Architecture Optimization and Practice Exploration of Big Data Analysis Platform for Industry Applications

Yinghua Zhang

Liaoning Communication University, Shenyang, Liaoning, 110000, China

33974932@qq.com

Abstract. When the traditional universal big data platform adapts to vertical industries such as finance, manufacturing and medical care, it faces the adaptation contradiction in real-time requirements, data scale and compliance standards, and needs a lot of secondary development resources. Therefore, this study focuses on balancing the universality and industry specificity of the architecture, and builds a big data analysis platform framework that can dynamically adapt to vertical scenarios. This paper studies the quantification of industry demand through feature matrix and establishes the mapping relationship between demand and architecture. A hierarchical decoupling architecture including industry adaptation layer, hybrid scheduling layer and accelerated computing layer is proposed to support plug-in loading of vertical components. The dynamic resource scheduling algorithm based on reinforcement learning (RL) is developed to improve the resource utilization rate, and the dynamic loading mechanism of industry knowledge graph (KG) is designed to shorten the cold start time of medical model. The experiments were carried out in the fields of finance, manufacturing and medical care. The results show that compared with the benchmark platform, the transaction delay of financial risk control is reduced by 53% and the accuracy of correlation analysis is improved to 99.3%. Manufacturing throughput increased by 59%, and failure prediction F1-score rose to 0.93; The cold start time of medical imaging model was shortened from 120s to 3.8s, and the time spent on GDPR compliance inspection was reduced by 89%, which verified the significant advantages of the optimized architecture in throughput, delay and business indicators. This research provides an architectural design and practical basis for the efficient application of big data analysis platform in vertical industries.

Keywords: Architecture Optimization; Practical Exploration; Industry; Big Data Analysis Platform; Reinforcement Learning; Knowledge Graph.

1. Introduction

The traditional general-purpose big data platform has obvious adaptation defects when facing the industry scene: the manufacturing industry needs to process the microsecond real-time data flow of millions of equipment sensors, while the financial risk control system needs to complete the correlation analysis of billions of transaction data within milliseconds; Medical image analysis should not only meet the storage challenges of PB DICOM data, but also meet the compliance requirements of privacy regulations such as GDPR. This contradiction between "general architecture" and "industry characteristics" leads enterprises to invest a lot of resources in secondary development [1]. In this context, architecture optimization has become the key path to break through the bottleneck of industry application. The landing cost of the industry can be significantly reduced by building an elastic and scalable hierarchical architecture, integrating the knowledge engineering methods of the industry and innovating the mixed load scheduling mechanism.

At present, the technical research shows the differentiation between universality and industry specificity. In terms of technology, the focus is on the innovation of computing paradigm, including batch stream fusion computing through Apache Flink and Spark Streaming to unify the syntax of off-line and real-time analysis, but it does not solve the data problem of specific industries [2]; Snowflake and Alibaba Cloud MaxCompute systems use the storage-computing separation architecture to improve resource utilization, but they encounter new bottlenecks in network I/O; Tools such as TensorFlow on Spark and Ray promote the collaborative training of deep learning and big data, but the integration in some manufacturing models is still not perfect [3-4]. In terms of industry, it pays more attention to vertical scene optimization. For example, Ant Group uses GeaBase for anti-fraud detection, Siemens MindSphere platform supports multiple devices to access, and NVIDIA Clara framework accelerates medical image processing, but each of them faces problems such as rule engine coordination, data compression distortion and encryption overhead [5-6]. Generally speaking, the existing research is limited by the lack of abstract hierarchy of architecture, the lack of industry knowledge fusion and the lack of dynamic adaptation mechanism.

The purpose of this study is to balance the universality and industry specificity of the architecture and build a big data analysis platform framework that can dynamically adapt to vertical scenarios. Quantify industry demand through feature matrix, and establish the mapping relationship between demand and architecture; A hierarchical decoupling architecture including industry adaptation layer, hybrid scheduling layer and accelerated computing layer is proposed to support plug-in loading of vertical components. On the key technology, the dynamic resource scheduling algorithm based on reinforcement learning (RL) is developed to improve the resource utilization rate, and the dynamic loading mechanism of industry knowledge graph (KG) is designed to significantly shorten the cold start time of medical model. Finally, empirical studies were carried out in the fields of finance, manufacturing and medical care to verify the significant advantages of the optimized architecture in throughput, delay and business indicators.

2. Architecture Optimization Method of Big Data Analysis Platform

2.1 Characteristic Matrix Modeling of Industry Demand

In order to realize the effective adaptation of big data analysis architecture in different industries, a modeling method of industry demand characteristic matrix is proposed. Structuring and quantifying the heterogeneous requirements of various industries in data processing, so as to provide scientific basis for the subsequent architecture design [7]. By constructing a quantitative model with multiple characteristic dimensions, industry demand is mapped into computable parameters to guide the dynamic allocation and optimization of system resources.

The modeling method adopts the weighted summation model, and the formula is as follows:

$$R_i = \sum_{k=1}^{n} w_k F_k(i) \tag{1}$$

Among them, R_i represents the comprehensive demand score of the i industry; $F_k(i)$ represents the value of industry i in the k-th characteristic dimension, which has been normalized to the interval of [0,1]; w_k is the weight of each feature dimension, which is scientifically determined by AHP (Analytic Hierarchy Process) to reflect the relative importance of different features to architecture design.

The feature dimension of the model can be extended according to specific application scenarios, such as:

F_1: Real-time requirements (for example, financial risk control is 0.95 and medical imaging is 0.3);

F_2: data scale (for example, manufacturing sensor data is 0.9 and financial transaction data is 0.8);

F_3: compliance level (if the medical industry needs to meet high compliance requirements such as GDPR, the score is 0.99, and the manufacturing industry is 0.4).

By calculating and comparing the R_i values of different industries, we can clearly identify their priorities and resource requirements in architecture design. For example, due to the high real-time demand, the financial industry will drive the system to allocate more streaming computing resources to it; The large-scale data characteristics of manufacturing industry will guide the system to optimize batch processing and edge computing capabilities. Finally, the output of the model will serve as an important basis for the dynamic configuration of the architecture, and improve the adaptability and efficiency of the platform in multi-industry scenarios.

2.2 Hierarchical Decoupling Architecture Design

A hierarchical decoupling architecture design scheme is proposed (Figure 1). The architecture consists of three core modules: industry adaptation layer, hybrid scheduling layer and accelerated computing layer, which undertake the tasks of industry feature support, resource dynamic scheduling and high-performance computing respectively, so as to ensure the universality of the system and realize the deep optimization of vertical scenes.

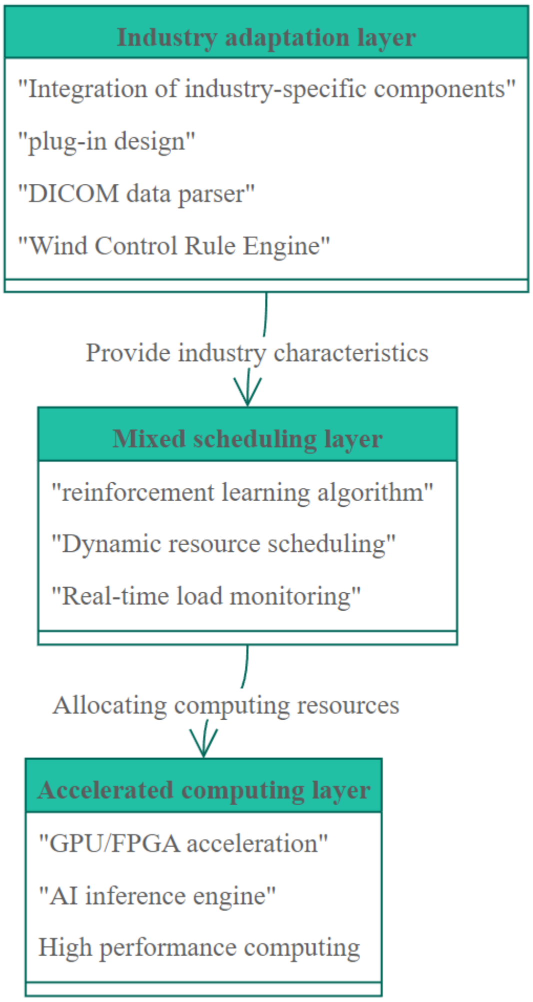

Figure 1. Hierarchical decoupling architecture

The industry adaptation layer is responsible for integrating various industry-specific components, such as DICOM data parser in medical field or risk control rule engine in financial industry, and adopts plug-in design to support rapid deployment and replacement. The hybrid scheduling layer constructs a dynamic resource scheduling mechanism based on RL algorithm, which can intelligently allocate computing resources according to real-time load changes and improve system efficiency. The accelerated computing layer is oriented to high-performance demand scenarios and integrates hardware acceleration modules based on GPU/FPGA, such as AI reasoning engine for medical image analysis, thus significantly improving the processing speed of key tasks.

In order to enhance the flexibility and expansibility of the architecture, the system adopts standardized plug-in interface specifications to ensure the unified access of components from different industries. Plug-in interface definition includes two core methods: initialization and data processing, which is convenient for context parameter injection and industry-specific logic execution. Define industry component interface specifications:

```
public interface IndustryPlugin {
    void init(IndustryContext ctx); // Context injection industry parameters
    DataBlock process(DataBlock input); // Industry specific data processing
}
```

The dynamic loading of plug-ins is realized by OSGi framework or custom class loader, which supports hot plug and online update of components, greatly improving the maintainability and real-time response ability of the system.

2.3 Resource Scheduling Algorithm based on RL

In this study, a dynamic resource scheduling algorithm based on deep Q-network (DQN) is proposed to improve the resource utilization efficiency and service quality of big data analysis platform in multi-industry scenarios [8]. The algorithm models the resource scheduling problem as a RL process, in which the state space is composed of key indicators such as CPU, memory, I/O utilization, task queue backlog and SLA default rate. The action space corresponds to the adjustment of resource allocation ratio, such as the dynamic allocation of stream processing and batch processing resources. Through the continuous exploration and learning of agents in the environment, the system can automatically identify the optimal resource allocation strategy.

In order to guide the optimization direction of the algorithm, a reward function integrating throughput, delay and resource adjustment cost is designed:

$$R = \alpha Throughput\text{-}\beta Latency - \gamma Cost_{overhead} \tag{2}$$

Among them, the weight coefficient α, β, γ of throughput, delay and adjustment overhead can be calibrated according to the SLA requirements of different industries. In the algorithm, the deep neural network is used as the approximation of the Q-value function, the current resource state vector is input, the expected benefits of all possible actions are output, and the action with the highest Q-value is selected as the scheduling decision, thus achieving efficient response and adaptive scheduling for complex dynamic loads.

2.4 Dynamic Loading Mechanism of Industrial KG

In order to improve the reasoning and response ability of big data analysis platform in vertical industries, a dynamic loading mechanism of industrial KG is designed, which is especially optimized for the cold start problem in medical field. By introducing cache acceleration and preheating mechanism, the first loading time of KG is significantly shortened, and the subsequent loading efficiency is further improved. Formula (3) describes the composition of loading time, in which the increase of cache bandwidth greatly reduces the loading delay, thus enhancing the rapid deployment capability of the system in the new scenario.

$$T_{load} = T_{base} + \frac{B_{cache}}{S_{KG}} \delta_{warmup} \tag{3}$$

After receiving the industry query request, the system first extracts key domain terms, then loads related subgraphs from KG, and injects these structured knowledge into the feature engineering module of AI model.

3. Practical Exploration

The experimental environment adopts a cluster architecture with 3 main nodes and 12 working nodes, and each node is equipped with 64-core CPU, 256GB memory and 4 Tesla T4 GPU to support large-scale data analysis tasks. The benchmark platform is built based on Apache Flink 1.14 and Spark 3.2, as a comparison of the unoptimized architecture. The optimization platform integrates the industry hierarchical architecture proposed in this study, including RL-based dynamic resource scheduler and industry KG loading module. The test data covers financial, manufacturing and medical fields, including 1 billion transaction records, time series data collected by 5,000 devices at 1Hz frequency and 100,000 DICOM images, so as to comprehensively evaluate the performance of the platform in different scenarios.

In the experiments of financial, manufacturing and medical scenarios, the optimized platform shows significant performance improvement compared with the benchmark platform based on Flink and Spark. Table 1 shows that the transaction delay of financial risk control is reduced by 53% to 42ms, and the accuracy of correlation analysis is improved to 99.3%; Manufacturing throughput increased by 59% to 1.24M events/s, and failure prediction F1-score rose to 0.93; The model cold start time of medical images was shortened from 120s to 3.8s, and the time spent on GDPR compliance inspection was reduced by 89% to 0.9s/ 10,000, which fully verified the advantages of the optimized architecture in efficiency, accuracy and compliance.

Table 1. Optimization effect of core indicators

Scene	Index	**Benchmark platform**	**Optimization platform**
Financial risk control	Transaction processing delay (P99)	89 ms	**42 ms**
	Accuracy of correlation analysis	97.1%	**99.3%**
Manufacturing industry	Throughput (events/s)	780K	**1.24M**
	Fault prediction F1-score	0.87	**0.93**
Medical imaging	Model cold start time	120 s	**3.8 s**
	GDPR compliance inspection is time consuming	8.2 s/ ten thousand sheets	**0.9 s/ ten thousand sheets**

In the financial risk control scenario, the RL scheduler of the optimized platform can dynamically adjust the resource allocation according to the transaction flow, automatically expand the flow processing resources during the peak period (t=15s), effectively suppress the processing delay from 120ms to less than 50ms, and significantly improve the real-time response capability of the system (Figure 2).

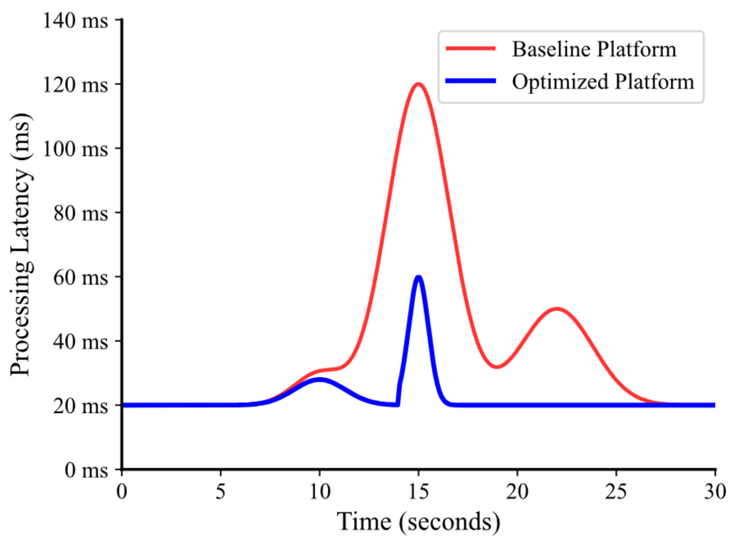

Figure 2. Dynamic scheduling reduces delay

Table 2. Composition comparison of cold start time (unit: s)

Stage	Benchmark platform	Optimization platform
Data loading	68.2	1.1
Characteristic calculation	41.3	1.9
Model initialization	10.5	0.8

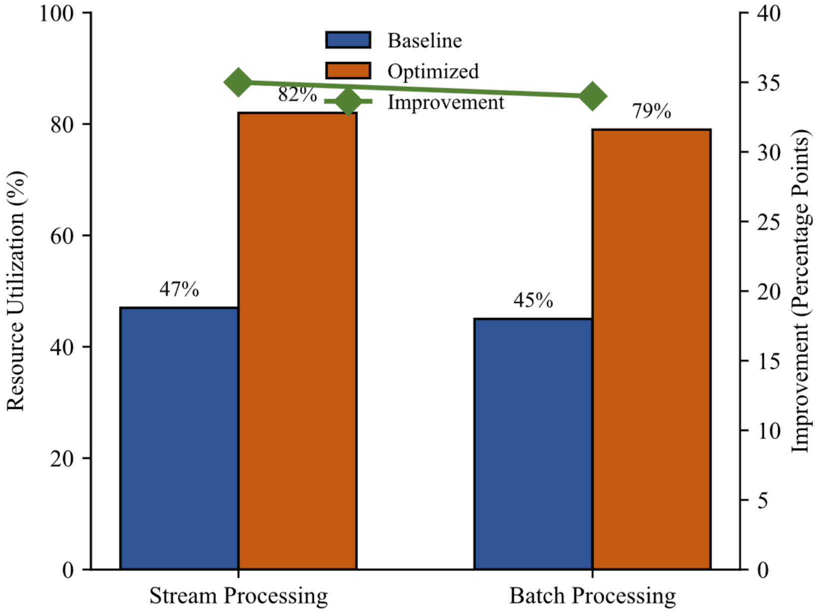

Figure 3. Hybrid load scheduling efficiency

In the medical field, the optimization platform greatly shortens the cold start time of the model through SSD cache and KG subgraph loading mechanism. The data shows that the data loading time is reduced from 68.2s to 1.1s, the feature calculation time is reduced from 41.3s to 1.9s, the injection of KG rules effectively reduces the feature dimension, and the overall cold start time is reduced from 120s to 3.8s, which significantly improves the system startup efficiency (see Table 2).

As shown in fig. 3, the dynamic balance between stream processing and batch processing resources is realized by applying the scheduler based on RL in the manufacturing scene, and the CPU utilization rate is increased from 47% to 82% on average. At the same time, the resource fragmentation rate is reduced by 64% through the container reorganization strategy, which significantly optimizes the resource utilization efficiency under mixed load.

4. Conclusion

By constructing an elastic and extensible hierarchical decoupling architecture, integrating industry knowledge engineering methods and innovating mixed load scheduling mechanism, the landing cost of the industry is significantly reduced. The experimental results show that the optimized architecture shows excellent performance improvement in key areas such as finance, manufacturing and medical care: the transaction delay of financial risk control is reduced by 53%, and the accuracy of correlation analysis is improved to 99.3%; Manufacturing throughput increased by 59%, and failure prediction F1-score rose to 0.93; The cold start time of medical imaging model was shortened from 120s to 3.8s, and the time spent on GDPR compliance inspection decreased by 89%. These results verify the remarkable advantages of the proposed method in efficiency, accuracy and compliance, and provide strong support for the application of big data analysis platform in vertical industries.

References

[1] Private Cloud Platform for Distributed Logging Big Data Based on a Unified Learning Model of Physics and Data.Applied Geophysics,22(2),499-510.

[2] Wang Xiaohui. (2025).AI-Based Big Data Platform for Sports Training Construction and Application. International Journal of Information System Modeling and Design (IJISMD),16(1),1-23.

[3] Changchao Qi, Lingdi Fu, Ming Wen, Hao Qian & Shuai Zhao. (2025). Research on a Simulation Platform for Typical Internal Corrosion Defects in Natural Gas Pipelines Based on Big Data Analysis.Structural Durability & Health Monitoring,19(4),1073-1087.

[4] Anton Ivanov, Abhijeet Ghoshal & Akhil Kumar. (2025). From Stars to Dogs: A Data Analytic Approach to Identifying "Out-Of-Favor" Products on E-Commerce Platforms. Production and Operations Management, 34(6),1346-1366.

[5] Yanzhao Jia. (2025). Application of Big Data Technology in User Behavior Analysis of E-commerce Platforms. Journal of Electronic Research and Application,9(3),104-110.

[6] Yousuf Hemani, Kilian Koch, Oscar Mendo Diaz, Anusch Bachhofner, Simone Baffelli & Davide Bleiner. (2025). The openBIS Digital Platform for Instrumentation and Data Workflow in the Analytical Laboratory. Chimia,79(1-2),36-45.

[7] Shumin Gao. (2025). Human resources big data management platform for listed companies through cloud computing technology. Journal of Computational Methods in Sciences and Engineering,25(3),2384-2395.

[8] Wang Yang, Wang Xuemei & Wang Wanyan. (2025). Hybrid Teaching Reform and Practice in Big Data Collection and Preprocessing Courses Based on the Bosi Smart Learning Platform.Journal of Contemporary Educational Research,9(2),96-100.