysis of large datasets to be quicker, more stable, and more precise. Its effectiveness in processing traffic data has been extensively demonstrated. Academic studies have shown that the MapReduce computing framework processes data approximately 40 times faster than standard Matlab tools (Zhang et al., 2015a). However, speed alone is not sufficient to ensure the accuracy or resolution of traffic issues. Algorithms, models, and systems are critical for addressing various traffic challenges. Therefore, the comprehensive integration of algorithms, models, and systems with Hadoop is essential to establish the fundamental technology of transportation big data. Currently, there is continuous development and improvement of new technologies that combine transportation with big data, gradually forming core technologies. These core technologies can be broadly categorized into four areas: large-scale parallel computing of road networks, path dynamic programming and intelligent search technology, big data analysis and optimization of transportation systems, and operation control and organizational scheduling. Each of these technologies has been extensively researched and is replete with models and techniques. However, there is a lack of systematic research to integrate the existing findings from an overall technology perspective, and most existing research has focused on the application of core technologies in a specific scenario. As a result, it is challenging to develop a comprehensive and robust big data technology system for transportation that uses Hadoop as the basic framework and various algorithm models as key components. Continuous research and development are required to ensure that core technologies can produce a more comprehensive and robust system.

(3) Research related to the integration of Hadoop and transportation big data can be classified into two trends, breadth and depth of integration. The breadth of integration refers to the widespread use of Hadoop as a big data analytics platform with significant advantages, but not the only one, as other data analytics frameworks such as Spark and Flink are also used to improve practicality. Therefore, the application of various big data analysis frameworks in transportation big data reflects the need for different frameworks that are applied to different degrees in transportation big data. As big data technology continues to integrate into the transportation field, richer big data technology will be integrated into transportation big data to solve transportation-related problems in the future. The depth of integration, however, presents a contradiction between universality and accuracy. While universally applicable algorithms or frameworks may show limitations or other contradictions when targeting different problems, a more precise algorithm and framework for solving specific problems can bring new breakthroughs

to the problem-solving approach. While most literature compiles Map functions and Reduce functions using Java language to achieve corresponding algorithmic functions in the integration of Hadoop and big data in transportation, there has been significant research on improving Hadoop core development source code for specific problems in the field of big data. This demand-oriented approach improves the Hadoop distributed architecture and becomes a precise algorithm and framework for solving specific problems, bringing new breakthroughs to the problem-solving approach. It is essential to note that many existing algorithms do not reflect the spot-on approach to traffic pain points when solving transportation problems, but only stay on the surface of big data technology applications. They do not delve deep into the intersection of the core development ideas of big data technology and traffic problem-solving ideas. As a result, the integration of Hadoop and transportation big data is not deep enough. To address this issue, more research is necessary to integrate big data technology more deeply into transportation and explore precise algorithms and frameworks for solving specific transportation problems.

7. Conclusions

The exponential growth of data resulting from the advancement of information technology has made big data analysis an inevitable outcome. With the widespread application of the internet and communication technology, the transportation industry generates massive real-time data that requires urgent analysis and processing. Undoubtedly, Hadoop has become one of the mainstream technologies for big data analysis and is widely studied in academic circles as a cloud computing platform. Therefore, researching the application of Hadoop in transportation big data is crucial. Due to the complex nature of transportation big data problems, using Hadoop big data technology to study them has become a hot topic. To gain a better understanding of research on Hadoop in transportation big data, we analyzed 98 collected papers and identified eight application scenarios of Hadoop in transportation big data. We also summarized the development process and latest achievements of established research in this area. Furthermore, we focused on literature that examined the optimization aspects of Hadoop and identified the latest research progress in the transportation field. Finally, we identified gaps in the current research. From our review and bibliometric analysis of the literature, we draw the following conclusions.

(1) Real-time data plays a crucial role in various transportation big data applications, including traffic status identification, real-time traffic control, dynamic route guiding, and real-time bus scheduling. However, Hadoop has limitations in handling real-time data. Therefore, integrating Hadoop with other big data frameworks designed specifically for real-time data processing, such as Apache Storm, Apache Flink, Apache Samza, and Kafka Streams, can provide effective solutions for real-time big data analytics in transportation. Further research on the integration and development of these frameworks with transportation big data can lead to new advances in the application of Hadoop big data technology to transportation.

(2) The fundamental mind behind big data is processing complex systems, and the traffic problem serves as a prime example of such a system. Traffic involves a broad range of interconnected factors, such as traffic flow, road conditions, driver behavior, and weather conditions, that are highly dynamic and subject to rapid changes, making it difficult to develop effective solutions. Therefore, a comprehensive approach is necessary to address traffic issues, one that takes into account all relevant factors and their interdependencies. This approach utilizes big data analytics to gain insights into the system and to develop data-driven solutions that can reduce congestion, minimize accidents, and enhance overall transportation efficiency. However, existing big data technologies, such as Hadoop, have limitations in dealing with relational data, particularly when it comes to analyzing multi-source and heterogeneous traffic data. Addressing this challenge requires the integration of cross-modal, multi-technology, and cross-domain processing to enable multidimensional correlations for large data sets.

(3) Hadoop distributed file system (HDFS) plays an integral role in facilitating distributed storage within the Hadoop ecosystem. However, leveraging HDFS for processing large datasets demands strict adherence to software and hardware requirements. Fortunately, data compression techniques can effectively reduce storage space requirements, thereby mitigating some of these limitations. Additionally, the statistical similarity between data compression and data analysis processes implies that encoding and decoding data through artificial intelligence can improve data analysis and even replace some HDFS functions. The transportation industry can reap significant benefits from this exciting prospect. However, it also poses significant challenges, demanding interdisciplinary expertise in artificial intelligence and transportation engineering to develop and implement efficient data compression and analysis approaches. Bridging the gap between these two fields is crucial for advancing the application of big data technology in transportation.

The objective of this study is to present a comprehensive and lucid overview of the advances in applying Hadoop technology to transportation big data during the past decade. Our research aims to offer a novel perspective that can assist scholars in comprehending the present state of the field and identifying future research directions. It is important to note that our investigation focuses exclusively on Hadoop cloud computing technologies due to certain limitations. Nevertheless, we believe that our study can be a valuable resource for researchers intending to conduct a similar analysis of other cloud computing technologies implemented in transportation big data. Such studies may uncover additional unique insights and facilitate future research in this domain.

Conflict of interest

All authors declare that there are no competing interests.

Acknowledgments

This research was supported by the Natural Science Foundation of China (No. 52062027), the Key Research and Development Project of Gansu Province (No. 22YF7GA142), Soft Science Special Project of Gansu Basic Research PIan (No. 22JR4ZA035), Gansu Provincial Science and Technology Major Special Project-Enterprise Innovation Consortium Project (No. 22ZD6GA010 and No. 21ZD3GA002), Lanzhou Jiaotong University Basic Research Top Talents Training Program (No. 2022JC02).

REFERENCES

Alexakis, T., Peppes, N., Demestichas, K., et al., 2023. A distributed big data analytics architecture for vehicle sensor data. Sensors 23 (1), 23010357.

Asadianfam, S., Shamsi, M., Kenari, A.R., 2021a. Hadoop deep neural network for offending drivers. Journal of Ambient Intelligence and Humanized Computing 13 (1), 659–671.

Asadianfam, S., Shamsi, M., Kenari, A.R., 2021b. TVD-MRDL: traffic violation detection system using MapReduce-based deep learning for large-scale data. Multimedia Tools and Applications 80 (2), 2489–2516.

Babar, M., Arif, F., 2018. Real-time data processing scheme using big data analytics in internet of things based smart transportation environment. Journal of Ambient Intelligence and Humanized Computing 10 (10), 4167–4177.

Cao, W., Bai, L., Nie, X., 2015. Map/Reduce based high value passenger discovery method for civil aviation. Computer Engineering and Design 36 (4), 1078–1083.

Cao, J., Ren, X., Xu, X., 2019. Research on frequent patterns of logistics paths based on parallel Apriori. Computer Engineering and Applications 55 (11), 257–264.

Cao, W., Zhai, P., Zhu, Y., 2017. Research on MapReduce-based revenue vulnerability rule extraction for civil aviation. Computer Simulation 34 (12), 9–13.

Chang, H., Huang, S., Lin, Y., 2014. An efficient cloud-assisted best-parking algorithm for BikeNet. International Journal of Ad Hoc and Ubiquitous Computing 16 (2), 136.

Chen, C., Liu, Z., Lin, W., et al., 2013. Distributed modeling in a MapReduce framework for data-driven traffic flow forecasting. IEEE Transactions on Intelligent Transportation Systems 14 (1), 22–33.

Daniel, A., Subburathinam, K., Paul, A., et al., 2017. Big autonomous vehicular data classifications: towards procuring intelligence in ITS. Vehicular Communications 9, 306–312.

Dong, Z., Yu, X., Cui, X., et al., 2014. Guanlan traffic data processing platform. Computer Research and Development 51 (S2), 129–133.

Feng, X., Liu, F., 2017. ADS-B data parsing and storage method based on Hadoop. Aerospace Control 35 (5), 80–86, 97.

Feng, X., Wu, X., Zhao, J., et al., 2017. Construction of QAR data warehouse in Hive. Computer Engineering and Applications 53 (11), 90–94.

Feng, X., Xu, B., Lu, M., 2015. Analysis of civil aviation passenger booking behavior segmentation and group characteristics. Computer Engineering and Design 36 (8), 2217–2222.

Gao, Z., Niu, Z., 2018. Analysis of traffic flow patterns based on big data. Journal of Harbin University of Technology 23 (6), 124–127.

Gui, Z., Xiang, Y., Li, Y., 2012. Parallel urban hotspot area discovery based on cab trajectory. Journal of Huazhong University of Science and Technology (Natural Science Edition) 40 (S1), 187–190.

Hong, Z., Tu, B., Tong, X., et al., 2023. A fast large-scale path planning method on lunar DEM using distributed tile pyramid strategy. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 16, 344–355.

Hua, X., Wang, J., Lei, L., et al., 2016. H-TDMS: a system for traffic big data management. In: 2016 Conference on Advanced Computer Architecture, Weihai, 2016.

Huang, J., Qie, J., Liu, C., et al., 2015. Cloud computing-based map-matching for transportation data center. Electronic Commerce Research and Applications 14 (6), 431–443.

Jan, B., Farman, H., Khan, M., et al., 2019. Designing a smart transportation system: an internet of things and big data approach. IEEE Wireless Communications 26 (4), 73–79.

Jiang, H., Zhang, J., Fang, R., et al., 2021. A method of train wheel pair fault diagnosis based on big data analysis. Journal of Hunan University of Science and Technology (Natural Science Edition) 36 (1), 91–98.

Jiang, H., Zhang, N., Li, K., et al., 2016. MapReduce-based intelligent recommendation algorithm for cab parking spots. Computer Application and Software 33 (2), 254–258.

Jing, W., Hu, L., 2016. Passenger recommendation algorithm based on Hadoop and cab history trajectory. Computer Engineering and Applications 52 (7), 264–270.

Kang, L., Wang, X., Bai, R., 2012. Analysis of MapReduce principle and its main implementation platform. Modern Library and Information Technology 2012 (2), 60–67.

Kim, Y., Araragi, T., Nakamura, J., et al., 2015. A distributed and cooperative NameNode cluster for a highly-available Hadoop distributed file system. IEICE Transactions on Information and Systems E98.D (4), 835–851.

Kong, Y., Jin, C., Wang, X., 2016. Analysis of population mobility based on cell phone trajectory data. Computer Applications 36 (1), 44–51.

Lei, Y., Lin, P., Yao, K., 2017. Network scheduling model of Internet customized bus and its solution algorithm. Transportation System Engineering and Information 17 (1), 157–163.

Li, T., 2016. Application of cloud computing in remote fault diagnosis of naval equipment. Ship Science and Technology 38 (4), 178–180.

Li, J., Chen, Z., Su, Z., 2019. Research on the analysis method of vehicle trajectory data based on area density division. Microelectronics and Computers 36 (2), 53–56.

Li, B., Hu, X., 2018. Effective vehicle logo recognition in real-world application using MapReduce based convolutional neural networks with a pre-training strategy. Journal of Intelligent & Fuzzy Systems 34 (3), 1985–1994.

Li, Y., Liu, C., 2016. A parallel detection method for registration plate cars based on historical license plate recognition data. Computer Applications 36 (3), 864–870.

Li, X., Ma, H., 2019. Vehicle mobile cloud reliability task scheduling. Computer Application and Software 36 (11), 78–85.

Li, Y., Qiu, H., Li, Y., 2015. A distributed video vehicle retrieval method based on Hadoop. Television Technology 39 (22), 95–99.

Li, Z., Yu, Y., Qing, Z., 2018. Optimization of tunnel construction plan for mountainous highway under Karst terrain. Highway 63 (11), 326–328.
Li, W., Zhu, J., Zhang, Y., et al., 2020. Design and implementation of intelligent traffic and big data mining system based on internet of things. Journal of Intelligent & Fuzzy Systems 38 (2), 1967–1975.
Liang, K., Tan, J., Li, Y., 2015. A MapReduce-based method for short-time traffic flow prediction. Computer Engineering 41 (1), 174–179.
Liu, X., Sun, M., Liu, Y., 2023. Research on logistics service recommendation model and application under mobile cloud environment. Optik 273, 170446.
Lu, H., Sun, Z., Qu, W., 2015. A review of big data and its application in urban intelligent transportation system. Transportation System Engineering and Information 15 (5), 45–52.
Lyu, T., Wang, P.S., Gao, Y., et al., 2021. Research on the big data of traditional taxi and online car-hailing: a system review. Journal of Traffic and Transportation Engineering (English Edition) 8 (1), 1–34.
Lyu, H., Xia, S., Yang, X., et al., 2016. A unified recommendation algorithm for cabs based on area division. Computer Applications 36 (8), 2109–2113.
Massobrio, R., Nesmachnow, S., Tchernykh, A., et al., 2018. Towards a cloud computing paradigm for big data analysis in smart cities. Programming and Computer Software 44 (3), 181–189.
Niu, L., Zhang, B., 2014. MapReduce to solve the shortest path of logistics and distribution single source. Electronic Technology Applications 40 (3), 123–125, 129.
Niu, L., Zhang, B., 2015. Research on solving the shortest path of urban logistics distribution based on cloud computing. Science and Technology Bulletin 31 (5), 184–188, 213.
Pan, W., Liu, J., Wang, R., et al., 2020. Research on the architecture of civil aviation air traffic control big data processing platform. Computer Application and Software 37 (6), 48–52, 113.
Praveen, D.S., Raj, D.P., 2023. Retraction note to: smart traffic management system in metropolitan cities. Journal of Ambient Intelligence and Humanized Computing 14 (S1), https://doi.org/10.1007/s12652-002-04064-9.
Rathore, M.M., Shah, S.A., Awad, A., et al., 2021. A cyber-physical system and graph-based approach for transportation management in smart cities. Sustainability 13, 7606.
Rathore, M.M., Son, H., Ahmad, A., et al., 2017. Real-time video processing for traffic control in smart city using Hadoop ecosystem with GPUs. Soft Computing 22 (5), 1533–1544.
Sadiq, A., El Fazziki, A., Ouarzazi, J., et al., 2016. Towards an agent based traffic regulation and recommendation system for the on-road air quality control. SpringerPlus 5 (1), 1604.
Shang, J., Liu, H., Li, W., 2022. Human-computer interaction of networked vehicles based on big data and hybrid intelligent algorithm. Wireless Communications and Mobile Computing 2022, 5281132.
Shi, X., Feng, L., Jiang, A., et al., 2022. Design of TPDS big data analysis system. Railway Rolling Stock 42 (1), 95–98.
Siangsuebchart, S., Ninsawat, S., Witayangkurn, A., et al., 2021. Public transport GPS probe and rail gate data for assessing the pattern of human mobility in the Bangkok metropolitan region, Thailand. Sustainability 13 (4), 2178.
Sun, Z., Qian, T., Ren, Y., et al., 2020. Study on key technologies and application of engineering big data management platform of tunnel boring machine. Tunnel Construction 40 (6), 783–792.
Tan, S., Cui, Y., 2021. Design and implementation of PHM fault screening for network control system of CRH2 EMU based on MapReduce. Locomotive Electric Transmission 2021 (1), 110–114.
Tang, Y., Chen, S., 2015. A closed frequent pattern mining method for distributed data streams. Computer Application Research 32 (12), 3560–3564, 3595.
Tang, D., Huang, J., Shi, W., 2018. Dynamic vehicle path scheduling algorithm based on big data platform. Computer Engineering 44 (1), 74–78.
Wang, Y., 2020. Effective storage method of big data of ship mobile network under cloud computing environment. Ship Science and Technology 42 (22), 154–156.
Wang, X., Ding, W., 2019. A short-time traffic prediction method for highway big data. Computer Applications 39 (1), 87–92.
Wang, M., Ding, Y., Wan, C., et al., 2020a. Big data platform for health monitoring systems of multiple bridges. Structural Monitoring and Maintenance 7 (4), 345–365.
Wang, Z., Li, T., Cheng, Y., et al., 2015. Prediction of taxi probability and waiting time based on empirical distribution. Computer Engineering and Applications 51 (24), 254–259.
Wang, B., Li, S., Zhu, P., 2020b. Design and implementation of highway intelligent traffic cloud platform. Journal of Shandong Agricultural University (Natural Science Edition) 51 (3), 503–506.
Wang, W., Liao, Z., Zhang, H., et al., 2017. Design and implementation of data storage and analysis system of railway signal system based on big data. Information Network Security 2017 (1), 29–37.
Wang, T., Shen, Z., Cao, Y., et al., 2021. Taxi-cruising recommendation via real-time information and historical trajectory data. IEEE Transactions on Intelligent Transportation Systems 2418, 1–13.
Wang, B., Tao, L., Gao, C., et al., 2014. Dividing traffic sub-areas based on a parallel k-means algorithm. Knowledge Science, Engineering and Management 2014, 127–137.
Wang, Q., Xue, T., 2022. Tunnel security management based on association rule mining under Hadoop platform. Mathematical Problems in Engineering 2022, 8508273.
Wei, F., 2013. ECLHadoop: an effective e-commerce logistics big data processing strategy based on Hadoop. Computer Engineering and Science 35 (10), 65–71.
Wu, J., Chen, Z., Yan, Z., et al., 2020. Design of Hadoop-based ship bearing health status monitoring system. Journal of Arms Equipment Engineering 41 (1), 140–144.
Wu, Q., Su, K., Zou, Z., 2018. MapReduce-based parallel projection method for massive bus passenger OD. Journal of Geoinformation Science 20 (5), 647–655.
Xia, Y., Chen, J., Lu, X., et al., 2016a. Big traffic data processing framework for intelligent monitoring and recording systems. Neurocomputing 181, 139–146.
Xia, D., Li, H., Wang, B., et al., 2016b. A MapReduce-based nearest neighbor approach for big-data-driven traffic flow prediction. IEEE Access 4, 2920–2934.
Xia, D., Lu, X., Li, H., et al., 2018. A MapReduce-based parallel frequent pattern growth algorithm for spatio temporal association analysis of mobile trajectory big data. Complexity 2018, 1–16.
Xia, D., Zhang, M., Yan, X., et al., 2020. A distributed WND-LSTM model on MapReduce for short-term traffic flow prediction. Neural Computing and Applications 33 (7), 2393–2410.
Xie, F., Gu, J., Zhang, S., et al., 2017. Bus arrival time prediction model based on MapReduce clustering and neural network. Computer Applications 37 (S1), 118–122.
Xu, X., Ding, Q., Bai, B., et al., 2016. Research on classification of road video image models based on MapReduce. Television Technology 40 (3), 111–115, 145.
Xu, J., Guo, J., 2018. Research on the path problem of logistics distribution vehicles based on big data platform. Transportation System Engineering and Information 18 (S1), 86–93.

Xu, J., Wang, Y., Lin, P., 2015. Dynamic shortest path algorithm in big data environment. Journal of South China University of Technology (Natural Science Edition) 43 (10), 1–7.
Yan, F., Zhang, X., Li, W., et al., 2019. Simulation study on data processing for bridge construction quality operation monitoring. Computer Simulation 36 (1), 441–444.
Yang, Z., Chen, H., Wang, C., et al., 2019. Urban short-time traffic flow prediction in the context of big data. Highway Traffic Science and Technology 36 (2), 136–143.
Yang, C., Chen, S., Yan, Y., 2017. The implementation of a cloud city traffic state assessment system using a novel big data architecture. Cluster Computing 20 (2), 1101–1121.
Yang, J., Li, S., Chen, S., 2014a. A population mining method based on incremental spatio-temporal trajectory big data. Computer Research and Development 51 (S2), 76–85.
Yang, Q., Mei, D., Han, Z., et al., 2013. Cloud-based ant colony algorithm for solving the shortest path of urban road network. Journal of Jilin University (Engineering Edition) 43 (5), 1210–1214.
Yang, Q., Mei, D., Zheng, L., et al., 2014b. Cloud-based genetic algorithm for solving shortest path of urban road network. Journal of South China University of Technology (Natural Science Edition) 42 (3), 47–51, 58.
Yang, F., Yang, T., 2021. Ship information control system based on cloud computing technology. Ship Science and Technology 43 (22), 136–138.
Yoo, A., Shin, S., Lee, J., et al., 2020. Implementation of a sensor big data processing system for autonomous vehicles in the C-ITS environment. Applied Sciences 10 (21), 7858.
Zhang, R., 2016. Research on cloud computing platform in data processing of ship terminal mooring monitoring system. Ship Science and Technology 38 (8), 163–165.
Zhang, T., Chen, Y., 2021. Hadoop-based design of a specific vehicle license plate retrieval platform. Highway 66 (1), 248–251.
Zhang, L., Chen, C., Wang, Y., et al., 2016. Exploiting taxi demand hotspots based on vehicular big data analytics. In: 2016 IEEE 84th Vehicular Technology Conference (VTC-Fall), Montreal, 2016.
Zhang, D., He, T., Liu, Y., et al., 2014. A carpooling recommendation system for taxicab services. IEEE Transactions on Emerging Topics in Computing 2 (3), 254–266.
Zhang, D., Lin, Y., Lu, K., et al., 2018. Shortest path algorithm for large-scale traffic road network based on cloud computing. Journal of South China University of Technology (Natural Science Edition) 46 (12), 139–146.
Zhang, R., Shu, Y., Yang, Z., et al., 2015a. Hybrid traffic speed modeling and prediction using real-world data. In: 2015 IEEE International Congress on Big Data, Santa Clara, 2015.
Zhang, H., Wang, X., Cao, J., et al., 2017. A hybrid short-term traffic flow forecasting model based on time series multifractal characteristics. Applied Intelligence 48 (8), 2429–2440.
Zhang, W., Xu, L., Duan, P., et al., 2015b. A video cloud platform combing online and offline cloud computing technologies. Personal and Ubiquitous Computing 19 (7), 1099–1110.
Zhao, Z., Ding, W., Han, Y., 2016a. A cloud-based architecture for traffic-aware data integration and processing platform. Computer Research and Development 53 (6), 1332–1341.
Zhao, H., Ding, Y., Li, A., et al., 2022. Digital modeling of vehicle load-bridge effect and system state monitoring. Journal of Southeast University (Natural Science Edition) 52 (2), 203–211.
Zhao, Z., Ding, W., Zhang, S., 2016b. A travel time calculation method based on spatio-temporal division on massive license plate recognition dataset. Journal of Electronics 44 (5), 1227–1233.
Zhao, H., Huang, C., 2015. Research and implementation of a Hadoop-based EPC IoT data analysis system. Computer Engineering and Science 37 (4), 657–662.
Zhao, H., Luo, J., Yang, J., et al., 2016c. Research and application of integrated BP neural network prediction model. Telecommunications Science 32 (2), 60–67.
Zhong, L., Tang, K., Li, L., et al., 2014. An improved clustering algorithm of tunnel monitoring data for cloud computing. The Scientific World Journal 2014, 630986.
Zhou, L., Li, Q., Tu, W., 2020. An efficient access model of massive spatio temporal vehicle trajectory data in smart city. IEEE Access 8, 52452–52465.
Zhu, J., 2022. Construction scheme of urban rail transit big data center based on Hadoop+MPP architecture. Urban Rail Transit Research 25 (5), 54–57.

Changxi Ma received the BS degree in traffic engineering from Huazhong University of Science and Technology in 2002 and the PhD degree in transportation planning and management from Lanzhou Jiaotong University in 2013. He is currently a professor in Lanzhou Jiaotong University. He is the author of three books and more than 100 articles. His research interests include ITS, traffic safety, and hazardous materials transportation.