Construction of Visualization Platform for Industrial Big Data Processing Based on Computer Technology

Hong-ya Wu*[1.2], Xiaoyu Zhou[1], Weijie Gu[1]

toni8110@163.com*

Changzhou Vocational Institute of Mechatronic Technology[1]
Jiangsu Internet of Things and Manufacturing Information Engineering Technology Research and Development Center (Jiangsu Changzhou 213164, China)[2]

Abstract: Based on computer technology and JavaWeb technology, a visualization platform for industrial big data processing and analysis is researched and designed. Users don't need to know the bottom details of the big data development platform, but only need to customize tasks on the web interface according to the data analysis requirements, then submit them to the platform for processing and wait for the results to be output. The differences between industrial big data and traditional data are analyzed. This paper briefly introduces how to use the machine learning algorithm of computer technology MLlib to analyze industrial big data. This paper introduces some technical problems and principles to be dealt with in the construction of visual platform, and formulates the process of platform construction. Finally, the visualization platform is tested and the problems in application research are summarized.

Keywords: Computer technology; Industrial big data; Visualization platform

1 Introduction

As the core big data analysis platform for testing and supervision, the platform environment of industrial big data processing and analysis research is complicated at present. The establishment of data processing platform and the programming of algorithms for testing and analysis all need professionals to execute, and the technical requirements are high, requiring operators to have certain technical knowledge. However, it is difficult for ordinary small and medium-sized enterprises to equip professional technicians in this line, so it is particularly important to make the data analysis platform easy to operate. We have designed the platform visually, shielding the bottom details of the platform functions, and the general staff can make good use of it with simple training. Users can customize the big data analysis tasks by themselves through the platform [1].

2 Analysis of Big Data Visualization Platform Technology

2.1 Introduction to Computer Technology

Computer technology is an open source cluster computing system based on in-memory computing, which was developed by AMP Lab using Scala language. At present, it is the top

ICBBEM 2023, May 19-21, Hangzhou, People's Republic of China

DOI 10.4108/eai.19-5-2023.2334314

open source project of Apache. Computer technology provides a faster and more universal data processing platform. Compared with Hadoop, computer technology can make your data analysis program run dozens or even hundreds of times faster. Some iterative algorithms that require machine learning, for example, are suitable for running on the computer technology platform [2].

The core of computer technology is composed of four libraries: SQL, Streaming, MLlib and GraphX, which can be simultaneously applied to one program. The big data analysis platform established in this paper is the machine learning library of computer technology MLlib, which provides a complete set of data analysis algorithms, such as classification and clustering, which can be used for industrial big data analysis. HDFS is a distributed file storage system in Hadoop ecosystem. It has many similarities with general distributed storage systems, but also has some advantages that other distributed storage systems do not have, the most important of which is its high fault tolerance, so that the system can be deployed on cheap machines, and the cost of deploying HDFS is not high [3].

At the same time, HDFS provides high-throughput data access, which is suitable for large-scale data sets. HDFS is selected as the file storage system in the development of industrial data platform.

2.2 Characteristics of Industrial Big Data

In the current data age, big data presents four characteristics. First, the data volume is huge, reaching PB level; Second, there are many types of data, which are expanded from traditional texts to various data formats; Third, the value density is low, and how to "purify" the huge data to obtain value is the key point; Fourthly, the processing speed is fast. With the decrease of hardware cost, the enhancement of performance and the introduction of distributed processing idea, the processing speed of these PB-level data is also very fast [4].

Industrial big data also has its own special features. First, the time dimension is not uniform, which is different from commercial big data. Industrial big data is collected by countless sensors that detect the condition of industrial equipment. As test and analysis data, we hope that the time dimension of the collected sensor data is consistent, that is, the original data can contain the values of all sensors at every time point. If the data at the time point is incomplete, it will inevitably affect the final data analysis results. However, in fact, in industrial equipment, due to many factors, such as temperature, aging of equipment, types of sensors, differences in sampling period, etc., it is difficult to ensure the consistency of time dimensions of the data collected by sensors, so we need to preprocess the original data and align the time dimensions. Industrial data are some straightforward data such as temperature, humidity, component speed, brightness, etc. collected by sensors, and it is difficult to reflect some conditions of equipment on the surface. Big data analysis needs to further process these data and convert them into meaningful data [5].

2.3 Introduction of classical algorithms

The core of data analysis is algorithm, and the selection and optimization of algorithm determines the quality of data analysis results, which is also the focus of data analysis. At present, the big data analysis platform chooses three simple algorithms for testing: linear regression algorithm, Bayesian classification algorithm and KMeans clustering algorithm. Naive Bayes classification algorithm is a relatively simple algorithm. The main idea is to solve the probability

of each category under the condition that the given item to be classified appears, and the one with the highest probability is the category to which the classified item belongs [6].

Bayesian algorithm in MLlib is mainly used for classification function, which classifies vectors according to different sample data, such as giving the following simple samples,as show in table 1:

Table 1 Test data set

0, 100	0,200	0, 300	0,400
1,010	1 ,020	1, 030	1, 040
2, 001	2,002	2, 003	1,004

In which each cell is a data sample, the first digit in each sample is the category, and the other three are the eigenvalues. After running through the Bayesian classification algorithm program in MLlib, a model will be generated according to the samples. Using this model, when we want to test which category a group of eigenvalue data, such as 500, belongs to, the computer returns a category 0, which means that the data with the eigenvalue of 500 belongs to category 0[7].

The simplest application of Bayesian classification algorithm in industrial equipment can make some preliminary judgments on industrial equipment: firstly, multiple groups of data collected from industrial equipment are obtained by sensors, then the sample data are classified according to the equipment health degree, the categories are marked into the sample data, and the data are modeled by Bayesian classification algorithm, so that a model about the status category of equipment corresponding to the data collected by sensors is obtained. Finally, the collected data is tested by this model to judge the equipment health degree [8].

3 Platform Building

3.1 Configuration of computer technology platform

The core of big data analysis platform lies in the establishment of computer technology platform, which is optimized based on Hadoop. Both of them are open source projects, and the installation packages can be downloaded in official website. First, configure the server. Two HPZ240 workstations are selected here, each equipped with Intel Zhiqiang series CPU, 32G memory, 128G solid state hard disk and 1T mechanical hard disk, which meet the hardware requirements of the platform. Each workstation OS is a LinuxCentos7 system, and the computer technology platform environment is deployed on Centos7. Firstly, JDK development environment and Scala development environment are installed, because the machine learning algorithm program used later is written in Scala, and the two hosts are set as a master machine and a salve machine [9]. IP subaddress is master：172.18.16.118, slave：172.18.16.107. Install computer technology on the master machine. This platform adopts the version of computer technology -1.6.3, which is free of installation. It only needs to be decompressed to the corresponding directory, then set the user name, modify the environment variables, and then configure the computer technology installation information. At the same time, install computer technology on slave machine in the same way. After the installation, start Hadoop cluster, start computer technology cluster and start Master and slave nodes respectively [10].

3.2 Visualization platform building

The visual interface program is written in Java, using JSP technology, and now there are many open source programming frameworks, which can save most of the coding time. Here, EOVA framework suitable for the project is selected. The program is uploaded on the HP server. The visualization platform transforms the user-defined data task into an instruction, which is submitted to the computer technology platform to run the data analysis task. The instruction includes computer technology connection request, task code, resource request path, etc. The visual webpage of the platform first displays the data insight of the platform's entrance, which is divided into four sub-modules: data modeling, model consulting, fault diagnosis and trend prediction [11].

Diagnosis and analysis is the ultimate goal of the platform, which is used to process the analysis data and judge the health status of the equipment. Diagnosis and analysis is to use the models in the model base to analyze the data. In the interface of fault diagnosis, the user selects the model and the time period of the diagnosis data, and then submits the task to run. All models generated in data modeling contain certain data description information. When a model is selected, the corresponding diagnostic data format should be consistent with the data format of the modeling. Here, the data characteristics specified by each model should be recorded[12]. As long as the model is selected, the characteristic values of the diagnostic data will be determined. When importing the diagnostic data set, only the time period of the data needs to be determined. Import the data set from MySQL into HBase, and select the corresponding analysis program. Here, a simple program is provided according to the model established by each algorithm. Submit the above tasks to Hadoop platform for execution, and save the results in HBase database and HDFS, and return them to the front page for display, the platform architecture process is shown in Figure 1[13].

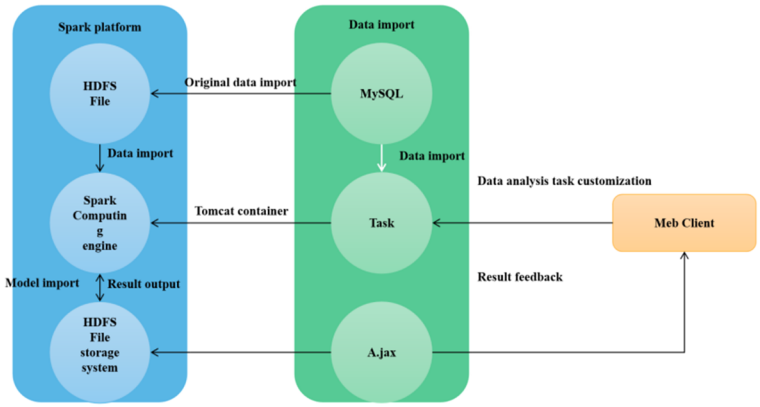

Figure 1 Platform architecture process

3.3 Script Processing

Give full consideration to different business script processing capabilities, and explain the names of each script, as shown in Table 2.

Table 2 Script Description

name	explain
Data Preload Scheduler- sh	Schedule usage in the overall scheduling script.
preP rocess-py	Handle the overall scheduling script
transferResult-py	Cache data to local files
bolCubeDefAgent- py Agent	Pull dimension mapping relation
dbUtil- py	Initialize data

4 Test

4.1 Visual Platform Startup

After the basic platform is built, Bayesian algorithm is selected to test the effect of the platform. First, the server-side computer technology platform serves, and then Hadoop environment, computer technology environment, Zookeeper environment of master/slave nodes and HBase service environment are started in turn. Second, start the Tomcat server, which is used to submit data analysis tasks to the computer technology platform. Here, all the environments on the whole server side have been loaded and started, and then the Javaweb program is running. The program is deployed on the HP server. In order to test the effect, we choose to run the web program locally. After the program is started, the whole visualization platform can start testing [14].

4.2 Platform testing

Enter 172.18.16.46: 81 on the browser to enter the visualization platform, enter the data insight module, and select data for modeling in data modeling. We select some data, such as data with ID values of 1, 2, and 3, and the data analysis period is from August 01, 2020 to August 31, 2020. The samples have been marked. 1.0 For the healthy state, select the Bayesian algorithm template for modeling, and the model is named "Bayesian Classification Model 1". Click the submit task button, and the modeling task will be executed after the submission. After the task is submitted, the computer technology platform processes the data modeling task, then saves the model on HDFS, and you can view the relevant information of the model by consulting the model in the interface bar. After the model is established, you can use the model to do data analysis. Click to enter the interface of fault diagnosis, and select the model to be used. Here, we select the newly established "Bayesian Classification Model I" model, select the data to be diagnosed, and select the data diagnosis from September 01, 2020 to September 02, 2020. Click the submit button, wait for some time, and then you can view the results below. If multiple 1.0' s are displayed, it means [15].

5 Conclusion

Through the establishment of the above-mentioned visualization platform and simple demonstration of technical analysis, it can be seen that the establishment process is complicated and involves a variety of complex technologies. When there is no visualization platform, it is still complicated to do data analysis. However, the establishment of the visualization platform

saves many operations and shields some underlying details. Users only need to customize their own data analysis task requirements on the interface, and the operation is very simple. At the same time, non-professionals can quickly master how to use it. Although the basic framework of the platform has been built, there are few configurable functions, and the focus of data analysis lies in the selection of algorithms. At present, the platform only provides some simple algorithms and can only make some simple data analysis. The value hidden in the data needs more and more perfect algorithms to mine. In addition, it is the preprocessing of collected data. After all, the environment of industrial equipment is complex and changeable, and there are many non-human uncontrollable factors that affect the acquisition of original data and the availability of data. Therefore, the preprocessing of raw data should eliminate these influencing factors. To sum up, the future improvement of big data visualization platform will focus on data preprocessing and application algorithm research. Kung Fu, gradually improve the platform function.

Acknowledgements

Jiangsu Construction Project of High Level Vocational Colleges

Jiangsu Higher Vocational Education Construction Project about High Level Backbone Professional

Project funded by Jiangsu “QingLan Project”

Changzhou Key Laboratory of High Technology Research（CM20183002）

References

[1] Jiang, L. . (2020). Retracted: research on the construction of new mode of mental health education for higher vocational college students based on computer big data technology. Journal of Physics: Conference Series, 1578(1), 012113 (4pp).

[2] Taniguchi, K. , Kubota, S. , & Yasumuro, Y. . (2022). Quantitative visualization of physical barriers for vulnerable pedestrians based on photogrammetry. Construction Innovation , 22(3), 604-623.

[3] Léon olde Scholtenhuis, Voordijk, H. , & Manen, M. V. . (2022). Empirically validating five propositions regarding 3d visualizations for subsurface utility projects. Engineering, Construction and Architectural Management, 29(6), 2535-2553.

[4] Sun, Y. , Zhao, J. , & Yin, W. . (2021). Construction of Big Data Security Early Warning Visualization Model for Smart Tourism,64(9),85-89.

[5] Di, X. U. . (2019). Construction of big data public opinion and judgment system based on spatial visualization. Information Science,65(3),77-79.

[6] Han, B. , Ma, J. W. , & Leite, F. . (2021). A framework for semi-automatically identifying fully occluded objects in 3d models: towards comprehensive construction design review in virtual reality. Advanced engineering informatics,6(50-), 50.

[7] Zhang, F. , Gao, Z. , Fan, R. , Fan, S. , Peng, X. I. , & Wang, M. . (2019). Construction and visual demonstration of the all-grid big data platform for dispatchingand controlling operation in state grid of hebei province. Power Systems and Big Data,5(9),98.

[8] Wang, S. , Min, D. U. , & Hospital, H. C. . (2019). Discussion on the construction of hospital medical data uploading platform under the background of big data. Microcomputer Applications,36(8),66-68.

[9] Yuan, Y. , Fuhua, J. I. , Zhao, B. , Jiang, H. , Wang, M. , & Fan, X. . (2019). Index method of farmland data based on solr and construction of big data platform. Transactions of the Chinese Society for Agricultural Machinery,76(2),47-49.
[10] Zi-Qian, L. I. , Zhu, Q. , Song-He, M. U. , & Zhang, Y. . (2019). Big data processing platform of smart grid based on cloud computing. Adhesion, 63(4), 65-69.
[11] Ren, Q. , & Wang, A. . (2019). Research on construction of college students' entrepreneurship and innovation platform based on green big data. Journal of Changchun University, 97(8), 34-35.
[12] Yanan, Q. U. . (2019). Construction of three-dimensional visual intelligent security three-dimensional protection comprehensive platform for large scientific research base. Mechanical and Electrical Equipment,51(7), 56-59.
[13] Juan, M. A. , Zhang, M. , Han, B. , Huang, Z. , Shi, A. , & Xue, Y. . (2019). Cloud platform construction for geological hazards' data from uav survey. The Chinese Journal of Geological Hazard and Control, 8(5), 55-57.
[14] Chang-Ji, H. U. , Dong, X. , Duan, C. Y. , & Polytechnic, F. . (2019). Construction of virtual training platform for photovoltaic power station. Computer Systems & Applications 31(1), 94-98.
[15] Yang, N. , Bai, H. J. , Zhang, Y. , Liang, Q. J. , Fan, Y. Y. , & Wang, C. R. . (2019). Analysis on the construction of chinese minority art plan platform under the background of "the belt and road". Journal of Dalian Minzu University, 33(6), 12-17.