IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 9, 2020, accepted January 18, 2020, date of publication January 23, 2020, date of current version March 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2969039

Financial Big Data Analysis and Early Warning Platform: A Case Study

YI LIANG[1,2], DAIYONG QUAN[3], FANG WANG[2,4], XIAOJUN JIA[1,5], MENGGANG LI[1,4,5], AND TING LI[2]

[1]National Academy of Economic Security, Beijing Jiaotong University, Beijing 100044, China
[2]School of Economics and Management, Beijing Jiaotong University, Beijing 100044, China
[3]Postdoctoral Programme of China Centre for Industrial Security Research, Beijing Jiaotong University, Beijing 100044, China
[4]Beijing Laboratory of National Economic Security Early-warning Engineering, Beijing Jiaotong University, Beijing 100044, China
[5]Beijing Center for Industrial Security and Development Research, Beijing Jiaotong University, Beijing 100044, China

Corresponding authors: Xiaojun Jia (xjjia@bjtu.edu.cn) and Menggang Li (mgli1@bjtu.edu.cn)

This work was supported in part by the Program of the Co-Construction with Beijing Municipal Commission of Education of China under Grant B18H100040 and Grant B19H100010, and in part by the Key Project of Beijing Social Science Foundation Research Base under Grant 19JDYJA001.

ABSTRACT In order to keep the bottom line of systemic financial risks and prevent the mitigation of major risks, this work focuses on the investigation of multi-source heterogeneous data fusion algorithms and cleaning technologies to establish a suitable style for data analysis and big data computation frame. In this paper, according to the above method, we provide the basis for early analysis of economic security. Utilizing the big data analysis, an emerging information technology method, we can be able to explore new risk early-warning methods, build a risk monitoring and early-warning platform and achieve scientific economic decision-making, so that the sources of economic risk in national economic security can be traced.

INDEX TERMS Big data, pre-warning, economic security, early-warning methods.

I. INTRODUCTION

The emergence of advanced data processing, storage, and computing technologies such as big data, cloud computing, and artificial intelligence (deep learning, knowledge atlas) has provided a way for national economic security risk early warning. Therefore, we carry out the prediction and early warning of economic security risks [1]–[3] research. In addition, we systematically, comprehensively and objectively assess the overall risk level of the national economy. This provides predictive and early warning for scientific decision-making, which not only helps prevent systemic risks, but is also an inherent requirement for China to maintain the soundness of its economic system, maintain economic and social stability, and promote national security.

Internet+ finance has brought explosive growth in economic data. Traditional mathematical statistics and fitting methods are difficult to meet the deep mining of massive data. Big data and machine learning methods are becoming standard on many economic Internet platforms. Therefore, how to quantify economic risks with the help of emerging information technology methods has become the primary issue of early warning. We focus on the fusion algorithms and cleaning techniques of multi-source heterogeneous data to form a style suitable for data analysis. Also, we study how to build a big data computing framework to provide a basis for early analysis of economic security.

Based on analyzing and summarizing the existing deficiencies, we build a big data processing platform and research how to collect real-time and high-speed processing of full-dimensional economic data to realize the early warning of national economic security risks. Our specific research result is the creative construction of a big data processing platform.

II. RESEARCH BACKGROUND

There are many works of literature on national economic security [4]–[7], but the work of early warning and simulation of economic risks based on big data has just begun. Most of the existing researches is qualitative risk analysis, and quantitative risk measurement research is extremely scarce. Specific work guidance for national economic security requires a set of systems that can be visualized, evaluated, and adjusted as a work support. Otherwise, blindly pursuing security will often become solidified or rigid.

The associate editor coordinating the review of this manuscript and approving it for publication was Sabah Mohammed.

The establishment of a complete assessment and early warning system can effectively guide and assist the specific work of China's national economic security construction. We use emerging information technology methods such as big data analysis and artificial intelligence to provide quantitative and visual history, status quo, and predictive analysis for various non-traditional security fields. In addition, the system can provide scientific decision-making support services to decision-makers at all levels.

III. MAIN METHOD

This section introduces the system's overall architecture, technical route, stream processing framework, and other aspects, as well as security early warning principles.

A. OVERALL ARCHITECTURE DESIGN

The advent of the era of big data has made massive data more evidence-based for macroeconomic analysis. The real-time processing of economic big data and the intelligence of data analysis methods bring many advantages to macroeconomic forecasting and early warning. It can overcome the shortcomings of traditional methods and improve the accuracy and timeliness of economic analysis.

Economic big data processing involves cross-disciplines such as cloud computing, artificial intelligence and machine learning, macroeconomic forecasting, and microeconomic analysis. To achieve the goal of collection and analysis of economic and security big data, on the one hand, we will research on economic and security big data collection, and focus on the research of multi-source heterogeneous data fusion algorithms and cleaning techniques to form a style suitable for data analysis. This provides the basis for early analysis of economic security. On the other hand, we will study the construction of different big data computing frameworks, real-time risk early-warning analysis algorithms for the real-time and delay needs of big data early-warning analysis of economic security, and deeply explore the relationship between different industries and regional economies. Also, the system can reveal the risks involved.

The data source in Figure 1 includes data collection, which mainly collects data at a deep level and realizes the convergence of different protocols and data at the grassroots level. It mainly relies on two aspects of capabilities, one is to rely on sensors, financial supervision systems, economic management environment-oriented, trust systems, and other elements of data for real-time collection. This enables the platform to directly integrate the underlying data with the help of big data platforms, large data storage centers, embedded software, and other infrastructure and connection technologies. The second is the use of new edge computing devices represented by smart gateways to achieve the aggregation of smart sensor and system data and the indirect integration of edge analysis results into cloud platforms. Various types of edge connection methods provide powerful help for the ubiquitous connection of the economic security risk early warning platform. Expand the data sources that the economic big data platform can collect and analyze.

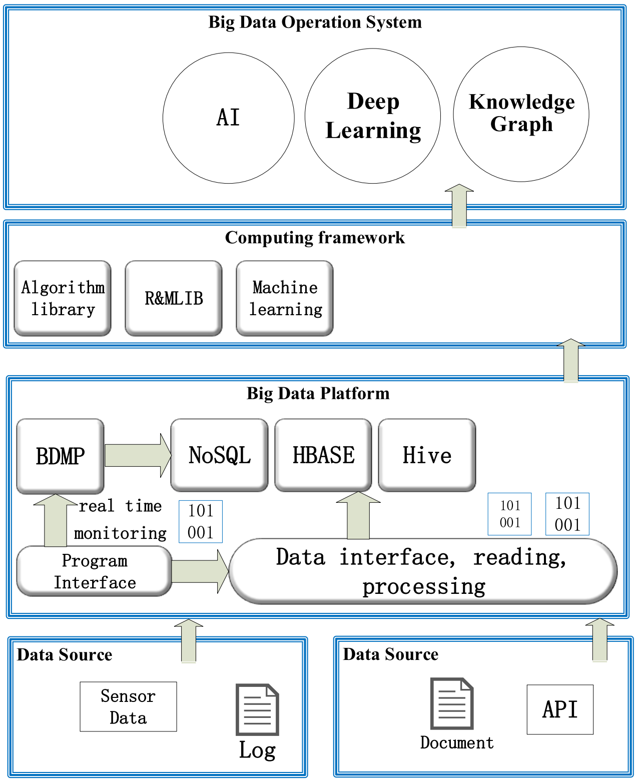

FIGURE 1. Overall architecture diagram.

The management service platform layer is the core. It is based on the PaaS architecture and integrates financial microservices, big data services, and application development. It is comparable to a mobile Internet operating system. First, the platform combines cloud computing, big data technology, and economic management experience and knowledge to form a basic analysis capability for economic data. The platform solidifies technology, knowledge, experience and other resources, and solidifies these resources into professional software libraries, application model libraries, expert knowledge bases and other portable and reusable development tools and microservices. Second, the platform provides a complete economic data service chain consisting of data storage, data sharing, data analysis, and supervision models. The platform brings together various traditional professional processing methods and cutting-edge intelligent analysis tools. It helps users to quickly and easily realize the integrated management and value mining of economic data. Third, it builds an application development environment based on economic data services. The platform provides various management microservices that contain financial knowledge and industry supervision experience, financial application [8] development tools, and comprehensive management methods for application development operations and maintenance. This helps users quickly build customized smart application apps and form business value.

The application service layer is key. Deploying applications based on the mixed industry regulatory environment and facing the various scenarios of regulation, this is the final output of the economic big data platform services. These typical application scenarios, such as intelligent supervision, networked collaboration, personalized customization, and service extension, provide users with various intelligent service solutions customized and developed in the platform.

B. DETAILED TECHNICAL ROUTE

The application of big data technology in finance has further promoted financial service model innovation, development model transformation, management innovation, and product innovation. The application of big data technology has provided effective support for financial innovation. At the same time, it also provides accurate information and knowledge for financial services and provides personalized services for consumers.

Compared with other industries, big data has more potential value for the financial industry. McKinsey research shows that the financial industry ranks first in the Big Data Value Potential Index.

''Big data'' undoubtedly has a wide range of applications in the financial field. Boston Consulting found 64 potential applications in seven major areas of banking. These applications are found in retail, corporate, and capital markets businesses. In addition, they include transaction banking, asset management, wealth management, and risk management.

From the overall situation of the application of ''big data'' by overseas financial institutions, one-half of the financial applications are in the stage of popularizing and understanding the concept of big data, one-third of the financial applications are in the pilot phase, and about one half in five financial institutions is already familiar with the application of ''big data''. It is stepping up its capabilities step by step and embedding the working mechanisms required by ''Big Data'' into business models and operating models. These applications have entered the stage of embedded transformation.

The application of big data technology can grasp the effective information of customers and business in many aspects. They can comprehensively analyze customers' assets and liabilities, liquidity and customer behavior. This helps financial institutions to carry out product innovation, precise marketing, and risk management, and to transform data assets into strategic assets and market competitiveness. This can enable big data technology to play a role in securities and futures, banking, insurance, and emerging Internet finance. These project construction schemes are divided into four areas: equipment domain, platform domain, application domain, and security domain.

1) DEVICE AREA

(1) The device domain implements remote access and connection management, connection monitoring, and configuration updates for IoT terminals. In addition, it also has software and system upgrades, troubleshooting, and lifecycle management functions. The device domain provides real-time device status monitoring and application status alarms and opens API call interfaces. It enables users to easily perform system integration and value-added function development. Device data for the device domain is stored directly in the cloud.

(2) Connection management technology for massive physical devices in the Internet of Things

Device domains enable massive, heterogeneous physical devices to orderly and seamlessly access the Internet of Things. This achieves wide-area interconnection and intelligent management of physical equipment. The device domain establishes global-based load balancing. It supports the nearby access of equipment and can meet the horizontal expansion of the platform. For the ubiquitous access cluster, according to the protocol type, the server load of the device domain can be scheduled in real-time to achieve the access of devices with different protocols and different loads.

(3) Equipment intelligent sensing monitoring and equipment interconnection

Based on the OID (Object Identifier) coding, the interconnection of different equipment and personnel is implemented to achieve the interconnection and interconnection of IoT terminals. We study the problems of interconnection and data collection of big data and multi-source heterogeneous devices. We study the technical characteristics of different types of interfaces and protocols. Finally, a unified data acquisition model and data transmission standard were developed to implement a method capable of accessing multiple signal sources of different standard protocols.

(4) Equipment information collection and processing

a) The device domain uses physical devices such as low-level sensors, electronic tags, and routers to collect and aggregate status data of target objects in real-time. It processes the collected high-dimensional time series data. The system analyzes and mines the standardized representation and organization sequence data to achieve the purpose of mining the hidden behavioral patterns of time series. At the same time, the system also provides users with a visual data query and sharing interface.

b) Heterogeneous data collection

The lack of a unified standard for the data format has brought severe challenges to the later database storage, as well as data cleaning and data integration and analysis. This system is ultimately based on a unified data collection model and needs to collect different types of equipment data, database data, and file data. The device domain provides an open interface for all types of structured and unstructured data through a unified data transmission standard, and finally achieves a unified and scalable heterogeneous data collection.

c) Data cleaning

An important feature of big data is the low density of value. Its data quality is also different from the quality of traditional data. There may be a lot of bad data, as well as erroneous data and missing data, but the quality of data has a huge impact on the effectiveness of decision-making based on big data mining. Therefore, the cleaning of data and the identification of invalid and erroneous data have become key prerequisites that affect the effectiveness of big data mining. This system needs to develop targeted data cleaning and screening models and algorithms based on the characteristics of data in different industries.

2) PLATFORM AREA

(1) Load balancing technology and stream processing system for real-time cloud processing

For real-time flow calculations for large flows, load balancing for data flow services is an important part of this system's research. In the real-time streaming computing environment, the data is not only increasing in size but also business diversity is becoming more and more complicated. The existing load balancing technology is facing the challenge of scalability bottlenecks and cannot meet the real-time reliability requirements of high-speed network stream processing.

This project studies the load balancing architecture and algorithms for real-time cloud computing platforms and proposes a high-availability real-time cloud computing system based on this. The system has built-in task-level and traffic-level load balancing mechanisms. As shown in Figure 2, for task-level load balancing, the system uses a real-time task scheduling model with low complexity. Based on this model, the system provides task scheduling and resource allocation algorithms that meet the dynamic scenarios of a streaming computing environment. It reduces the data traffic between nodes while the tasks are dynamically balanced. For traffic-level load balancing, the system uses a load balancing algorithm that maintains session consistency without a global session table. The system realizes the balance among the data flow of cloud tasks and improves the system availability and balance under the environment of massively connected data flows. This enables the system to analyze complex business data.

(2) Real-time data storage and query

Real-time data storage and querying use a shared-nothing architecture. This architecture supports a real-time data storage and query engine that searches billions of rows in seconds by supporting an advanced index structure, and the engine supports horizontal scaling. A cluster consists of different types of nodes, and each node performs a specific function.

The composition of the cluster is shown in Figure 3. Real-time nodes are responsible for data injection, storage, and response to recent event queries. Similarly, the historical compute node is responsible for loading and responding to queries for historical events. The data is stored in a storage node. The storage node can be a historical compute node or a real-time node. A query first visits the broker node, which is responsible for discovering and routing the query to various storage nodes containing relevant data, and then the storage nodes execute the parts of their query in parallel and return the results to the broker node. The broker node receives these results and merges them. Finally, the combined final result is returned to the requester of the query. The brokers node, compute node, and real-time node are all considered queryable nodes. There is a set of coordination nodes to manage load distribution and replication. The coordination node is a non-queryable node. It is mainly used to maintain the stability of the cluster. The Coordination node needs to rely on an external MySQL database. It requires Apache Zookeeper to perform cluster collaboration. Although queries are forwarded through Hadoop, intra-cluster communication is through Zookeeper.

3) APPLICATION DOMAIN

a: APPLICATION FUNCTION

AEP (Application Enabled Platform) provides two major functions: application development and unified data storage. It provides application development tools, middleware, data storage functions, and business logic engines. AEP can connect to third-party system APIs. Enterprises and individual users can quickly develop, deploy, and manage applications on AEP. AEP does not need to consider issues such as lower infrastructure expansion, data management, and aggregation, communication protocols, and communication security, thereby reducing development costs and development time.

The platform provides industry Saas services. This service provides professional, rich and comprehensive service functions for different users in vertical industries. Users use smart terminals to customize required services at the application layer and perform key technology verification required by the enterprise.

b: DATA OPENING AND INTELLIGENT DECISION-MAKING

The business analysis subsystem provides secure, stable, real-time, and practical technical big data analysis services. It can not only perform classification processing on the basis of various related data but also analyze and provide visual data analysis results. The business analysis subsystem uses real-time dynamic analysis to monitor the stock situation and provide early warning. It helps users perform multi-dimensional data monitoring and services in areas such as operations. Intelligent decision-making refers to the process of using uniformly expressed user intents to match in knowledge aggregation to solve problems and give an ordered set of recommended solutions.

The business analysis subsystem uses a matching method based on the search problem framework and defines some common search problem types. This is called the search problem framework. The subsystem builds a search intent decomposition fusion rule knowledge base based on the problem framework. It introduces decomposition rules in the decomposition of search intent. The subsystem structured the complex search intent into a logical combination of interacting atomic search problems that can be matched to the search problem framework to solve the atomic search task, and reversely fuse search results according to the decomposition relationship.

The granularity of the atomic search task is adaptively abstracted, reified, and synonymously restructured according

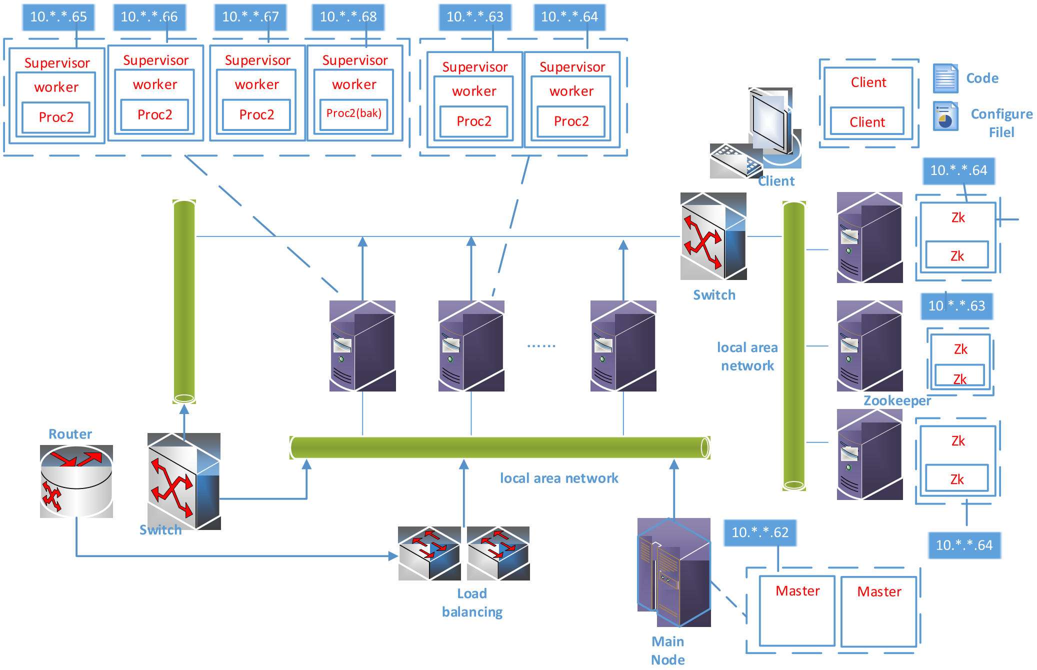

FIGURE 2. Schematic diagram of load balancing and flow processing system.

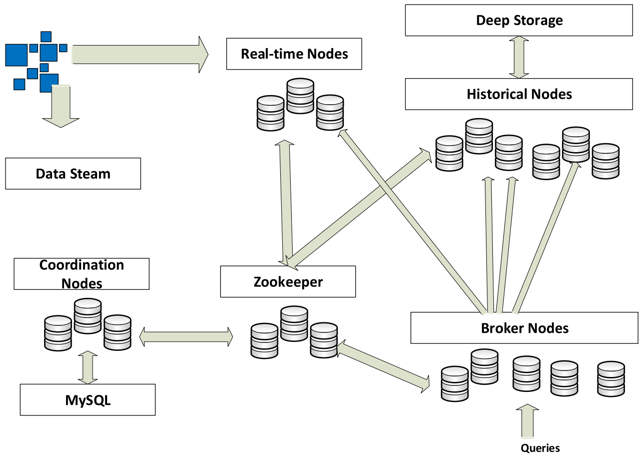

FIGURE 3. Cluster organization diagram.

to the user's own needs and network conditions to improve search efficiency and reduce search overhead. First set up a buffer pool for the search task. Annotate the source of information for each atomic search task. The subsystem fuses homogeneous atomic search tasks based on the relationship between the information source and the atomic search tasks. It also formulates a concurrent scheduling strategy for atomic search tasks. Finally, the subsystem uses a data fusion algorithm to process the search results. This method can reduce ambiguity and uncertainty, and analyze and reason out the optimal solution.

4) SECURITY DOMAIN

Open platform security management functions include access control, authentication and authentication, transmission encryption, vulnerability detection, attack monitoring, and security operation and maintenance management.

a: TRUSTED IDENTITY AUTHENTICATION PLATFORM

Big data analysis will likely perform user identity authentication and authorization in the fields of information security including SIEM (Information Security Event Management), network monitoring, and other fields. In addition, the security information field also includes identity management, fraud detection, and most products, which are enough to bring changes to the market.

In fact, the changes brought about by big data have already begun. Since 2014, leading security agencies have begun deploying big data solutions to support their security operations. The data analysis tools deployed in the soc (security operation and maintenance center) were all customized, but 2014 marked the beginning of the real commercialization of big data technology in the security field.

In terms of network supervision, the competent government department must perform the supervision function under the support of relevant laws, regulations, systems, standards, norms, processes, risk control and traceability mechanisms. Network supervision objects include network users, identity certificate issuing agencies, and identity authentication service agencies. In addition, network supervision objects also include network business systems and various types of data generated during network activities.

In order to meet the unified management requirements of the identity authentication platform on the collection, processing, and sharing of ID copies, the construction of the identity authentication system will refer to the software

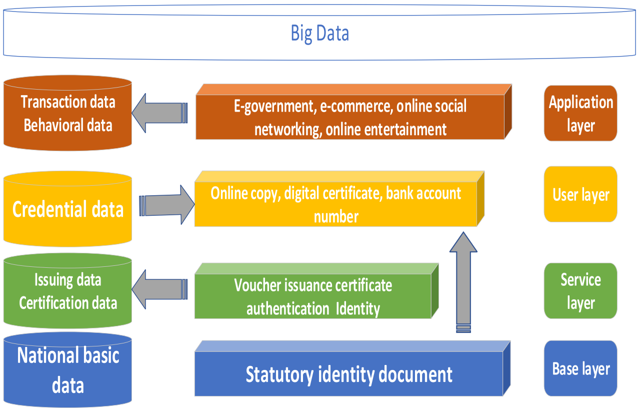
Big Data
Transaction data
Behavioral data
E-government, e-commerce, online social networking, online entertainment
Application layer
Credential data
Online copy, digital certificate, bank account number
User layer
Issuing data
Certification data
Voucher issuance certificate authentication Identity
Service layer
National basic data
Statutory identity document
Base layer

FIGURE 4. Overall framework of trusted identity authentication platform.

and hardware architecture of the public security business management system. This system is developed based on the public security population and certificate database resources. As shown in Figure 4, the platform realizes effective management and comprehensive application of identity authentication and copy collection. This system also realizes the real-time exchange of national authentication information data of the first-level platform and the second-level platform.

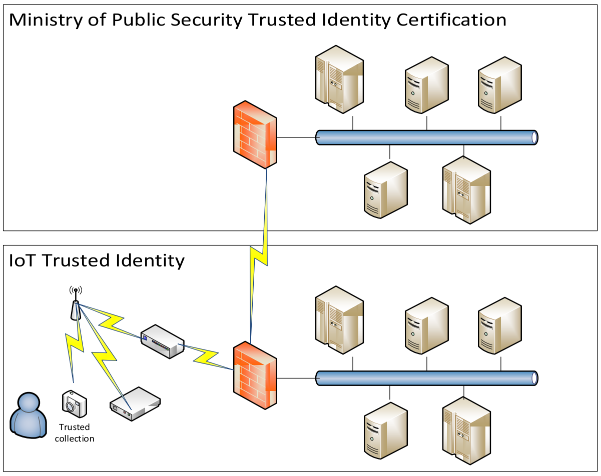

FIGURE 5. Trusted identity authentication platform structure diagram.

As shown in Figure 5, the platform is divided into two layers. The Ministry of Public Security has established a copy of the national ID card to gather the authentication information platform, which is the authentication node of the national ID card copy. In addition, the Ministry of Public Security has also carried out the construction of an industrial Internet trusted ID card authentication platform, and undertaken tasks such as information collection and identity authentication in the industrial Internet industry.

b: CLOUD SECURITY AND CREDIBILITY OF A TRUSTED ECONOMIC BIG DATA PLATFORM

In order to implement platform security protection technology to ensure that users will not be harmed by attacks in the platform, the platform faces a large number of malicious acts such as unauthorized access and data theft from outside the platform and other users. These actions bring security risks to the industrialization and commercialization of the platform. This system will study the realization of efficient virtual machine malicious behavior detection technology from the following two aspects.

a) Dynamic, multi-dimensional, fine-grained platform identity authentication and access control technology

The platform is facing security threats such as identity theft or user impersonation and unauthorized access, in addition to malicious insider threats.

From the perspective of data query, on the one hand, it is necessary to prevent the third-party service platform from mining user access patterns, that is, to obfuscate the mapping relationship between authorized users and their accessible data from the perspective of the third-party service platform. On the other hand, third-party service platforms are required to implement fine-grained access control to data that authorized users can access in the traditional binary structure.

In order to solve the above problems, we need to study platform identity authentication and access control mechanisms that meet the requirements of efficient, dynamic, and fine-grained access control. Through the above research, the trust needs of the platform to meet users and the security of the platform can be guaranteed.

b) Cloud security management and control based on virtual machine introspection technology

Virtual machine introspection (VMI) technology can support the monitoring of the operating status inside the virtual machine outside the virtual machine so that user behavior can be easily detected. However, the implementation of VMI determines the following aspects. The first is the degree of system coupling and versatility. The second is the difficulty of obtaining semantic information. The third is self-security and concealment. The fourth is the loss of platform performance.

Although the highly coupled VMI method can provide accurate semantic information, it also causes a considerable loss in platform performance. Therefore, the all-weather high-intensity VMI technology built on virtual machines and virtual machine managers runs counter to the platform's ‘‘efficiency first’’ design philosophy and is not highly usable in practice. And the behavior analysis technology built on the network level independently of the host can provide a view of the network behavior running inside the virtual machine without affecting the performance of the user host. In particular, it has a better detection effect on malicious network attacks, which is a powerful supplement to the use of VMI technology.

c: BACKUP MECHANISM

The open platform is based on distributed NoSQL cluster storage and implements a primary 2 backup mechanism for separate storage.

As shown in Figure 6, there are three modes of data storage: master-> slave, slave <-> slave, and cyclic. These three data storage methods allow the Trusted Industrial Internet

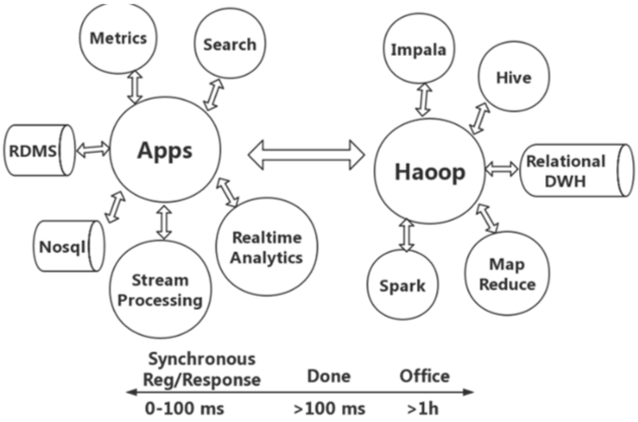

FIGURE 6. Flat backup schematic.

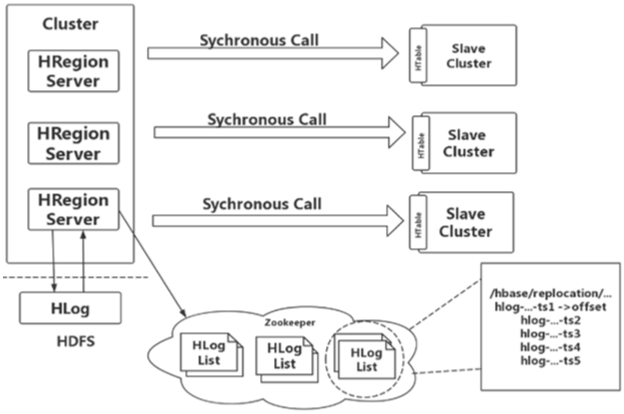

FIGURE 7. Streaming framework.

Open Platform to obtain data from any server and ensure that it can obtain all copies on other servers. In the event of a catastrophic failure in a data center, client applications can use DNS tools to redirect to another alternate location. The system provides ''eventual consistency'', which means that at any moment, the data is finally guaranteed to be consistent.

C. STREAMING FRAMEWORK

This system is based on the real-time requirements of the business. There is Storm that supports online processing, and Cloudar Impala. In addition, there are Spark and stream processing framework S4 that supports iterative calculation.

The storm is a distributed, fault-tolerant, real-time computing system developed by BackType and later captured by Twitter. The storm is a stream processing platform, which is mostly used to calculate and update the database in real-time. The storm can also be used for ''Continuous Computation''. Storm performs continuous queries on the data stream and outputs the results to the user in the form of a stream during the calculation. It can also be used in ''distributed RPC'' to perform operations in parallel.

Cloudera Impala is an open-source Massively Parallel Processing (CMPP) query engine developed by Cloudera. It has the same metadata, SQL syntax, ODBC driver, and user interface (HueBeeswax) as Hive. It can provide fast, interactive SQL queries directly on HDFS or HBase.

Impala is a query engine developed under the inspiration of Dremel. It no longer uses slow Hive + MapReduce batch processing but uses a distributed query engine (composed of 3 parts: Query Planner, Query Coordinator, and QueryExec Engine) similar to commercial parallel relational databases. Impala can query data directly from HDFS or HBase using SELECT, JOIN and statistical functions, thereby greatly reducing latency.

Figure 7 is a stream-centric architecture built on top of Apache Kafka.

In Figure 7, Kafka is only used as a general-purpose data pipeline. Each system is able to feed data to kafka, and it can also feed data from it. Applications or stream processors can access it and generate new, derived streams. These streams, in turn, can be provided to these systems. Data processing models are more diverse, and Hadoop is no longer an inevitable choice for building big data platforms.

In the application mode, the big data processing mode continues to be enriched. Batch processing, streaming computing, interactive computing, and other technologies are geared towards different demand scenarios and will continue to be enriched and developed.

In terms of implementation technology, in-memory computing will continue to be the main means to improve the performance of big data processing. Compared to traditional hard disk processing methods, memory computing has improved significantly in performance.

The open-source project Spark has been widely used in practical business environments and has developed into the largest open source community in the field of big data. Spark has a variety of computing frameworks such as stream computing, interactive querying, machine learning, and graph computing. Spark supports Java, Scala, Python, R, and other language interfaces, which greatly improves the efficiency of data use. These advantages of Spark have attracted the attention of many developers and application developers. It is worth noting that the Spark system can be built based on the Hadoop platform or run independently without relying on the Hadoop platform.

Many new technology hotspots are continuously integrated into the diversified model of big data, forming a more diverse and balanced development path, and also meeting the diverse needs of big data. The author proposes to consciously link and integrate big data research and development into the big data technology ecology, or use the results of the technology ecology, or give back to the technology ecology.

In terms of learning technology, deep analytics will continue to be a representative, driving the application of big data intelligence throughout. The intelligence mentioned here especially emphasizes the extension of related capabilities, such as decision prediction and accurate recommendation. These extensions involving human thinking, influence, and understanding will become the key application directions for deep data analysis.

Compared with traditional machine learning algorithms, deep learning proposes a method for a computer to automatically generate features and integrates feature learning into the process of building models, thereby reducing incompleteness caused by artificial design features. With the help of deep neural network models, deep learning can more intelligently extract features at different levels of the data. This method enables a more accurate and effective representation of data. And the larger the number of training samples, the more advantageous the deep learning algorithm is over traditional machine learning algorithms.

At present, deep learning has made major breakthroughs in areas where training sample data is easy to accumulate, such as image classification, speech recognition, question-answering systems, and other applications, and has achieved successful commercial applications.

It is predicted that as more and more industries and fields gradually improve the collection and storage of data, the application of deep learning will become more widespread. Due to the complexity of big data applications, the fusion of multiple methods will be an ongoing norm.

D. PRINCIPLES OF SAFETY EARLY WARNING

Evolution method [9], as a search method using the difference of economic data, this system uses this method to realize financial data security early warning under big data analysis.

The parameters of the differential evolution method mainly include population size, scaling factor, and cross probability. In the standard differential evolution method, the scaling factor and crossover probability are fixed. If it is too large or too small, it will affect the searchability of the method, resulting in low quality of economic data security early warning under big data analysis. We need to optimize the differential evolution method so that the scaling factor and crossover probability can be adjusted at any time according to the current convergence of economic data, so as to improve the convergence performance of the method and ensure the effectiveness of early warning of economic data under big data analysis.

In order to optimize the network basis function σ_i and connection weight w_{ji} of the differential evolution, we use the method of real number encoding and set it as follows.

$S = \{X_1, X_2, \ldots, X_n\}$ represents the initial population of the data, where N is the number of populations of the data, Xi is any individual in the data population, and the data is initialized. Then,

$$\begin{cases} \sigma_i = \sigma_{\min} + rand\,(0,1) \times (\sigma_{max} - \sigma_{min}) \\ w_{ji} = rand\,(0,1) \end{cases} \tag{1}$$

In Equation (1), rand (0,1) represents a uniformly distributed random number between (0,1),

$\sigma_{max} = \underset{i\neq i, i=1,2\ldots h}{argmax}\,(abs\,(c_i - c_j))$, c_i and c_j represents the center of a data node. The number of data nodes is h.

The basic principle of the differential evolution method is to randomly select 3 different individuals, perform differential scaling processing on 2 of them, and fuse the processed result with the remaining 1 individual to finally realize the recombination and mutation of the population individuals. And obtain,

$$V_i(g+1) = X_{r1}(g) + F \times (X_{r2}\,(g) - X_{r3}\,(g)) \tag{2}$$

In Equation (2), V_i (g) represents the i-th data individual in the g-th population. For three randomly selected individuals $X_{r1}(g)$, $X_{r2}(g)$ and $X_{r3}(g)$, F is the scaling factor. In the standard differential evolution method, F is a fixed value. If the value of F is not reasonable, the problem of premature population maturity will occur, which will seriously affect the convergence speed of the algorithm. In order to accelerate the convergence speed and obtain the global optimal solution, a strategy of dynamically adjusting F based on the degree of population difference is proposed.

The connotation of population difference is that in a population space, all individuals in the space are clustered to obtain the number of population clusters. Individuals in this space have certain differences. If the difference is larger, it means that the more uniform the individual distribution in the space is, the more likely it is to obtain a globally optimal solution. In the initial stage of differential evolution, in order to diversify the individual population, F is dynamically adjusted by Equation (3).

$$\mathrm{F}\,(\mathrm{g}) = \begin{cases} F_{max} - (F_{max} - F_{min}) & X_i\,(g) > \tau_1 \\ F\,(g-1) & otherwise \end{cases} \tag{3}$$

In Equation (3), F_{max} and F_{min} The maximum and minimum values of the scaling factor. They represent the set iteration threshold. Cross-operation is performed on the target data individual X_i and the mutant individual V_i in the data, population to produce a new individual U_i, then:

$$\mathrm{U}_{ij}\,(\mathrm{g}+1) = \begin{cases} v_{ij}\,(g+1)\; rand < CR, & j = j_{rand} \\ x_{ij}\,(g+1) & otherwise \end{cases} \tag{4}$$

In Equation (4), rand is a random number of random distributions between (0,1). j_{rand} is a random integer distribution between [1, N]. CR represents cross-probability. Then,

$$\mathrm{CR} = \begin{cases} CR_{min}, & \dfrac{g}{g_{min}} < \tau_2 \\ CR_{max}, & otherwise \end{cases} \tag{5}$$

In Equation (5), CR_{min} and CR_{max} represents the maximum and minimum values of the cross-probability. τ_2 represents the original cross over the probability of the setting.

The differential evolution method mainly uses a one-to-one competition strategy to select the optimal economic data set. Based on the fitness, the competitive candidate Ui (g + 1) and the corresponding matching individual Xi (g) are eliminated. The winning individual will be selected for the next population. This process is repeated until the global optimal solution of the method is obtained. This can realize advanced security early warning of economic data under big data analysis. Then,

$$X_i\left(g1\right)=\begin{cases}X_i\left(g\right) & f\left(U_i\left(g+1\right)\right)<f(X_i(g))\\ U_i\left(g+1\right) & otherwise\end{cases}\tag{6}$$

In Equation (6), $f\left(\cdot\right)$ a function that represents an individual's fitness.

According to the above discussion, by adjusting the scaling factor and crossover probability of the differential evolution method, the convergence performance of the method is improved, and economic data security early warning under big data analysis is realized.

IV. BANK CREDIT RISK CASE ANALYSIS

The rapid development of the Internet and its applications has produced a variety of structured and unstructured mass data. Their storage and processing have encountered unprecedented challenges, and many related technologies have never been studied.

With the explosive development of society and production, new types of smart cities and smart communities have emerged. In addition, informatization in transportation, communications, energy, and other industries has also produced a large amount of data.

Among the large amounts of data generated by various industries, the size and strength of data in the financial sector occupy the first place in the industry. Therefore, the application of big data technology in the financial industry is very necessary and has great potential.

At present, the big data sources that can be used for credit risk analysis are mainly divided into three categories [10], which are customer-related information within the bank, industry and supply chain information, and external data. The bank's internal and customer-related information data includes customer basic information, transaction data, financial data, etc. Industry and supply chain information includes the industry's prospects and status, social public relations, and the operation of supply chain companies. External data includes information on social networks, capital markets, search engines, and related websites.

A. APPLICATION BACKGROUND OF CREDIT RISK ANALYSIS IN THE FINANCIAL INDUSTRY

Credit risk analysis based on big data can integrate as much information as possible. Comprehensive analysis of the analysis object has the characteristics of mass, changeable form, fragmentation, etc. [11].

Applications in the financial sector allow for a comprehensive analysis of individuals. These applications implement credit analysis, credit risk analysis, and internal risk early warning based on big data. In addition, there are value customer mining and industry risk analysis.

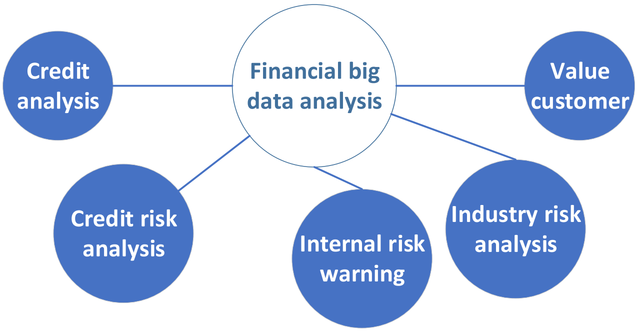

FIGURE 8. Application of big data in the financial industry.

As shown in Figure 8. Credit analysis can conduct a comprehensive and systematic analysis of the debtor's moral character, repayment ability, capital strength, guarantee, and environmental conditions. Convenient for banks to evaluate and analyze lenders before lending.

Big data-based credit risk analysis refers to the ability of banks to analyze their ability to make repayments on time and predict the probability of bad loans based on comprehensive information from customers in real-time after the loan, in order to expect the timely detection of bad loans and early warning.

Existing credit risk analysis focuses on credit management. This kind of risk analysis lacks dynamic analysis and timely warning. The main methods are expert system method, feature analysis method and credit scoring method. The expert system method has strong subjective factors, but the uneven level of experts leads to unstable and inaccurate credit risk discovery and management. The feature analysis method is based on a one-sided analysis of basic customer information, financial information, and internal bank data. Therefore, the accuracy of this method needs to be improved. The credit scoring method is based on the borrower's different credit indicators, weighted differently, and calculated using a simple mathematical statistical model. The credit scoring method has a single analysis method and is only suitable for static analysis. It cannot deal with sudden or potentially complicated situations.

Big data-based internal risk warning refers to analyzing abnormal behaviors of financial insiders based on employee personal information, information on internal communication tools, information sent by email, access to external networks, and network traffic. This internal risk early warning can prevent risks inside financial companies and ensure the stable operation of financial markets [12].

The main method to prevent internal threats in the financial system is to build an audit model that can be updated iteratively and has a learning function based on big data. This model comprehensively analyzes the suspicious behavior of users and conducts internal risk detection. This method specifically starts from two aspects of horizontal and vertical. Among them, horizontal analysis refers to machine learning methods based on big data to calculate outliers of people with abnormal behavior from the overall and local perspectives. Vertical analysis is based on the historical behavior of internal employees, statistics of behavior patterns and then found whether there is abnormal activity. The model uses internal risk analysis based on big data to formulate corresponding control procedures and measures for key personnel in sensitive positions in the financial industry to reduce threats and save costs.

Value-based customer discovery based on big data is to discover potential customers through the customer's transaction information and selected transaction locations, as well as other external information. It also conducts financial product marketing and recommendation based on customer behavior characteristics [13]. For example, according to the customer's education level, work industry, position, age, gender, past transaction history, and other characteristics, analyze the customer's preferences and make corresponding financial product recommendations. This method can determine whether the customer has the potential to become a key customer in the future. On this basis, different service strategies are formulated for different types of customers. This method can retain existing customers and develop potential customers. In addition, this method can also achieve refined management, improve the resource allocation of the financial industry, and make fine-grained contributions to the stable and robust development of the financial market.

Big data-based industry risk analysis means that banks understand the risks of various industries by understanding the current status and prospects of various industries and combining government, market, enterprise, and network information. This method formulates different sales, recommendations, and stop-loss strategies based on differences between industries while maximizing profits while avoiding possible risks. Banks can change the credit strategy and the distribution ratio of funds in various industries in real-time based on the analysis results of the industry. This is an effective measure for banks to effectively reduce credit risk, which can effectively increase the entry of funds.

B. CREDIT RISK ANALYSIS AND EARLY WARNING METHODS

At present, internationally advanced banks have comprehensively integrated customer information for comprehensive evaluation based on big data analysis, and dynamically and in real-time calculate the probability of bad loans. This greatly reduces the reliability of bank credit collection and strategy implementation [14]. The main implementation strategy is that the bank analyzes the relevant data of existing credit enterprises before the loan, evaluates the credit risk of the enterprise, and conducts the inspection through the integrated analysis of big data in the later stage.

Big data is very different from traditional risk analysis methods. These differences include the source, dimensions, form, and method of analysis of the data.

From the data source and dimensional analysis, credit risk analysis based on big data introduces third-party data in addition to the existing data in the bank. For example, third-party transaction platform information, third-party payment information, online shopping and logistics information, product evaluation information, and social network data. This information can improve the lack of internal information of the bank, improve the accuracy and timely effectiveness of credit risk analysis.

Analyzing from the form of data, big data has data in multiple formats. It includes structured data and includes unstructured data. This poses challenges for both data storage and analysis.

From the perspective of analysis methods, because the data is particularly huge, the data is multi-sourced, and the forms are diverse and fragmented, in terms of data feature extraction, it is necessary to exclude interference data extraction related data. These characteristics make the feature extraction method especially important. In addition, it is also necessary to analyze the extracted relevant data using appropriate machine learning methods, in order to discover the connections between the massive data and the underlying meaning. Therefore, big data analysis methods are more challenging than traditional analysis methods [15].

Big data-based credit risk analysis uses multi-source big data and advanced big data analysis techniques to help the financial industry provide more accurate early warnings of credit risks. And this method can ensure the timeliness of the forecast results and it can provide banks with accurate guidance-oriented analysis and effective support for the formulation of credit policies.

C. BIG DATA STORAGE AND COMPUTING TECHNOLOGY

Hadoop is a simple distributed storage and management platform suitable for processing big data. It can be developed in combination with a framework written in the Java language and can be integrated with the Java SSH framework developed by the credit analysis and early warning platform. Hadoop is open-source with the advantages of low cost, high reliability, and fast processing speed. This makes Hadoop very suitable for credit risk analysis and early warning systems. This system platform uses Hadoop as the support for big data storage and computing.

1) HDFS-BASED DATA STORAGE TECHNOLOGY

Hadoop Distributed System (HDFS) is used to store large amounts of data [16], and it is a distributed file system component. The so-called distributed storage refers to an efficient and reliable storage method in which multiple servers can perform storage and computing operations in parallel. Hadoop can dynamically allocate storage resources through real-time calculation during the storage process, which significantly increases resource utilization.

HDFS's overall architecture follows modular, interactive, and hierarchical storage. First, modular storage means that HDFS stores system metadata and application data in the NameNode server and the DataNode server, respectively. Metadata refers to data describing the characteristics of the data and attribute information. All servers communicate with each other through transmission control protocols. Second, interactive storage means that the DataNode server stores data backups on multiple other DataNode servers for data protection. Hadoop uses interactive storage to achieve storage reliability and does not use data protection mechanisms like the Parallel Virtual File System (PVFS). This strategy has the additional advantage of doubling the data transmission bandwidth while ensuring data durability and stability.

Hierarchical storage refers to that the HDFS namespace is represented by the inode using a hierarchical structure based on files and directories. Files and directories in different inodes record data operations and space attribute information.

An HDFS client creates a new file as shown in Figure 9. The client needs to query the NameNode server first to get the file blocks in the first DataNode. Then based on the point-to-point strategy, the next file block is obtained to form a file writing channel until the file writing is completed. The above storage strategy increases the throughput while ensuring the reliable storage of big data, so the HDFS-based storage architecture is very suitable for multi-source big data storage in this system.

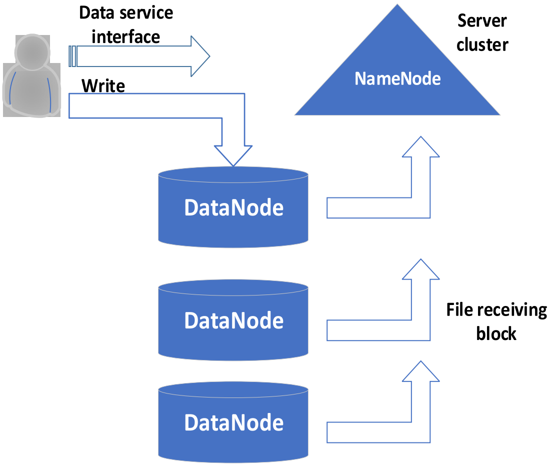

FIGURE 9. File storage process based on HDFS framework.

2) DATA REDUCTION TECHNOLOGY BASED ON MAP REDUCE

MapReduce is used to process and generate large data sets. It can compute these data sets in parallel on multiple large clustered machines to handle machine failures. MapReduce effectively uses the network and disks by scheduling communication between machines to meet a variety of real-world tasks.

It has been agreed that MapReduce is used to complete big data computing tasks. On average, 100,000 tasks are executed on MapReduce every day. Google's MapReduce cluster can complete about 20pb of data per day.

MapReduce calculations use a set of input key-value pairs and a set of output key-value pairs. It expresses calculations as two functions, map and reduces. MapReduce uses key-value pairs to pass data to the reduce function. The function can concatenate all data key-value pairs, and through calculation can form a small set of values, which can deal with the problem that the data cannot be operated in memory because the data is too large.

This example first introduces the relationship between big data and credit risk and analyzes it from five aspects: credit analysis, credit risk analysis, internal risk warning, value customers, and industry risk analysis. The thesis focuses on the application of big data in credit risk analysis. In addition, the paper also explains the necessity and challenges of credit risk based on big data. The second section is data storage and computing technology. This paper introduces Hadoop Distributed Storage Technology (HDFS) and MapReduce's big data computing technology and solves the big data storage and computing problems of the credit distribution analysis system based on multi-source big data.

V. SUMMARY

This paper introduces the emerging information technology methods used in national economic security risk early warning analysis.

We explained the overall architecture design to the detailed technical route. We focused on platform security technologies and a brief summary of the streaming framework technology genre. In addition, we conducted preliminary research on the application of emerging information technologies such as big data and cloud computing in economic security simulation and early warning practice. In addition, we analyzed the application of big data analysis technology in bank credit risk early warning. This provides the possibility and way for national economic security risk early warning and simulation.

In the future, more technologies will be selected for early warning platform, for example, we will consider adding real-time data processing [17], [18], expert decision support [19], etc. In addition, we also consider opening the platform [20]–[25] to be used by more researchers and policymakers as a bridge for academia, industry, and government.

REFERENCES

[1] Y. Zhang, Y. Xu, Z. Y. Dong, Z. Xu, and K. P. Wong, "Intelligent early warning of power system dynamic insecurity risk: Toward optimal accuracy-earliness tradeoff," *IEEE Trans. Ind. Informat.*, vol. 13, no. 5, pp. 2544–2554, Oct. 2017.

[2] M. A. Abiad, *Early Warning Systems: A Survey and a Regime-Switching Approach*, nos. 3–32. Washington, DC, USA: IMF, 2003.

[3] Y. Z. Wang, J. J. Hou, and Z.-Z. Liu, "A survey of economic earlywarning models," *J. Hum. Univ. (Social Sci.)*, vol. 2, pp. 27–31, Feb. 2004.

[4] R. Ong, *China's Security Interests in the 21st Century*. Evanston, IL, USA: Routledge, 2007.

[5] Z. H. W. Lei, "The measurement system of economic security of China in the context of globalization," *World Economy Study*, vol. 1, pp. 8–13, Jan. 2011.

[6] L. D. Xu and X. Hu, "The feasibility and the countermeasure of reducing environmental injustice of the underdeveloped regions," *J. Kunming Univ. Sci. Technol. (Social Sci. Ed.)*, vol. 1, pp. 7–11, Jan. 2012.

[7] E. M. Akhmetshin and V. L. Vasilev, "Control as an instrument of management and institution of economic security," *Acad. Strategic Manage. J.*, vol. 15, p. 1, Jan. 2016.

[8] H. Zhou, Y. Qiu, and Y. Wu, "An early warning system for loan risk assessment based on rare event simulation," in *Proc. Asian Simulation Conf.* Cham, Switzerland: Springer, 2007, pp. 85–94.

[9] X. Li, Q. Wang, L. Yang, and X. Luo, "Network security situation awareness method based on visualization," in *Proc. 3rd Int. Conf. Multimedia Inf. Netw. Secur.*, Nov. 2011, pp. 411–415.

[10] S. Piramuthu, "Feature selection for financial credit-risk evaluation decisions," *Informs J. Computing*, vol. 11, no. 3, pp. 258–266, Aug. 1999.

[11] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. J. Patil, and D. Barton, "Big data: The management revolution," *Harvard Bus. Rev.*, vol. 90, no. 10, pp. 60–68, 2012.

[12] F. L. Greitzer, J. R. Strozer, S. Cohen, A. P. Moore, D. Mundie, and J. Cowley, "Analysis of unintentional insider threats deriving from social engineering exploits," in *Proc. IEEE Secur. Privacy Workshops*, May 2014, pp. 236–250.

[13] M. Zineldin, "Quality and customer relationship management (CRM) as competitive strategy in the Swedish banking industry," *TQM Mag.*, vol. 17, no. 4, pp. 329–344, Aug. 2005.

[14] B. Fang and P. Zhang, "Big data in finance," in *Big Data Concepts, Theories, and Applications*. Cham, Switzerland: Springer, 2016, pp. 391–412.

[15] M. Bennett, "The financial industry business ontology: Best practice for big data," *J. Banking Regulation*, vol. 14, nos. 3–4, pp. 255–268, Jul. 2013.

[16] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop distributed file system," in *Proc. MSST*, vol. 10, 2010, pp. 1–10.

[17] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[18] D. Zhang, "High-speed train control system big data analysis based on the fuzzy RDF model and uncertain reasoning," *Int. J. Comput., Commun. Control*, vol. 12, no. 4, pp. 577–591, 2017.

[19] D. Zhang, J. Sui, and Y. Gong, "Large scale software test data generation based on collective constraint and weighted combination method," *Tech. Gazette*, vol. 24, no. 4, pp. 1041–1049, 2017.

[20] S. Sagiroglu and D. Sinanc, "Big data: A review," in *Proc. Int. Conf. Collaboration Technol. Syst. (CTS)*, May 2013, pp. 42–47.

[21] R. A. A. Habeeb, F. Nasaruddin, A. Gani, I. A. T. Hashem, E. Ahmed, and M. Imran, "Real-time big data processing for anomaly detection: A survey," *Int. J. Inf. Manage.*, vol. 45, pp. 289–307, Apr. 2019.

[22] H. Atlam, R. Walters, and G. Wills, "Fog computing and the Internet of Things: A review," *Big Data Cognit. Comput.*, vol. 2, no. 2, p. 10, 2018.

[23] D.-H. Shih, H.-L. Hsu, and P.-Y. Shih, "A study of early warning system in volume burst risk assessment of stock with big data platform," in *Proc. IEEE 4th Int. Conf. Cloud Comput. Big Data Anal. (ICCCBDA)*, Apr. 2019, pp. 244–248.

[24] S. Li and H. Yu, "Big data and financial information analytics ecosystem: Strengthening personal information under legal regulation," *Inf. Syst. E-Bus. Manage.*, vol. 17, pp. 1–19, Jan. 2019.

[25] Z. Lv, X. Li, and K.-K.-R. Choo, "E-government multimedia big data platform for disaster management," *Multimedia Tools Appl.*, vol. 77, no. 8, pp. 10077–10089, Apr. 2018.

YI LIANG is currently pursuing the Ph.D. degree in applied economics with the School of Economics and Management, Beijing Jiaotong University, Beijing, China. His current research involves in industrial economics and industrial security.

DAIYONG QUAN received the Ph.D. degree in computer science from the University of the Chinese Academy of Sciences, Beijing, China, in 2016. He is currently a Senior Engineer of archives of University of Science and Technology Beijing, China. His research interests include intelligent algorithms, economic risk prediction model and digital asset protection.

FANG WANG received the Ph.D. degree in statistics from Beijing Jiaotong University, Beijing, China. She currently holds a postdoctoral position with Beijing Jiaotong University. Her main research involves in game theory, machine learning, industrial economics, and industrial security.

XIAOJUN JIA received the Ph.D. degree in economics from the Renmin University of China, Beijing, China. She is currently an Associate Professor with the Beijing Center for Industrial Security and Development Research, Beijing Jiaotong University. Her main research involves public finance, finance, industrial economics, and industrial security.

MENGGANG LI received the Ph.D. degree in applied economics from Beijing Jiaotong University, Beijing, China. He is currently the Dean of the National Academy of Economic Security of Beijing Jiaotong University, the Director of the Beijing Laboratory of National Economic Security Early-warning Engineering, the Director of the Beijing Philosophy and Social Science Beijing Industrial Security and Development Research Base, and the Chairman of the IEEE Professional Committee in Logistics, Informatics and Industrial Security System. His current research concerns national economic security, industrial economics, and industrial security.

TING LI is currently pursuing the Ph.D. degree in applied economics with the School of Economics and Management, Beijing Jiaotong University, Beijing, China. Her current research involves in industrial security and industrial structure.