Research and Application of Cloud Computing and Big Data Technology

Bingran Hui *
School of management, Tianjin University of Technology, Tianjin, China

*Corresponding author e-mail: huibingran@tjut.edu.cn

Abstract. The combined application of new network technologies with the traditional Internet and industry has enabled new businesses such as the Internet of Things and cloud computing to be gradually derived. To realize the construction of a modern society in China, great attention should be paid to the application of new network technologies such as cloud computing in actual work. Now through data integration and correlation analysis, the accuracy and innovation of information are further realized, and the overall work efficiency and work quality are improved. This paper systematically analyzes the network big data technology and applications of cloud computing and Internet of Things, and conducts research from data collection, data integration, data processing, and data governance. Its main purpose is to make the value and function of these new network technologies fully utilized.

Keywords: Cloud Computing; Data Collection; Data Integration; Data Center; Research and Analysis

1. Overview

Under the supervision requirements of modernization and informatization and the development requirements of the new network market, the "smart city supervision" based on cloud computing information technology is a new type of supervision that conforms to the current wave of informatization and responds to the new situation and new challenges of market supervision model.

In the market supervision big data model, due to the characteristics of big data, the big data supervision model has many points worthy of promotion and reference compared with the traditional market supervision. First, "information islands" can be fatally harmed through data sharing platforms, enables smooth coordination of vertical levels, horizontal departments, and even cross-regional and cross-border within the government, which is conducive to promptly discovering and solving problems in supervision. Second, through the collection and analysis of big data, it can more accurately describe the “Subject portrait”, “Object portrait”, "Behavior portrait", etc. of the market supervision field, which can be precise and personalized for government departments when formulating regulatory policies or adopting regulatory measures. The third is based on the sharing and opening of government data, which can improve the past asymmetry of information, and greatly increase the transparency of the government, and the improvement of administrative efficiency of the government will help to increase the credibility of the government. Cloud computing technology, as a necessary means for the verification of the big data model, has been combined with big data as a whole and cannot be separated.
At the inaugural meeting of the China Big Data Expert Committee, Academician Huai Jinpeng, the director of the committee, described the relationship between big data and cloud computing with a formula:

$$G = f(x)$$

(1)

Among them, G represents our goal; f represents cloud computing; x represents big data. Big data is a demand, and cloud computing is a means of processing big data. Big data and cloud computing are like the pros and cons of a coin. The two complement each other and are indispensable.

2. Collection, integration, and governance of market supervision big data

2.1. Data collection

At present, the Market Supervision Administration has a wide range of data sources, including data from the original industry and commerce, food and drug supervision, quality supervision, intellectual property, price, and other departments. Due to the wide range of data sources, the data format is more complicated. The data standards are not uniform, and the technical system is difficult. There is still room for greater improvement in the degree of integration and utilization of Internet data. At this stage, the main ways of collecting business data in the field of market supervision are as follows.

1) Market supervision business system. The province currently uses the same system. Deploying in two levels at the provincial and prefecture levels. The system data of each city will be collected regularly to the provincial bureau data center. 2) The enterprise fills in the data online. At present, business entities can fill in data such as annual report information and real-time information through the corporate credit system publicity system, government service network market supervision flagship store, etc., and collect them regularly from the government affairs extranet to the provincial data center. 3) Market supervision information platform. Mainly refers to the system deployed on the e-government extranet and shared by provincial departments. Relevant data will be used and supplemented by various departments. Among them, the market supervision bureau mainly provides market subject information, and at the same time, it is regularly collected from the government extranet to the provincial data center. 4) Internet data. Aiming at the massive amount of information existing on the Internet, based on artificial intelligence and semantic analysis technology, the Internet big data collection and structural transformation are realized. Internet data mainly includes consumer online public opinion, Internet advertising, e-commerce platform online stores and their operating data, etc. By purchasing third-party services, you can obtain corporate public opinion, business operations, e-commerce product reviews and other data, enriching market supervision big data analysis. For this reason, under the premise of maintaining stable operation of the data collection methods of traditional market supervision and administration services, the project uses the Internet to build new data collection methods in the era of big data, continuously enriching data sources, and gradually standardizing data standards. This is the key issue of the first breakthrough.

2.2. Data integration

1) Network interconnection based on cloud computing platform. Establish data standards for Internet partners, set up a data collection server on the public network, and transmit the data of Internet partners to the government extranet through a gatekeeper device. The construction provides a variety of mainstream data collection interfaces, including text files, Excel files, databases, message queues, Webservices, etc., to meet the data upload needs of Internet partners, as shown in Figure 1.
2) Data notes and classification. By labeling business data and combining with the mainstream classification methods of the industry, establish the standards of Internet data industry and market supervision and administration. By using machine learning, semantic analysis, and artificial intelligence achieve the automatic merging of new data types on the Internet into the standards of the State Market Supervision Administration, achieve the integration of Internet data and internal business data of the Market Supervision Administration. By combining industry standards in various sectors, establish the mapping relationship between the standards of various departments and the standards of the State Administration of Market Supervision, achieve the integration of market supervision data in various departments.

2.3. Data governance
The control of data quality can start from all aspects of the data processing process such as data integration, data preprocessing, resource storage, resource monitoring, resource utilization, etc., to enhance the use value of data in the analysis and decision-making of the market supervision and administration bureau, and to give play to it. The real role of data as an asset.

It can be seen from Figure 2 that data quality is oriented to the entire business process, from the business library to the ETL to the data warehouse can be controlled by user-defined data quality rules. The verification rules are divided into three types: accuracy, completeness, and consistency, which are oriented to three different verification schemes. The existing data in the business database and data warehouse can be directly checked for online quality rules. In the ETL process, data cleaning can be performed through the data cleaning service provided by the quality rule definition. At the same time, data can be compared between two or more data sources. Users can select the comparison data source, customize the comparison rules, and perform online data comparison. The result of the comparison can reflect the data differences between tables in different libraries, and further solve the problem of data
quality. The data quality verification methods are divided into full verification and sampling verification, and the execution methods are divided into manual execution and timing execution. The following is a detailed breakdown of each link.

1) Formulation of data governance rules. The quality rule is divided into two parts: the rule list and the object list. Through these two parts, the data quality can be checked in the forward and reverse directions. The so-called "forward" refers to the definition of rules first, and after selecting the data directory and rule conditions on the rules, performing the detection. "Reverse" refers to directly selecting a data directory for data quality verification. The customized rule can choose the recommended processing method (manual processing and timing processing), which is convenient for finding the best processing method when using the rule for quality inspection in the future. 2) Automatically monitor data flow. Through the above series of data identification, data comparison, and data detection, the system can automatically generate personalized data quality inspection reports on a regular basis, vividly describing the location of data quality problems and processing conditions, which are specifically divided into business data quality reports and technical data quality report. 3) Data detection and comparison. Data testing mainly includes four aspects: data integrity testing, data consistency testing, data accuracy testing and data timeliness testing. To ensure the consistency between the synchronized database and the source database, a series of comparison rules need to be established to check whether the data is consistent. Generally, the incremental comparison method is adopted, which can reduce repeated comparisons. The workload of some comparisons can be directly repaired, and some comparisons may need to notify the user by sending an email to notify the user to manually repair the data. 4) Data quality rating. In response to data quality issues such as missing data sets, mismatched metadata, and inaccurate data, the data management platform has formulated objective scoring rules, uniformly scoring, and rating each data catalog, and providing a comprehensive online preview of the management system Data quality score ranking, problematic data, details of various data quality problems, etc.

3. Application of Cloud Computing Information Technology in Market Supervision Data Center

The data center is built based on mainstream big data processing and Internet construction technologies. These technologies are necessary to build PB-level data processing and concurrent access by millions of users. The open system of the data center technical system architecture is shown in Figure 3, including five levels.

![Figure 3. Data center technology system architecture](image)

1) A big data storage system based on the Hadoop system, including a distributed file system and a distributed columnar database, provides a unified, secure, flexible, and scalable storage system. In-memory storage technology is also an important part of the storage system. In-memory storage involves
technical fields such as distributed caching and in-memory databases. 2) In the data acquisition and data exchange system, different tools are used according to different data types and source databases, including sqoop software for collecting structured data to Hadoop, ETL tool software kettle, and flume for collecting logs Tools etc. 3) Data processing technology system, including offline computing, streaming computing, memory computing, machine learning, search engine, etc., to respond to different usage scenarios. 4) Cluster management and control platform. The entire distributed cluster needs to be managed uniformly. A unified management and control platform has been developed in the platform technology ecosystem, including a distributed task scheduling system, resource management system, data security and access control system, service monitoring and management system. 5) Big data processing management system and tools, used for data life cycle management, including processes from data modeling, metadata definition, data relationship, data visualization, and data service.

3.1. Data storage
1) Distributed file system. Market supervision big data involves a large number of unstructured files such as images, audios, and videos, and with the development of applications, it needs to be stored through a distributed file system. By using the Hadoop Distributed File System (HDFS) to store unstructured files such as images, audios, and videos related to the field of market supervision, it can realize the massive efficiency and reliability of unstructured files such as big data related images, audio, video and so on by the Market Supervision Administration Storage, build a distributed cluster of HDFS. 2) Distributed database. The database is the core value of the platform. All kinds of valuable data will be stored in the database. An efficient, safe, and reliable database is the basis for the stable operation of the platform. In today's mainstream large-scale Internet big data architecture model, relational databases and NoSQL databases are generally mixed. Relational databases are used to deal with transactional core businesses, and NoSQL databases are used to deal with massive low-value-density businesses. Distributed database support services will be built from two aspects: relational database clusters and distributed NoSQL databases. 3) Distributed cache. To reduce the frequent reading of storage devices and improve the access efficiency of peak users, the platform provides unified distributed caching software, which adds high-speed cache between the database and the application for data resources that change less but need to be read frequently. It can reduce the pressure on the data access layer, greatly improve the system performance, and at the same time can avoid the downtime of the server based on the horizontal expansion architecture, through the distributed cache can ensure the high availability of the platform.

3.2. Distributed Computing
The analysis of the Market Supervision Administration involves various types of data, including internal data of the Market Supervision Administration, data from other government departments, Internet/mobile Internet data, etc. These are various types, complex relationships, and large amounts of data. They rely on traditional data query technology and analysis methods are difficult to meet the calculation and analysis needs of tourism data. The big data support platform needs to provide calculation and analysis capabilities for massive data. 1) Non-real-time mass data statistics are mainly used to generate aggregated data of the Market Supervision and Administration Bureau. Generally, the data of the Market Supervision and Administration Bureau on a monthly/quarterly/annual basis is not required for real-time performance. 2) Real-time data calculation processing is to calculate and process the data obtained in real time and be able to respond to the results in real time. The main characteristics are large data throughput and high data processing timeliness. 3) Real-time data query is to query different responses in real time according to the different input by the user. The system does not limit the query conditions, and the user can customize various combinations of conditions. Real-time data query enables users to face the market supervision and administration big data at any time and quickly obtain desired results. 4) Deep data mining and utilization refers to the discovery of valuable information hidden in massive data through artificial intelligence, machine learning, natural language processing and other related technologies.
In view of the above four scene classifications, this article proposes corresponding computing and analysis technologies and solutions: using distributed batch processing technology to achieve non-real-time computing of massive data; using streaming computing technology to achieve computing and processing of real-time data; through massive data interaction Inquiry technology realizes real-time inquiry of big data of market supervision and administration; through data mining technology, realizes in-depth mining and utilization of big data of market supervision and administration.

The cloud computing-based big data platform provides the necessary support for the analysis and verification of the market supervision administration by integrating the internal system data of the market supervision administration, the data of the relevant departments of the market supervision administration, and the Internet data. Through the analysis and mining of the complex big data from many aspects, find out the valuable information contained therein, and provide the basis for the overall analysis, research and judgment, and index evaluation of the market supervision and administration department.

4. Conclusion
This article aims to explore the application of cloud computing technology in market supervision business with the field of market supervision as the research object, with a view to further accelerating the construction of market supervision and management informatization and enhancing modern market supervision capabilities. Through the case analysis of the application of cloud computing storage technology in the market supervision process, the current problems of data application in the market supervision of our country are studied and analyzed. Case studies are carried out from the aspects of data collection and integration, data processing and quality control management.

References
[1] Fuyin Chu, Lin Zhang, Kunpeng He. Fast query and realization of massive data based on hadoop platform[J]. Computer Knowledge and Technology, 2016, 12(21): 3-5.
[2] Xiaojuan Zhou. Implementation of a lightweight big data analysis system [J]. Electronic Design Engineering, 2016, 24(8): 40-43.
[3] Wei Zhang. Design and analysis of key technologies for high-concurrency medical big data platform data[J]. Information Technology and Network Security, 2018, 37(4): 18-22.
[4] Jialong Luo, et al. Design and implementation of user behavior analysis platform based on big data analysis technology [J]. Computer Knowledge and Technology, 2019, 15(35): 54-56.
[5] Yuanwu Zeng, et al. Prospects for the construction of geographic information public platforms in the era of big data[J]. Bulletin of Surveying and Mapping, 2015(11):84-87,105.
[6] Youqun Ren. Supporting university governance with smart campuses: the role, consensus and vitality of university informatization from the perspective of "education governance"[J]. Journal of Hubei Correspondence University, 2014, 27(4):1.
[7] Xin Sun. Design and Implementation of Smart Campus in Higher Vocational Colleges [D]. Jinan: Shandong University, 2014: 1-2.
[8] Tong Zhou, Wen Zhou. The current situation and thinking of the construction of smart campus[J]. Information and Computer (Theoretical Edition), 2011(20): 86.
[9] Jingao Li, Zhenguo Qi, Ying Zhao. Research on Smart Campus and Its Construction[J]. Software Guide, 2016, 15(2): 140-142.
[10] Changhong Yu, Yunwu Wang, Ma Wu. The status quo, problems and countermeasures of smart campus construction[J]. Teaching and Management, 2015(6): 48-51.
[11] Helong Xu. Thoughts on the application of big data in the construction of smart campuses[J]. Fujian Computer, 2020, 36(4): 76-78.
[12] Lei Zhou. The current situation, problems and countermeasures of smart campus construction[J]. Modern Vocational Education, 2017(3): 105.
[13] Chuandi Tang. Research and realization of smart campus in colleges and universities[J]. Knowledge Economy, 2019(34):147,149.