International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 5434-5440

https://doi.org/10.70517/ijhsa464382

Research on Distributed Computing Optimization for Real-time Risk Control of Enterprise Financial Big Data

Xin Zheng[1], Lei Zhang[1,*], Chenlu Jia[1] and Hongmei Yue[1]
[1] Department of Management and Media, Shenyang Institute of Science and Technology, Shenyang, Liaoning, 110167, China
Corresponding authors: (e-mail: yekongxingchen001@163.com).

Abstract This study focuses on the problems of poor real-time and limited accuracy of traditional data processing methods in the context of enterprise financial risk management, and builds a set of real-time enterprise financial big data processing and risk management system based on a distributed computing framework. The system adopts Spark as the core distributed computing architecture, and achieves a peak processing capacity of up to 47,500 items/second and an average processing delay of only 267 milliseconds in a 12-node configuration, which greatly improves the efficiency of data processing. The real-time risk management module of the system shows significant advantages in risk identification, prediction, control and feedback, with the risk identification rate reaching 91.2%, the warning accuracy rate reaching 87.4%, and the length of warning advance extending to 18.6 days. By building a multi-level and multi-dimensional risk assessment modeling system and integrating static financial analysis and dynamic transaction monitoring, the system achieves the function of dynamic detection of risk status and accurate early warning. At the level of data security and privacy protection, the system utilizes data watermarking technology in distributed environment, computing task security sandbox, and privacy computing framework based on homomorphic encryption to ensure data security and privacy.

Index Terms distributed computing, enterprise financial big data, risk management, data security, privacy protection

I. Introduction

I. A. *Background of the study*

With the rapid development of the digital economy, the amount of data generated in the business activities of enterprises has been growing, especially the complexity and processing speed of financial data requires more [1]. Traditional financial data processing often faces problems such as poor real-time, low efficiency, and insufficient risk warning, and a more efficient solution is urgently needed [2]. Distributed computing frameworks, especially technologies such as Apache Spark, provide effective tools for processing massive data through parallel computing and fault-tolerant mechanisms [3]. Distributed computing architecture can not only improve the efficiency of data processing, but also solve the bottleneck of big data analysis faced by enterprises. Enterprise financial data includes a variety of structured and unstructured data from transaction details, assets and liabilities to contract documents, etc., and traditional data processing methods can no longer meet real-time requirements [4]. Therefore, how to realize accurate financial risk management while efficiently processing has become a key issue in enterprise transformation.

I. B. *Status of research*

Currently, many studies focus on improving the efficiency of enterprise financial data processing and risk management capabilities based on big data technologies and distributed computing frameworks [5]. Most domestic studies focus on how to combine big data technology and financial risk management to develop real-time risk early warning systems [6]. For example, tools such as Hadoop and Spark are utilized to improve the real-time analysis of financial data through streaming data processing [7], [8]. However, traditional batch processing methods have obvious shortcomings in real-time and accuracy, which are difficult to meet the rapid response needs of modern enterprises. Foreign research focuses on how to improve risk prediction accuracy through data stream processing and machine learning algorithms, especially by introducing deep learning and real-time data stream analysis technology to build a more accurate risk identification system [9], [10]. In addition, data security and privacy protection have become important issues in the application of big data, especially how to ensure data privacy while conducting effective data processing and analysis [11].

Although existing research has made some progress in enhancing financial data processing and risk management, there are still problems such as poor real-time performance and poor risk prediction accuracy [12]. This study proposes a real-time financial data processing and risk management method based on a distributed computing framework, aiming to solve the deficiencies of traditional methods in processing complex financial data and implementing accurate risk management by optimizing the algorithms and technical architecture.

I. C. Research ideas

This research adopts a real-time financial data processing and risk management method based on distributed computing framework, which realizes real-time identification and accurate early warning of corporate financial risks by constructing a multi-level and multi-dimensional risk assessment model, combining static financial analysis and dynamic transaction monitoring. The research adopts Apache Spark as the core computing platform, combined with streaming processing technology to improve data processing speed and real-time performance. In addition, machine learning and deep learning algorithms are used to improve the accuracy of risk identification, while blockchain and homomorphic encryption are utilized to guarantee the security of financial data in terms of data security and privacy protection. Through these innovative approaches, this research seeks to solve the technical bottlenecks in financial data processing and risk management for enterprises, and to provide them with more efficient and secure financial risk control solutions.

II. Research methodology

II. A. Data collection and integration

In order to build a real-time processing and risk management system for enterprise financial big data based on distributed computing framework, an efficient data collection and integration mechanism is first designed. This study adopts multi-source heterogeneous data collection technology and establishes a financial data integration platform through distributed data organizing architecture. The platform can comprehensively collect and organize financial risk-related data from both internal and external sources of the enterprise to ensure the quality and consistency of the data. The data collection process is divided into a batch layer and a velocity layer. The batch layer regularly collects historical data to support risk model training, while the velocity layer is responsible for real-time data collection and processing to ensure that the enterprise can monitor financial risks in a timely manner.

In terms of data collection, a combination of “pull” and “push” strategies is used. Internal systems such as ERP, CRM, etc. push data through Apache Kafka, while external data is collected through targeted crawlers and API interfaces. In the data collation and fusion stage, the on-the-fly data processing architecture built by Spark realizes the standardized transformation of heterogeneous data and ensures the consistency of financial data from different sources and structures. To improve processing speed, the system uses Spark Streaming for micro-batch processing, optimizing the time window to 2 seconds to provide near real-time data processing capabilities.

II. B. Data cleaning and processing

To ensure the quality of financial data, this study designed a multi-stage Spark-based data cleaning and processing process. The data preprocessing stage solves the problem of format inconsistency and redundancy, and the system uses hash value de-duplication technology to efficiently remove duplicate data. In the anomaly detection and processing session, combining statistical and machine learning methods, the system identifies anomalous data from three aspects: time series, numerical distribution and business rules, and corrects them according to the set rules. For missing data, a hierarchical processing strategy is adopted, with different filling methods based on the importance of data and the degree of missingness [13].

Data consistency validation is accomplished through a three-level validation mechanism to ensure that the data conforms to business rules and quality standards. For real-time data, the system adopts an incremental cleaning strategy to process only new or changed data, which significantly reduces the computational burden. These methods ensure the accuracy and reliability of financial data and provide high-quality data support for subsequent real-time risk analysis and decision-making.

II. C. Real-time data analysis and mining

This study designs an efficient real-time data analysis and mining system relying on the Spark distributed computing framework. The system cuts and parallelizes real-time data streams through Spark Streaming's micro-batch processing technology to achieve second-level risk identification and prediction capabilities. The processing of data streams includes real-time financial data acquisition, streaming computation, complex event processing, machine learning modeling and other links, forming a complete data processing chain.

In terms of real-time analysis and mining, the system utilizes rule-based complex event processing and time window technology to construct a risk pattern library, which is capable of identifying abnormal patterns in financial data in real time. The system also combines machine learning algorithms for real-time training and optimization of risk assessment models, improving the accuracy of risk prediction. Risk results are pushed to the decision support system in real time and the risk status is displayed through an interactive dashboard, providing immediate decision support for the enterprise.

II. D. Risk model construction and algorithm implementation

In order to accurately assess financial risk, this study constructs a multi-level and multi-dimensional financial risk assessment model, which contains multiple dimensions such as macroeconomic factors, industry characteristics, enterprise financial status and micro trading behavior. The contribution of each risk indicator to the overall risk is calculated through machine learning algorithms, especially random forest and SHAP value analysis. We designed a comprehensive scoring model to weight each risk indicator to calculate the risk score and provide risk warning through different risk warning mechanisms.

The risk early warning system includes three levels: level 1 based on static risk scoring, level 2 based on dynamic threshold monitoring, and level 3 based on anomalous pattern identification [14]. Through these multi-level early warning mechanisms, timely warning and response to different types of financial risks can be realized.

II. E. Real-time monitoring and feedback

This study designs a comprehensive real-time monitoring and feedback system that incorporates a distributed computing architecture to dynamically monitor corporate financial risks. The system collects and processes financial data in real time, utilizing Spark Streaming for streaming computation and anomaly detection. The risk monitoring engine continuously tracks the changes of financial indicators through state management and realizes dynamic warning by using multi-dimensional warning threshold management module.

The risk event processing unit categorizes and responds according to risk level. The visualization display module uses WebSocket technology to realize real-time data push, and displays the risk status in the form of multiple charts. Intelligent decision support module combines historical experience and rule engine to provide targeted response strategies for enterprises. In addition, the system is designed with a risk feedback assessment model that can assess the accuracy, recall rate and lead time of the warning, and continuously improve the accuracy of risk monitoring and warning.

III. Findings and analysis

III. A. System performance testing

This study builds a set of enterprise financial big data real-time processing system based on distributed computing framework, and verifies its processing capability and stability through comprehensive performance tests. The test environment consists of 12 computing nodes, each node is equipped with Intel Xeon E5-2680 v4 processor, 128GB of memory and 10Gbps network connection device, running CentOS 7.9 operating system, Apache Spark 3.3.0 as the computing architecture, using HDFS 3.3.1 to implement data storage. Kafka 3.2.0 is used as the message queuing facility. The testing methodology utilizes an incremental load strategy, starting with a single-node baseline performance to build a baseline, progressing to horizontal scaling tests (incrementally increasing to 2/4/8/12 nodes), then advancing to load incremental tests (increasing from 5000 to 50,000 items/sec), and finally conducting a 72-hour medium load (15,000 items/sec) stability test. With the help of Spark Web UI, Ganglia cluster monitoring and other tools, we captured the performance parameters of the system at all levels, including processing speed, latency, throughput, resource utilization, risk identification accuracy, and scaling efficiency. Table 1 presents the results of the system performance tests.

The test results show that the system performance is excellent, under the configuration of 12 nodes, the peak processing capacity reaches 47,500 items/second, which is 14.8 times higher than the traditional batch system; the average processing delay is only 267 milliseconds, which is much lower than the 15-second delay of the traditional system. The accuracy of risk identification reaches 92.7%, which is 18.5 percentage points higher than the traditional method. Different data processing modules show different performances, the data collection and integration module has a higher throughput capacity, however, it is constrained by network I/O limitations, and the data cleaning and processing module shows a positive relationship with the number of CPU cores. The real-time analysis and mining module relies on computing resources and the size of memory capacity, and the risk model calculation module is characterized by excellent parallelism. The system shows excellent resilience in response to unexpected traffic conditions, and can automatically schedule resources to control the growth of latency. The processing efficiency of different types of financial data varies, with structured data being the fastest, semi-structured data the second fastest, and unstructured text being relatively inefficient.

Table 1: System performance test results

Index	Traditional system	Ours system (4 Node)	Ours system (8 Node)	Ours system (12 Node)
Data processing speed	3200	18600	32400	47500
End-to-end processing delay (ms)	15000	485	342	267
System throughput (MB/s)	42	245	427	625
Accuracy rate of risk identification (%)	74.20	89.50	91.80	92.70
False alarm rate (%)	18.30	9.20	7.50	6.80
Underreporting rate (%)	12.50	5.80	4.20	3.50
Resource utilization rate (%)	45.20	72.60	76.80	78.30
Expansion efficiency (%)	-	92.50	88.30	85.70

In addition, the results of the performance test verified that the system is outstanding in scalability, and the processing capacity shows a nearly linear growth trend with the increase of the number of nodes, and the scaling efficiency is always maintained at more than 85% when scaling up from 2 nodes to 12 nodes, which indicates that the system has an excellent level of scaling capability. The system scalability test also shows that the expansion efficiency is maintained above 85%, which confirms that the design of the system architecture is reasonable and effective. During the long-time stability test, the system's performance indexes fluctuated by no more than 5% during 72 hours of continuous operation, with no service interruption or data loss, which demonstrates that the system runs stably and reliably. Compared with the traditional batch processing method, the system has significant advantages in terms of risk management effectiveness, with the risk warning time shortened from an average of 24 hours to less than 5 minutes, the accuracy of risk identification greatly improved, the false alarm rate reduced from 18.3% to 6.8%, and the omission rate reduced from 12.5% to 3.5%. The system can meet the requirements of modern enterprise financial risk management for high real-time, high reliability and high accuracy standards, providing a solid technical support for enterprises to build a real-time risk prevention and control system, and the experimental results have fully proved its comprehensive advantages in processing speed, latency control, accuracy and scalability and other aspects.

III. B. Risk management effectiveness assessment

The real-time risk management module, as the core component of this study, is directly related to the improvement of enterprise financial risk prevention and control capability. In order to comprehensively assess the actual status of enterprise financial risk management system based on distributed computing architecture, this paper plans a set of systematic evaluation methods, builds up multi-dimensional evaluation indexes, and carries out an in-depth evaluation of the effectiveness of the system by means of comparative experiments and case analysis.

The evaluation was conducted in three enterprises in different industries, namely, manufacturing enterprise A (annual revenue of about 5 billion yuan), retailing enterprise B (annual revenue of about 3.5 billion yuan), and financial services enterprise C (asset size of about 12 billion yuan), and the evaluation period was set at 6 months, so as to make an objective judgment on the value of the system through the comparison of the differences in the effectiveness of the risk management system before and after the implementation. The evaluation was conducted through a comprehensive approach of “comparison test + case analysis + user feedback”. In the comparison test, the financial risk management of each enterprise is divided into an experimental group and a control group, with the experimental group utilizing the system developed in this study and the control group adopting the traditional approach. Case studies are conducted to analyze the performance of the system in the process of risk lifecycle management by selecting typical risk events. User feedback is collected from enterprise risk managers and decision makers through questionnaires and interviews. The evaluation index system is built on four levels: risk identification capability, prediction accuracy, control efficiency and feedback mechanism. Data collection relies on risk event log records, automatic statistics and manual review to ensure accuracy and comprehensiveness. The assessment of risk management effectiveness is shown in Table 2.

The results show that the system is significantly better than the traditional approach in all indicators. The risk identification ratio reached 91.2%, an improvement of 33.1%, the identification speed was shortened from 26.4 hours to 0.3 hours, and the classification accuracy rate increased by 24.1%. The early warning accuracy rate reached 87.4%, with an improvement of 32.8%, the length of early warning advance was extended from 5.2 days to 18.6 days, and the false alarm rate was reduced from 21.5% to 8.3%. The length of risk disposal was shortened from 32.6 hours to 9.4 hours, the loss reduction ratio was increased from 25.3% to 62.8%, and the control cost was reduced by 1.39 million yuan annually. The timeliness of system feedback was improved from 48.2 hours to 3.5 hours, and the scores of feedback completeness and continuous optimization ability were improved by 46.8% and 50% respectively. Case analysis confirms the significant application value of the system. The outstanding performance of the system in real-time and accuracy enables the enterprise to detect risks earlier, predict the development trend more accurately, deal with risk events more efficiently, and achieve the transformation from passive response to active management, which is of great significance to the sound development of the enterprise in the complex economic environment.

Table 2: Assessment of risk management effect

Dimension	Index	Traditional method	This system	Increase amplitude
Risk identification ability	Risk identification rate/%	68.5	91.2	33.1%
	Risk identification speed/h	26.4	0.3	98.9%
	Risk classification accuracy rate/%	72.3	89.7	24.1%
Accuracy of risk prediction	Early warning accuracy rate/%	65.8	87.4	32.8%
	Early warning time/day	5.2	18.6	257.7%
	False alarm rate/%	21.5	8.3	61.4%
Risk control efficiency	Risk disposal time/h	32.6	9.4	71.2%
	Risk loss reduction rate/%	25.3	62.8	184.2%
	Risk control cost/10000 ¥	385	246	36.1%
Risk feedback mechanism	Feedback timeliness/h	48.2	3.5	92.7%
	Feedback completeness score	6.2	9.1	46.8%
	Continuous optimization ability score	5.8	8.7	50.0%

III. C. Analysis of the effect of data security and privacy protection

While providing efficient risk management, the enterprise finance big data real-time processing system must guarantee data security and privacy protection, which is extremely critical to the value of the system. This section analyzes the effectiveness of the enterprise financial data security and privacy protection system built on distributed computing architecture. We have designed and implemented a multi-level deep defense architecture to ensure the security of the whole life cycle of data through the synergy of technical and management measures, and verified the effectiveness of these measures through empirical research. Table 3 shows the evaluation of the effectiveness of data security and privacy protection measures.

A machine-learning-based anomalous access detection system has been set up at the monitoring level, capable of recognizing 95.7% of anomalous access actions, with an average detection time of only 1.2 seconds. Destruction control ensures that data is safely erased at the end of its lifecycle, which is in line with data compliance standards. Evaluation results show that the system's security protection ability score of nine points two (out of ten), compared with the traditional approach to improve 46.0 percent, the accuracy of abnormal access to detect the degree of 95.7 percent, an increase of 32.0 percent, anomalous detection of the response time from thirty-five point six seconds to one point two seconds, an increase of 96.6 percent. What is especially critical is that while the system provides advanced security, the loss of encrypted data processing performance is only 24.5 percent, which is a significant reduction compared to the 68.3 percent of the traditional approach, achieving a good balance between security and performance. In terms of compliance, the system covers the requirements of the Network Security Law, the Data Security Law, the Personal Information Protection Law, and other regulations, with a compliance coverage ratio of 98.4 percent, ensuring that the processing of corporate financial data is reasonable and legal.

In the actual application, we conducted a six-month tracking and evaluation of the security effectiveness of three enterprises. In the application example of manufacturing enterprise A, the system successfully resisted an Advanced Persistent Threat (APT) attack on the financial database, in which the attacker attempted to obtain core financial data through the loopholes of distributed computing nodes, and the system detected and blocked the attacker in time through anomalous behavior detection, thus protecting intellectual property rights and trade secrets valued at about 2.3 billion yuan. In the application example of retail enterprise B, the system's data watermarking technology successfully traced back an internal data leakage incident involving customer payment information, promptly reducing losses and improving internal control processes. In the case of financial services company C, a homomorphic encryption-based privacy computing framework enabled it to share risk assessment models with partners while protecting customer privacy, expanding the value of data and meeting regulatory compliance requirements.

Table 3: The effectiveness of data security and privacy protection measures

Security dimension	Traditional method	This system	Increase amplitude
Data Breach Prevention Capability Score	6.3	9.2	46.0%
Abnormal access detection accuracy rate/%	72.5	95.7	32.0%
Abnormal detection response time/s	35.6	1.2	96.6%
Performance loss of encrypted data processing/%	68.3	24.5	64.1%
Security compliance coverage rate/%	75.2	98.4	30.9%
Privacy Protection Technology Maturity Score	5.8	8.9	53.4%

IV. Conclusion and outlook

IV. A. Conclusions of the study

This study proposes and implements a real-time processing and risk management method for enterprise financial big data based on distributed computing framework. By building a distributed computing architecture with Spark as the core, this study achieves a peak processing capacity of 47,500 items/second, which significantly improves the real-time financial data processing. The real-time risk management module of the system demonstrates excellent performance, with a risk identification accuracy of 91.2% and an early warning accuracy of 87.4%, and is able to issue an early warning 18.6 days in advance, which greatly improves the risk identification and response capability. By introducing data watermarking, homomorphic encryption and other technologies, the system effectively guarantees the security and privacy of financial data, while ensuring a balance between computing efficiency and data protection. Overall, the system based on the distributed computing framework can significantly improve the timeliness and accuracy of enterprise financial data processing, thus effectively improving the efficiency and ability of financial risk management and helping enterprises realize the transformation from passive response to active prevention.

IV. B. Research limitations and future prospects

Although this study has made significant progress in real-time financial big data processing and risk management, there are still some limitations. The system has some bottlenecks in throughput when facing extreme high concurrency scenarios, especially during the financial quarter-end settlement period. The adaptability of the risk model in the face of new types of risk patterns still needs to be strengthened, and the system's ability to process unstructured data is relatively limited, failing to dig deeper into the potential risk signals in the text of financial statements, audit reports and other texts. In addition, the homomorphic encryption technique for privacy protection still faces the problem of high computational overhead, which makes it difficult to support the real-time computation of complex financial models.

Future research will focus on the following aspects: first, optimize the throughput and stability of the system by combining streaming and batching integrated processing architecture; second, improve the adaptive capability of risk models, and use deep reinforcement learning to improve the identification of new risk patterns; further strengthen the processing capability of unstructured financial data, and improve the semantic comprehension capability by combining with natural language processing technology; and lastly, explore light-weight privacy Finally, we will explore lightweight privacy protection techniques to reduce computational overhead and improve model efficiency. The research will also develop more cost-effective solutions for small and medium-sized enterprises (SMEs) to promote the popularization of this technology.

References

[1] Ha, L. T. (2023). Digital business and economic complexity. Journal of Computer Information Systems, 63(1), 162-175.

[2] Liang, Y., Quan, D., Wang, F., Jia, X., Li, M., & Li, T. (2020). Financial big data analysis and early warning platform: a case study. IEEE Access, 8, 36515-36526.

[3] Khan, M. A. A., Bhushan, C., Ravi, V., Rao, V. S., & Orsu, S. S. (2022). Nowcasting the financial time series with streaming data analytics under apache spark. arXiv preprint arXiv:2202.11820.

[4] Fernández-Gómez, A. M., Gutiérrez-Avilés, D., Troncoso, A., & Martínez-Álvarez, F. (2023). A new Apache Spark-based framework for big data streaming forecasting in IoT networks. The Journal of Supercomputing, 79(10), 11078-11100.

[5] Ren, S. (2022). Optimization of Enterprise Financial Management and Decision-Making Systems Based on Big Data. Journal of Mathematics, 2022(1), 1708506.

[6] Jiang, L. (2021, July). Early Warning Mechanism and Countermeasures of Enterprise Financial Risk in the Era of Big Data. In 2021 7th Annual International Conference on Network and Information Systems for Computers (ICNISC) (pp. 652-654). IEEE.

[7] Chang, Y., & Wang, J. (2022, December). Research on Optimization of Enterprise Financial Management System Based on Big Data Hadoop. In 2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS) (pp. 1456-1461). IEEE.

[8] Boyanov, L. K. (2021). Financial data processing in big data platforms. Economic Alternatives, 27(4), 534-546.

[9] Bello, A., & Ibrahim, F. (2024). Real-time Machine Learning: Algorithms and Applications in Stream Processing. Journal of Innovative Technologies, 7(1), 1-9.

[10] Bousbaa, Z., Sanchez-Medina, J., & Bencharef, O. (2023). Financial time series forecasting: a data stream mining-based system. Electronics, 12(9), 2039.

[11] Salami, I. A., Adesokan-Imran, T. O., Tiwo, O. J., Metibemu, O. C., Olutimehin, A. T., & Olaniyi, O. O. (2025). Addressing Bias and Data Privacy Concerns in AI-Driven Credit Scoring Systems Through Cybersecurity Risk Assessment. Asian Journal of Research in Computer Science, 18(4), 59-82.

[12] Hong, S., Wu, H., Xu, X., & Xiong, W. (2022). Early warning of enterprise financial risk based on decision tree algorithm. Computational Intelligence and Neuroscience, 2022(1), 9182099.

[13] Al-Okaily, M., & Al-Okaily, A. (2024). Financial data modeling: an analysis of factors influencing big data analytics-driven financial decision quality. Journal of Modelling in Management.

[14] Koyuncugil, A. S., & Ozgulbas, N. (2012). Financial early warning system model and data mining application for risk detection. Expert systems with Applications, 39(6), 6238-6253.