Journal of Electronics and Information Science (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/jeis.2023.080304
ISSN 2371-9524 Vol. 8 Num. 3

# *Application Scenarios and Practice of Data Science in the Context of Big Data*

**Jianwei Ren**

*University of Wisconsin-Madison, Seattle, 53711, USA*

***Keywords:*** Big Data, Data Science, Practical Application, Interdisciplinary Disciplines

***Abstract:*** With the development of the Internet, the era of BD (Big Data) is getting closer and closer. A country with mature BD has a future, and many enterprises cannot compete without BD. For example, BD can accurately position people's hobbies, the sales industry or service industry can use BD for precision marketing, and the development trend of BD includes data resource, data science, and the establishment of data alliances. Data science is a specialized discipline, a discipline born in the era of BD. It is at the intersection of statistics, machine learning and domain knowledge, and is an obvious interdisciplinary discipline. With the development of BD, data science must also develop with it. How data science develops and in which scenarios it can be applied remains to be studied. Through the research on BD and its development trend, and the theoretical research and analysis of data science, this paper aims to explore the specific application of data science, a new discipline, and practice it. Experiments have shown that applying data science to filtering spam and malware has a filtering rate of up to 95%. When applied to the sales industry, the predicted results are almost identical to the actual results. It has been confirmed that data science can collect, process, analyze data, and make predictive inferences. Data science can be applied to personalized content, navigation, and other scenarios that require prediction of results.

## 1. Introduction

In the era of BD, the value of data is immeasurable. Even insignificant data may have great value. However, data science is a discipline born from the development of BD. Its knowledge system includes basic theory, data processing, data technology, etc. It is a discipline used to process data, including reducing the complexity of computing data, improving data quality, etc. For BD, many scholars have conducted research on BD applications and data science. Scholars Yu Canqing and Li Liming introduced the basic concepts of data science, and combined with the characteristics and development trend of large-scale Cohort study, analyzed the content and structure characteristics of Cohort study data. They explored the application and value of data science in different research stages of large population cohorts [1]. Wang Kai, Zhang Shaojie and Ma Juan identified 1944 landslide macro deformation stages from the GNSS (Global Navigation Satellite System) surface displacement monitoring database of four lithology regions, constructed the BD sample environment of deformation stage, and analyzed the spatial distribution law and early warning criteria of landslide macro displacement stages in four lithology regions under the BD environment

[2]. HUO Cun-xiao and HOU Yu discussed the possibility and advantages of BD in product development for the elderly, and proved that BD can be used as an effective data tool and analysis method to help designers build a highly continuous user data model [3]. Although data science is an emerging discipline, there are still many studies on it.

Data science is born from BD, which may have great value. Therefore, in order to dig out the potential value of data science, this paper first studies BD, and then studies and analyzes the development trend of data science and BD today, aiming to discover the application of data science in practice, what aspects it can be specifically applied to, and how effective it is. Experiments have shown that data science can be used for prediction, as it can collect, process, and analyze data to draw inferences. When applied to filtering spam and malware, the filtering rate can reach up to 95%, indicating that its prediction accuracy is very high.

## 2. Big Data

BD is also called huge data, which refers to data sets that cannot be captured and processed by conventional software within a certain time range. The characteristics of BD include: capacity, diversity, speed, variability, value, etc. BD includes structured, semi-structured and unstructured data. BD processing data can be roughly divided into three steps: data acquisition, data calculation and data usage. The corresponding BD core technologies can be divided into BD acquisition, BD preprocessing, BD storage and BD analysis [4-5].

(1) BD collection is to collect data from various sources, such as database collection, and collect data stored in the database; Network data collection, borrowing web crawlers or public API documents to obtain data from web pages. The collected data has a lifecycle, and if the data expires, it would be automatically cleared and destroyed [6]. The data lifecycle diagram is shown in Figure 1.

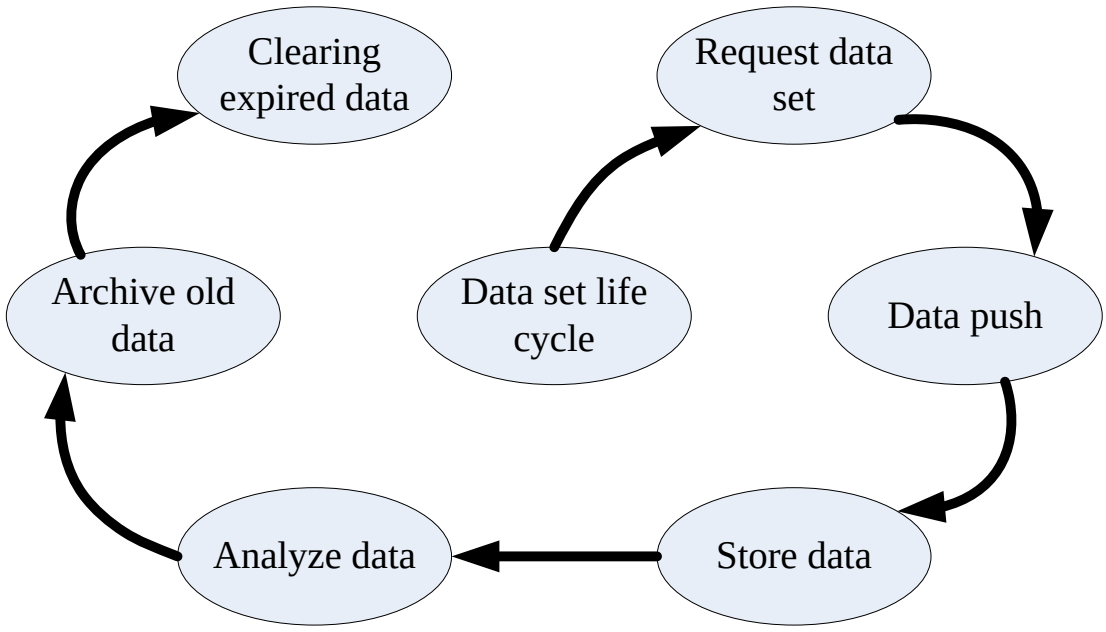


Figure 1: Data Lifecycle

(2) Before analyzing the data, it is necessary to clean, standardize, and verify consistency of the collected data, aiming to improve the quality of the data. Data preprocessing mainly consists of four parts: data cleaning, integration, transformation, and specification [7].

(3) BD storage is to use memory to store collected data in the form of database. The main methods of storage include using MPP (Massively Parallel Processing) architecture or using

Hadoop technology. BD storage is shown in Figure 2.

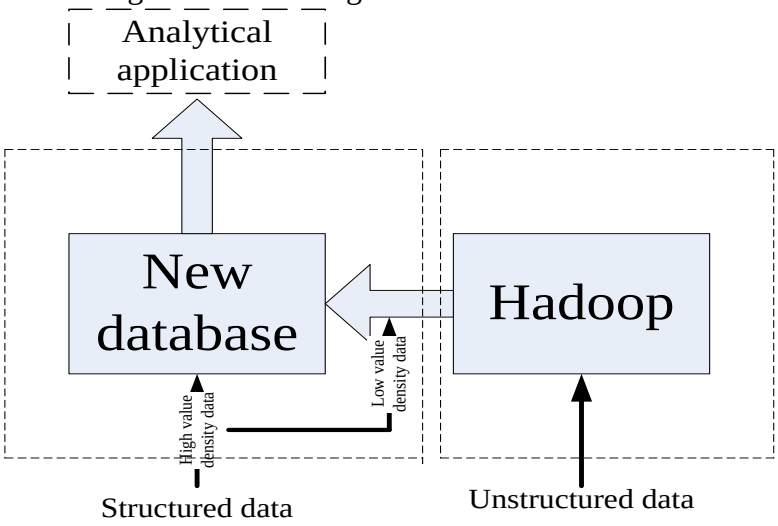


Figure 2: BD Storage

(4) BD analysis and mining, data mining is to extract valuable data that is hidden and unknown from a large number of data. Predictive analysis allows analysts to make forward-looking judgments based on the results of graphical analysis and data mining [8].

## 3. Data Science

Data science is an interdisciplinary discipline generated by BD. It is a discipline that transforms data into decision-making and action [9]. The essence of data is the traces left by the world's operation, and it can understand the laws of world operation by analyzing and understanding data in order to transform the world. Data science is a tool for humans and computers to work together to transform data into knowledge by collecting, describing data, and discovering knowledge [10-11].

As an interdisciplinary discipline, databases involve many aspects of knowledge, such as machine learning, visualization, statistics, mathematics, computers, and so on. Its basic theories are mainly statistics, machine learning and Data and information visualization.

The knowledge system of data science is shown in Figure 3.

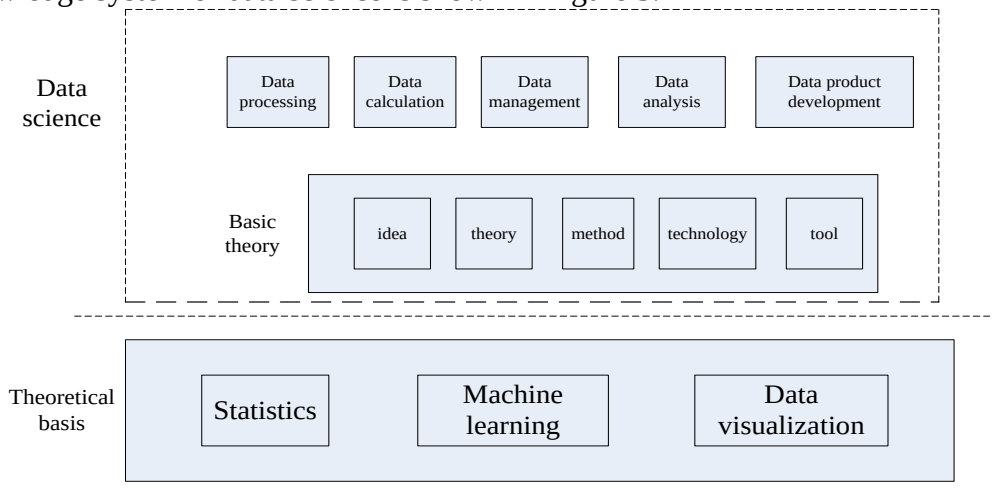


Figure 3: Data Science Knowledge System

### 3.1 Statistics

Statistics is a science that collects, processes, analyzes, and interprets data, and ultimately draws conclusions about it. Its common data analysis methods include descriptive statistics, such as Central tendency analysis, off center trend analysis, correlation analysis, inferential statistics, etc.; Hypothesis testing, including parametric testing and non-parametric testing; Regression analysis, such as univariate linear regression and multivariate linear regression analysis. Today, the combination of statistics, Information theory, system theory, etc., has further developed statistics, making the collection, processing, analysis, etc. of statistical data easier and faster [12].

### 3.2 Machine Learning

Machine learning is also an interdisciplinary field that involves disciplines such as statistics, probability, and algorithm complexity, and is the core of artificial intelligence [13]. Machine learning can further analyze complex and diverse data and utilize it more efficiently. Machine learning can be divided into machine learning that simulates human brain, such as symbol learning and neural network learning; Machine learning using mathematical methods, such as statistical machine learning, has three elements: model, strategy, and algorithm; There are also supervised and non-Supervised learning, structured and unstructured learning, Reinforcement learning, rule learning, and so on. Machine learning is also widely used, such as data analysis and mining, traffic prediction, network security defense (filtering spam and malware, etc.), pattern recognition, and so on [14-15].

### 3.3 Data and Information Visualization

Data and information visualization, that is, data is converted into chart form, which can display data more intuitively. Visualization can enable complex data to accurately, effectively, and efficiently express the implicit information in the data. Strategic decision-makers can quickly and accurately obtain the information behind a large amount of data by using Data and information visualization, which saves a lot of time and provides a reliable basis for formulating strategies [16]. The process of Data and information visualization mainly includes: (1) defining goals, what types of data need to be used, and what problems need to be analyzed. (2) Collect data and collect relevant data based on defined goals. (3) Clean up data by removing redundant or low value data from the collected data. (4) Select the visual effects of Data and information visualization, such as dynamic visualization or static visualization. (5) Create data visual effects [17].

## 4. Specific Applications of Data Science

According to the research and analysis of BD and data science, in theory, data science can be applied to personalized content, filtering spam and malware, consumption and service industries, navigation, supply chain management, etc. [18].

### 4.1 Application of Data Science to Personalized Content

It can collect various data from users' daily lives, analyze these data, and make predictive inferences. For example, short video software can collect data on the types of videos users browse, likes and favorites, and then analyze and predict the results to push videos that match the predicted results to users [19].

### 4.2 Filtering Spam and Malware

Data science involves machine learning, which can learn by inputting data from various spam and malware, increasing the ability to distinguish between spam and malware. Data science can collect and analyze data on spam and malware. When receiving emails and downloading software, predictive inference can be made based on the analysis results. If it is spam and malware, reminders or direct processing can be provided.

### 4.3 Application of Data Science in the Consumer and Service Industries

It is possible to collect and analyze past operational data, predict future demand, and then calculate to obtain the optimal solution. For example, in the sales industry, collect past sales data, analyze the data, understand the market situation, which products have good sales performance, and which products have achieved significant profits, etc.; Predict future market changes, sales volume, and inventory requirements; Calculate the optimal strategy.

### 4.4 Application of Data Science in Navigation and Supply Chain

It can analyze the collected road data and driver driving data, calculate the length of the road, calculate the number of traffic lights, predict traffic conditions, calculate estimated arrival time, etc., and standardize the appropriate road for driving.

For the distribution of goods in the supply chain, it can perform predictive analysis based on the assigned tasks, calculate the optimal route solution, and so on. Data science can visualize these Data and information visualization into roads, and drivers can intuitively see which road they are driving on [20].

## 5. Data Science Application Practice

According to the theoretical analysis of data science, it can collect, process, and analyze data for predictive analysis. To test whether data science can be accurately applied to navigation, filtering spam and malware, and whether predictive analysis can be conducted accurately. This article conducts experiments on its application in the sales industry and filtering spam and malware [21].

### 5.1 Application of Data Science in the Sales Industry Experiment

Divide the sales data of the three companies into the first half of the year and the second half of the year, and divide them into three groups: a, b, and c. Analyze the sales data of the first half of the year using data science to predict the sales results of the second half of the year. Compare the predicted analysis results with the company's sales data for the second half of the year. The experimental results are shown in Figure 4.

From Figure 4, it can be seen that the revenue predicted using data science is similar to the actual revenue of the three companies. It can be proven that using data science for prediction is indeed feasible and effective in applying it to the sales industry.

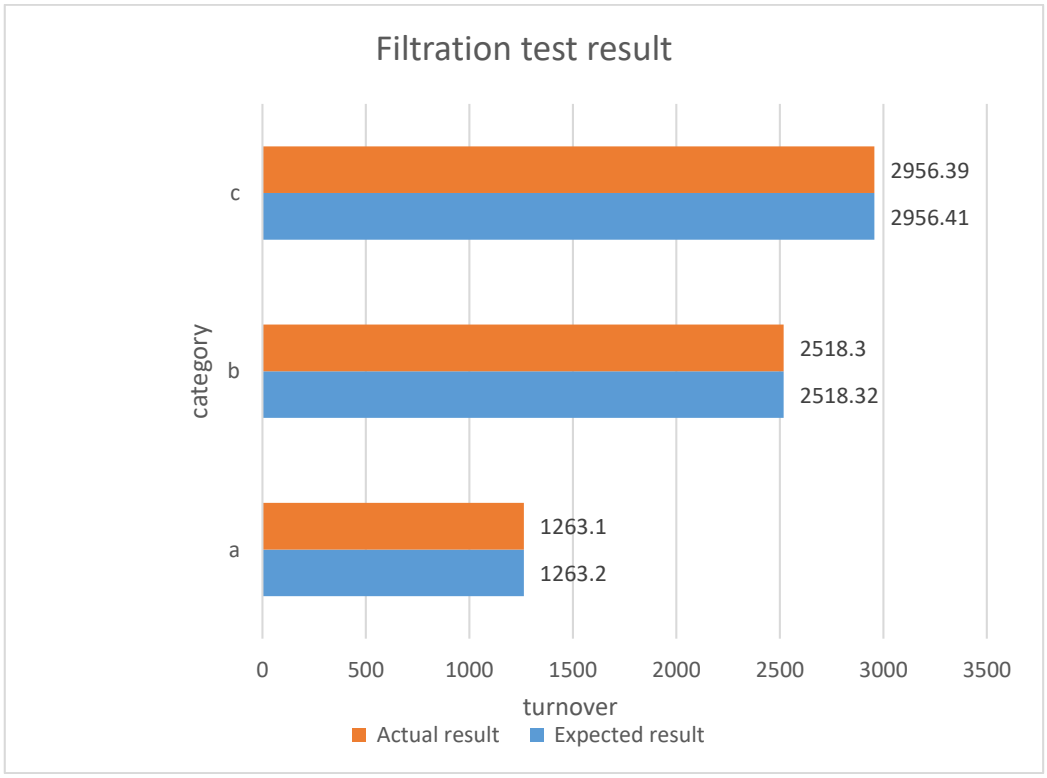


Figure 4: Expected and Actual Turnover

## 5.2 Filtering Spam and Malware Experiments

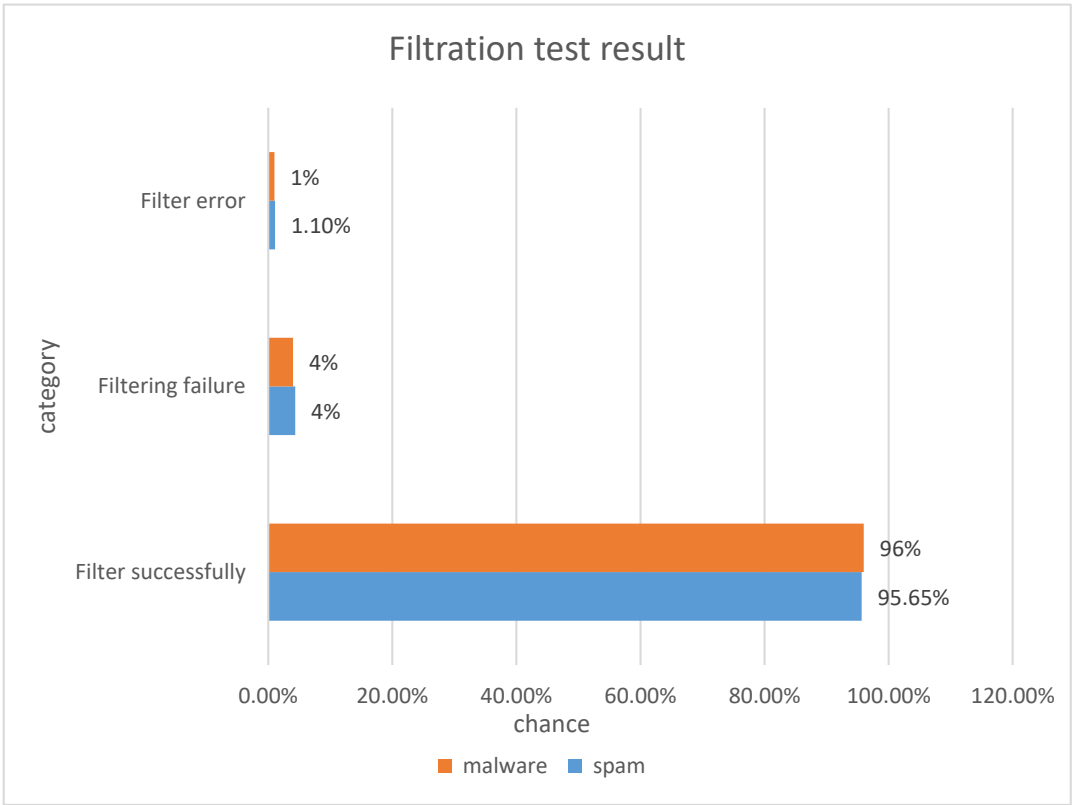


Figure 5: Filter Results

Mix a large amount of spam and malicious software with normal emails and software, and observe whether data science can accurately filter them and what is the success rate of filtering. Prepare 2000 normal emails and 2000 spam emails, mix 100 malicious software and 100 normal software, and observe how many have been successfully filtered, how many have not been filtered, and how many have been filtered incorrectly. The experimental results are shown in Figure 5 [22].

From Figure 5, it can be seen that the application of data science to filtering spam and malware is indeed very effective, with a success rate of up to 95%.

### 5.3 Experimental Summary

For the practical application of data science, this chapter conducts experiments on its application in filtering spam and malware, as well as in the sales industry, with the aim of verifying whether data science can use data for prediction and whether the prediction results are accurate. Divide the sales data of the three companies into two parts, the first half of the year and the second half of the year. Use data science to analyze the data of the first half of the year and predict the sales data of the second half of the year. Compare the predicted results with the actual results of the second half of the year for experimentation. Mix spam and normal emails, and mix malware and normal software to conduct filtering experiments using data science. Experiments have shown that data science can accurately predict, and the expected results of experiments in the sales industry are almost identical to the actual results. The filtering rate for spam and malware is as high as 95%, which can effectively distinguish and filter junk information.

## 6. Conclusions

For the application scenarios and practice research of data science in the context of BD, this paper first briefly introduces BD and data science and their relationship. Data science is an interdisciplinary discipline that has emerged from the development of BD. With the development of BD, data science has also developed along with it. However, the specific applications of data science and whether it is effective still need further research. This paper studies BD and data science, and finds that in theory, data science can also make predictive inference like BD, which can be applied to personalized content, navigation, sales service industry, spam filtering and malware filtering. Starting from practice, experiments were conducted on the specific application of data science to the sales industry and the filtering of junk information. The conclusion was drawn that the prediction accuracy of data science is very high, with a filtering rate of up to 95% when applied to the filtering of junk information. When applied to the sales industry experiment, the predicted results are almost identical to the actual results.

## References

*[1] Yu Canqing, Li Liming. Data science in large cohort studies. Chinese journal of Epidemiology, 2019, 40(1):1-4.*

*[2] Wang Kai, Zhang Shaojie, Ma Juan, et al. Study on spatial distribution and warning criteria of landslide macro displacement stage in big data environment. Advances in earth science, 2022, 37 (10): 1054-1065. The DOI: 10. 11867 / j. i SSN. 1001-8166. 2022. 042.*

*[3] Huo Cunxiao, Hou Yu. Research on product design for the elderly in the era of Big Data. Packaging Engineering, 2019, 040(012):147-150.*

*[4] Qiu Zixun, Zhou Yahong. Digital economy development and regional total factor productivity: Based on the analysis of national big data comprehensive pilot zone. Journal of Finance and Economics, 2021, 047(007):4-17.*

*[5] Yao Na, Wang Xiao, Yang Chuanjiang, et al. Big data grid control equipment in the technical support system monitoring research. Journal of hydroelectric and water conservancy, 2022, 6 (7): 65-67. The DOI: 10. 12238 / HWR v6i7. 4509.*

*[6] Qiu Huijun, Yuan Lianxiong, Huang Xuecun, et al. Study and analysis of symptom characteristics of Novel*

*coronavirus pneumonia based on Internet big data. Chinese otolaryngology head and neck surg, 2020, 55 (6): 569-575. The DOI: 10. 3760 / cma. J. c. n115330-20200225-00128.*
*[7] Jin Chensheng. Discussion on risks and preventive measures of small and micro enterprise credit business in commercial banks under the background of big data. Knowledge Economy, 2021, 590(023):11-12.*
*[8] Xing Luyu, Hu Runhong, Tang Chensong. Computer Knowledge and Technology, 2021, 017(009):244-246. Python Big Data Mining Guest experience in Yixian County, Anhui Province. Computer Knowledge and Technology, 2021, 017(009):244-246.*
*[9] Qu Jingquan. Challenges of macroeconomic analysis in the era of big data. Investment and Entrepreneurship, 2020, 031(017):26-28.*
*[10] Yu Canqing, Li Liming. Data science in large cohort studies. Chinese journal of Epidemiology, 2019, 40(1):1-4.*
*[11] Compiled by Zhang Juan. World Economic Forum released the report "Data Science in the new Economy". Research Information Technology and Application, 2019, 10(4):93-94.*
*[12] Yang Yin, Huang Yunqing, Liu Shaoyue. Data in local colleges of science and technology of data professional personnel training mode research. Journal of education modernization, 2019, 6 (4): 23-25. DOI: CNKI: SUN: JYXD. 0. 2019-04-007.*
*[13] Sun Jiling Li Xiaojing. Construction and innovation of Disability Statistics System on a new journey: Summary of the 6th Symposium on Disability Data Science and the Founding Conference of Disability Statistics Branch of China Statistical Society. Statistical Research, 2022, 39(2):158-160.*
*[14] Cheng Shuo, Liu Guifeng, Liu Qiong. When Library and information science meets Data science: intersection and expansion. Library Forum, 2022, 42(11):94-100.*
*[15] Singh A, Garg S, Kaur K,et al. Fuzzy-Folded Bloom Filter-as-a-Service for Big Data Storage in the Cloud. IEEE Transactions on Industrial Informatics, 2019, 15(4):2338-2348. DOI: 10. 1109/TII. 2018. 2850053.*
*[16] Huang Y, Sheng K, Sun W. Influencing factors of manufacturing agglomeration in the Beijing-Tianjin-Hebei region based on enterprise big data. Acta Geographica Sinica, 2022, 32(10): 2105-2128. DOI: 10. 1007/ s11442-022-2039-9.*
*[17] Diller G P, Baumgartner H. Impact of Adequate Provision of Care Models and Big Data Analysis for Adults with Congenital Heart Disease. Aktuelle Kardiologie, 2021, 10(05):403-407. DOI: 10. 1055/a-1556-0210.*
*[18] Wang T, Zheng Z, Rehmani M H,et al. Privacy Preservation in Big Data from the Communication Perspective — A Survey. IEEE Communications Surveys & Tutorials, 2019, 21(1):753-778. DOI:10. 1109/COMST. 2018. 2865107.*
*[19] Kulkarni A R, Kumar N, Rao K R. Efficacy of Bluetooth-Based Data Collection for Road Traffic Analysis and Visualization Using Big Data Analytics. Big Data Mining and Analytics, 2023, 6(2): 139-153. DOI: 10. 26599/ BDMA. 2022. 9020039.*
*[20] Trenti T, Pecoraro V, Pirotti T,et al. IgM anti-SARS-CoV-2-specific determination: useful or confusing? Big Data analysis of a real-life scenario. Internal and emergency medicine, 2021, 16(8):2327-2330. DOI:10. 1002/jmv. 26830.*
*[21] Hao L A, et al. Big data analysis of the internet of things in the digital twins of smart city based on deep learning. 2021, 128:167-177.*
*[22] Lv Z, Lou R, Li J, et al. Big Data Analytics for 6G-Enabled Massive Internet of Things. IEEE Internet of Things Journal, 2021, PP (99):1-1.*