Exploring the Application of Databases and Data Science in Logistics Enterprises

Yumin Hu[1*]

[1]Guangzhou Foreign Language School, 511455 Guangzhou, China

Abstract. This paper explores the innovative applications of databases in the logistics industry through a literature review and clarifies the background and significance of information digitization in contemporary logistics enterprises. It systematically sorts out several innovative data modeling methods proposed by researchers, and based on the feasibility, limitations and application prospects of these methods, conducts analysis and evaluation on databases such as My Structured Query Language (MySQL) and Neo4j graph database used in the data models. The paper finds that existing models have certain shortcomings, mainly reflected in the mismatch between database performance or operation principles and the datasets required in the contemporary era. Targeting the characteristics of different models, the paper puts forward relevant improvement suggestions: for example, introducing databases more suitable for storing and calling logistics information and combining them with the databases cited by researchers; or optimizing the index algorithms for logistics distribution routes. Finally, the paper looks forward to the prospects of the practical application of big data models in the logistics field in the future.

1 Introduction

In this era of rapid information technology advancement, big data and data science are increasingly becoming pivotal forces driving transformation across all industries. As a vital link connecting production and consumption, the logistics sector's operational efficiency and service quality profoundly impact economic development and social life [1]. In recent years, the explosive growth of e-commerce has led to a surge in data volumes that logistics companies must process, while simultaneously demanding higher processing efficiency. Against this backdrop, the IT department—as the core driver of digital transformation within logistics enterprises—faces the urgent need to leverage big data and data science knowledge to achieve technological upgrades and functional transformation.

Big data and data science technologies present unprecedented opportunities for logistics IT departments. Research indicates that by efficiently collecting, storing, and processing massive datasets, IT departments can provide robust data support for enterprise functions including transportation management, warehouse optimization, route planning, and resource scheduling. For instance, real-time traffic data, weather information, and historical transport

* Corresponding author: angiehu236@gmail.com

records can help logistics companies build dynamic route optimization models, significantly improving delivery efficiency. IoT devices and sensor data enable dynamic inventory monitoring and automated management, reducing operational costs in warehouse management [2].

Furthermore, big data technology provides crucial support for networked logistics distribution. Research indicates that real-time collection of multi-source data—such as cargo location, transport status, and inventory levels—via IoT devices and information systems, coupled with efficient management through distributed storage technology, establishes a data foundation for logistics optimization. At the analytical level, techniques like time series analysis, machine learning, and intelligent algorithms enable precise demand forecasting, dynamic route planning, and facility site optimization. These methods significantly enhance delivery efficiency, reduce operational costs, and strengthen the network's resilience against uncertainty [3].

Moreover, big data has gradually become an integral part of our lives. We can reasonably predict that in the future, it will further enhance transportation infrastructure and service quality within the logistics industry, promoting its efficient management. However, as the digitalization process advances, the logistics sector will also face new challenges. The demand for digital logistics operations continues to grow. Digital transformation can help improve logistics efficiency and freight transportation effectiveness, which also means digital data will become a key factor in generating logistics activities [4].

Based on the current research landscape, this study suggests prioritizing innovation in data storage and processing technologies while exploring data management approaches better suited to the logistics sector's characteristics. Concurrently, research should investigate how to integrate new technological methods with logistics operations more effectively, enabling data to become a true driving force for logistics enterprise development. Research in these areas will empower logistics companies' information departments to better apply data science in solving transportation and warehouse management challenges, thereby advancing the industry's development and management efficiency.

2 Related Work

Regarding the application of data analysis techniques in logistics transportation, researchers have proposed specific and efficient algorithms and models to address challenges in the traditional logistics industry, thereby enhancing data analysis practices within this sector. This paper will specifically introduce algorithms that help logistics companies improve operational efficiency. The methods involved include, but are not limited to, accessing databases and organizing and analyzing transportation data.

2.1 Transportation scheduling optimization

Renowned engineer He Jie proposed an algorithm capable of effectively evaluating optimal transportation routes. Its calculation formula is:

$$E_i = w_T \times T_i + w_D \times D_i + w_C \times C_i \quad (1)$$

In Formula (1), E_i represents the comprehensive evaluation index for the i transportation route. w_T, w_D and w_C represent the weighting coefficients for time. T_i, D_i, C_i, distance, and cost, respectively; This formula enables rapid calculation of precise transportation costs for each route—data is stored in a structured format via SQL, presenting information more clearly and objectively. Relevant personnel can analyze the transportation

database and obtain the transportation cost index for each route by querying the SQL index [5]. Comparative analysis allows for the intuitive and efficient selection of the optimal transportation route. This method not only accurately reduces transportation costs and improves efficiency but also saves decision-making time, representing a significant optimization for logistics companies' transportation scheduling operations.

2.2 Logistics Management System

Additionally, researcher Duan Kai designed a logistics management system using the Struts2 + Spring + Hibernate (SHH) framework and MySQL database based on the structure of the logistics supply chain. Traditional logistics management systems cover a broad range of aspects and involve relatively complex operations. However, database-based management systems effectively address this issue by dividing complex systems into distinct submodules, each with its own operational methodology. Modules are developed independently. This system processes diverse data types through information technology, meeting the operational requirements of modern physical enterprises [6].

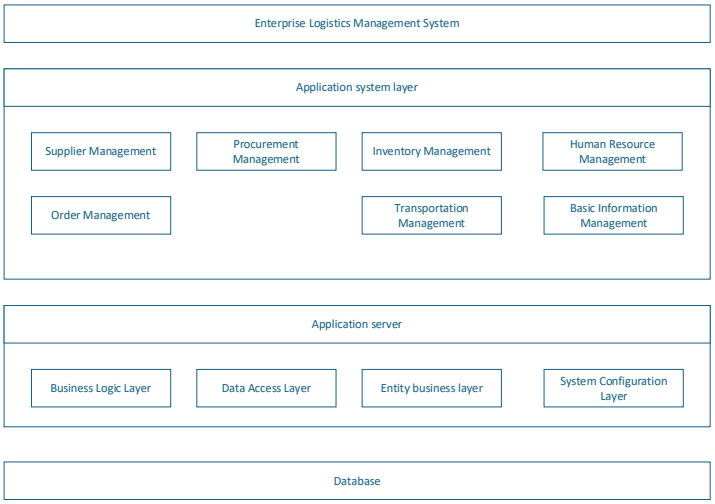

Fig. 1. Caption of the overall structure of logistics management system.

Figure 1 is the overall structure of the logistics management system. The entire system is divided into two major components: the enterprise logistics management system and the database. The enterprise logistics management system consists of the application system layer and the application server. The application system layer primarily manages direct information, while the application server handles deeper-level tasks such as logical processing and data access [7].

Furthermore, the database plays a crucial role within the logistics management system. Therefore, the author constructed an Entity-Relationship (E-R) diagram to organize and analyze the connections among the diverse data within the entire logistics enterprise. By converting different entities, entity attributes, and their relationships into a relational model, data can be organized and summarized in a clear and intuitive manner, significantly enhancing the efficiency of data processing and storage. Below is a data table compiled based on the data within this logistics system.

Table 1. Staff information.

	Data Type	Size	Attribute
Login Name	varchar	20	NOT NULL
Login Pwd	varchar	20	NOT NULL
Real Name	varchar	10	NOT NULL
Gender	varchar	4	NOT NULL
E-mail	varchar	20	NOT NULL
Position	varchar	4	NOT NULL
Entry Time	varchar	4	NULLABLE

Table 1 contains data such as the registrant's name, passport number, real name, gender, and address, with the data type being varchar. The table also includes data size and attributes.

The database-based logistics management model proposed in this study can be practically applied to logistics company operations, enhancing work efficiency, reducing costs, and generating greater profits for the company.

2.3 Neo4j database node and relationship modeling

Another researcher utilized the Neo4j database to store logistics member user information and express delivery order details, employing the Floyd-Warshall algorithm to complete route planning across 31 provincial capital cities nationwide. As a graph database, Neo4j offers several advantages over relational databases. First, graph databases are more efficient. When handling data with numerous relationships and increased depth, they can store data beyond structured formats. In other words, graph databases execute more efficiently, effectively reducing processing overhead. Second, graph databases offer flexibility. Users can adjust data models in real-time as business needs evolve. This means relationships don't need to be fully defined upfront; adjustments can be made later based on requirements. Third, agility. Agile Database's graph icons enable functional and performance testing with each build. The graph also provides a unified database language and operational guidelines, supporting development by diverse application software developers [8].

The construction process is as follows.

2.3.1 Member data structure modeling

The member node store can retain all member information within the software and establish connections with shipping orders. The data modeling is shown in Table 2.

Table 2. Membership Node Data Keywords, Data Types, and Their Meanings

Field	Data Type
vip_name	String
vip_tel	String
vip_sex	String
vip_age	String
vip_password	String
vip_city_address	String
vip_district_address	String
vip_detail_addr	String
vip_balance	Int

Table 2 stores member user information, including member name, phone number, gender, age, passport number, city name, and detailed address. Data types are String or Int.

2.3.2 Express waybill node database modeling

This model stores various details of express waybills and establishes relationships with senders and recipients. The data modeling is shown in Table 3.

Table 3. Key Words, Data Types, and Their Meanings of Express Delivery Order Node Data

Field	Data Type
logistics_id	String
consignor_name	String
consignor_shengaddress	String
consignor_shiname	String
consignor_quname	String
consignor_xiangxiname	String
consignee_name consignee_tel	String String
logistics_id	String
consignor_name	String
consignee_tel	String

Table 3 stores user Identity Document (ID), names, and phone numbers, all of which are stored as String data types.

2.3.3 City node data modeling

City nodes comprise four categories: provincial nodes, provincial capital nodes, and ordinary urban area nodes. This project constructed a total of 31 provincial nodes, 31 provincial capital nodes, and 323 non-provincial capital nodes.

2.3.4 Provincial capital node data modeling

Provincial capital node data modeling serves as the central node for implementing the Floyd-Warshall algorithm. Paths between each provincial capital consist of relationships and attribute values. The data modeling is shown in Table 4.

Table 4. Keywords, Data Types, and Meanings of Road Relationship Data Between Provincial Capital Nodes

Field	Data Type
Distance	Int

As shown in Table 4, use the Neo4J database to model the data listed above. The provincial capital node serves as the central node for implementing the Floyd-Warshall algorithm.

2.3.5 Visualization analysis

After obtaining key attribute relationships, Baidu Maps Application Programming Interface (API) can be invoked to visualize the listed routes, facilitating various user services such as map display, navigation, positioning, and route planning [6].

3 Application & Challenge

3.1 Transportation dispatch optimization algorithm

Regarding the application of transportation dispatch algorithms, this single-route cost comparison method is more focused on small and medium-sized logistics enterprises. Such logistics companies typically specialize in specific routes long-term, leveraging familiarity with route conditions and loading/unloading characteristics. He Jie's route evaluation algorithm optimizes route selection, reducing workload and potential errors from manual analysis. This method comprehensively considers factors to directly identify optimal segments, enabling rational planning of transportation time and routes while minimizing the probability of transfer points or unexpected events.

This algorithm has limitations, however, as it considers only transportation time, distance, and cost. The complete transportation process involves additional critical factors. Cargo characteristics are particularly important; logistics companies often must select routes based on the specific properties of the goods being transported. For instance, transporting heavy or bulky cargo requires choosing specific routes to avoid roads with weight, height, or width restrictions. The scheduling optimization algorithm does not incorporate basic cargo information, potentially causing operational inefficiencies. Additionally, cargo itself presents constraints. For perishable goods, pharmaceuticals, and other items requiring rapid delivery, logistics companies must prioritize fast, stable routes even if they incur higher costs. Consequently, this algorithm lacks flexibility in such scenarios, and a single algorithm cannot sufficiently support route decisions for specialized cargo.

3.2 Logistics management system

Design primarily targets internal operations within logistics companies. It focuses on processing general logistics management tasks and regulating large volumes of logistics information. This model is suitable for medium to large-scale logistics enterprises that require systematic control over comprehensive, long-term processes. This logistics management system divides the overall design into multiple modules, each capable of independent design and development. This feature is highly suitable for large logistics companies involving multiple departments and diverse operational models. By categorizing massive data volumes into patterns and organizing them based on their interrelationships, large logistics companies can operate more efficiently.

The limitation of this model lies in the insufficiently tight connections between modules, preventing cross-module data analysis. Furthermore, MySQL is a relational database that stores data in an ordered format. It lacks robust information retrieval capabilities and struggles to handle unstructured data, making it ineffective for storing the unstructured data common in logistics scenarios [9]. Storing such data requires significant costs for data reprocessing and conversion.

3.3 Neo4j database node and relationship modeling

This model is geared toward handling complex relational data and is currently widely adopted in e-commerce logistics and national third-party logistics enterprises. It provides precise location tracking services, enabling users to conveniently query and monitor shipments. As a graph database, Neo4j balances efficiency and flexibility, allowing easy adaptation to business needs. It effectively resolves the performance issues inherent in traditional relational databases, which suffer from large data volumes and the need for extensive join tables [10].

4 Recommendation

4.1 Transportation scheduling formula

To address the issue of insufficient parameters in the proposed algorithm, additional parameters related to cargo characteristics and their corresponding weights can be incorporated into the formula. For example, new parameters W_i and V_i could be introduced to represent cargo weight and volume, respectively. First, determine whether it falls within the restricted range of the transportation route. After excluding these factors, normal comparison can proceed through the comprehensive evaluation index. Simultaneously, add a new parameter R_i represent the road coefficient. This provides real-time information on whether the road is under maintenance. When the road is impassable, R_i decreases, lowering the overall index, enabling the algorithm to bypass this section.

Regarding the issue of limited application scenarios, this paper proposes integrating this method with other data models. By combining this algorithm with existing diverse data modules, transportation scheduling data can be efficiently organized and analyzed—directly storing different data into modules for comparison and analysis to more rapidly derive optimal transportation solutions.

4.2 Logistics management system

To address MySQL's limitations in handling unstructured data, this paper proposes adopting a hybrid data architecture. This involves integrating existing data models with databases suited for storing unstructured data, such as Humongous Database (MongoDB). Concurrently, a data association mechanism should be established to store index information within the MySQL database. This indexing enables rapid location of corresponding unstructured data in the MongoDB database, facilitating conversion between structured and unstructured data.

Addressing the lack of detail in database table normalization and unclear relationships between tables in the original text, this paper advocates for enhanced database normalization design to clarify inter-table relationships. Primary tables of various relationships are linked via foreign keys established through a common entity attribute, thereby explicitly defining associations between tables. Constraints are also set for key fields: for instance, designating the order number as a primary key ensures uniqueness, while setting the customer ID as a foreign key prevents data redundancy caused by invalid entries.

The original text does not explicitly address data permission management, which may lead to insufficient data security, causing data leaks and potentially exposing the database to attack risks. To enhance security, this paper proposes using a hash algorithm to encrypt and store user information, strengthening data confidentiality.

To address the issue of insufficient integration among existing modules, this paper proposes building a unified data interaction platform to enable real-time data sharing and exchange between modules. Furthermore, leveraging cross-module data analysis models to

integrate data from all modules can enhance overall operational efficiency and prevent inventory backlogs or stockouts caused by conflicting information.

4.3 Neo4j Database Node and Relationship Modeling

The Neo4j database node and relationship modeling data structure shares the same data encryption deficiency as described above. This can be resolved by employing salt values within the hash algorithm, repeating the identical encryption process.

To address the database's limited capacity for processing massive datasets, this paper proposes adopting a distributed deployment approach, specifically the Causal Clustering architecture. This configuration includes one core node and two read-only replica nodes. As data grows, this architecture allows scaling the number of read-only replica nodes, reducing query pressure on staff.

Additionally, Baidu Maps' API functionality can be fully leveraged to convert users' textual addresses into specific numerical values (e.g., latitude and longitude) for storage in the database. Furthermore, real-time traffic APIs can be enabled during delivery to obtain road conditions and avoid congested routes. Such challenges can also be addressed using the optimized transportation scheduling algorithm formula mentioned earlier. By inputting specific parameters, it rapidly calculates precise values, compares time costs across different routes, and optimizes route selection.

5 Conclusion

This paper integrates database technology with the logistics industry and proposes relevant recommendations for the sector. Within logistics operations, there is a need to collect and characterize vast amounts of complex, diverse data. Databases serve as an excellent container, linking disparate data through primary and composite keys to clearly present and analyze large datasets for effective transmission.

Based on the research and analysis of the above scholars, current domestic studies on database applications in logistics primarily focus on optimizing data analysis algorithms and creating data models for logistics company operational workflows. Utilizing database modeling with systems like MySQL and Neo4j has promoted automation and enhanced precision in logistics company operations. Replacing manual tasks with data modeling effectively reduces errors caused by operational mishandling and improves operational efficiency. Simultaneously, database modeling integrates logistics information with big data, enabling users to track order progress in real time and thereby enhancing the user experience to a certain extent.

References

1. L. Min, Big Data - Driven Brand Building and Optimization Innovation in the Logistics Industry: A Case Study of Logistics Distribution Networks. China Brand & Anti - Counterfeiting. **04**, 185 - 187 (2025)
2. J. Gou, Audit Risk Exploration for Logistics Enterprises in the Big Data Era. Chin. Shipp. Wkly. **16**, 63 - 65 (2025)
3. Y. Yao, Research on Information Management Strategies for Logistics in the Big Data Context. China Storage & Transportation. **02**, 113 - 114 (2025)

4. A. G. Fareed, F. Felice, A. Forcina, Role and applications of advanced digital technologies in achieving sustainability in multimodal logistics operations: A systematic literature review. Sustain. Futures. **8**, 100278 (2024)
5. J. He, Chen Mengyang. Research on Factory Production Logistics Optimization Methods Based on Computer Databases. Intell. IoT Technol. **56**, 78 - 81 (2024)
6. K. Duan. Design and Application of a Novel Logistics Management System Based on MySQL Database Technology. Microcomput. Appl. **34**, 126 - 128 (2018)
7. X. Zhang, Q. Zeng, M. Tang, Design and Implementation of Express Logistics System Based on Graph Database and Baidu Maps. J. Beijing Inst. Graph. Commun. **29**, 141 - 149 (2021)
8. F. Li. Application of Hash - Based Salted Encryption Algorithm in GIS. Geophys. Explor. Equip. **33**, 058303 (2023)
9. B. Jose, S. Abraham. Performance analysis of NoSQL and relational databases with MongoDB and MySQL. Mater. Today: Proc. **24**, 2036 - 2043 (2020)
10. A. Perçuku, D. Minkovska, Stoyanova L. Modeling and processing big data of power transmission grid substation using Neo4j. Procedia Comput. Sci. **113**, 9 - 16 (2017)