Data Processing System Based on Computer

Peng Liao

1Enrollment and Employment Division, Shangrao Preschool Education College, Shangrao, Jiangxi, China

*Corresponding author e-mail: LiaoPeng@srygz.com

Abstract. With the popularization of computer and Internet, software technology, signal processing technology and real-time processing technology have been applied to all aspects of life, which has caused a surge of data. Therefore, big data (hereinafter referred to as BD) has become the focus of attention all over the world, which requires improving data application and processing technology. Through BD, countries can obtain corresponding knowledge, which will improve the software and comprehensive application of BD. With the in-depth study of deep learning algorithm, we can continuously improve the application of BD, which is a Data Processing (hereinafter referred to as DP) method with high precision, fast speed, flexible use and strong scalability. Through the DP system, we can realize the post demodulation and processing of signals in various equipment, which can achieve the state of data availability. In many ways, we can obtain the required knowledge through the DP system, which will improve the processing ability of the computer. Firstly, this paper puts forward the challenges faced by BD. Then, this paper analyzes the algorithms in DP. Finally, this paper designs a DP system.

Keywords: Computer, DP System, System Design

1. Introduction
With the continuous increase of network BD, BD has gradually developed in the direction of complexity, uncertainty and emergence. Therefore, data analysis system brings great challenges to it system architecture and computing power, which also gives birth to big DP frameworks, such as Google GFS, MapReduce, disco, Google Dremel, Dryad, etc [1]. Chinese Internet giants have also gradually deployed applications on Hadoop and other systems, which will design targeted computing models and algorithms according to different system applications. For example, storm streaming model is more suitable for real-time query of massive streaming data [2]. With the increase of the scale of big DP system, we must strengthen the research on platform independence, which can avoid the linkage with specific operating system and hardware platform [3-6]. Through a similar pipe protocol, we can replace the Java execution logic in the map and reduce phases with binary executables, which is also a step towards localization in China. Facing the challenges brought by BD, the research on computer-based DP system is of great significance [7].

2. Challenges of BD computing
2.1. Complexities of data
When dealing with computing problems, BD provides unprecedented large-scale samples. At the same time, people have to face more complex data objects. Therefore, the typical characteristics of BD are gradually transformed into diverse types and modes, complex relationships, and uneven quality [8]. The inherent complexity of BD makes many links such as data perception, expression, understanding and calculation face great challenges, which leads to the surge of computing complexity in the space-time dimension. At present, people lack understanding of the internal mechanism of BD complexity, including the distribution and collaboration of BD, and data complexity, which greatly restricts people's ability to design efficient computing models and methods of BD [9]. Therefore, we must formally or quantitatively describe the complexity of BD, which can improve the understanding of the essential characteristics of data and measurement indicators. By analyzing the complexity law of BD, we can understand the essential characteristics of the complex pattern of BD, which will obtain better knowledge abstraction. Therefore, the complexity of BD will guide the design of BD computing model and algorithm, which needs to clarify the reduction of BD on demand and reduce the complexity [10].

2.2. Computational complexity
BD has many complexities, including multi-source heterogeneity, huge scale, rapid and changeable, which makes the traditional methods of machine learning, information retrieval and data mining can not effectively support the processing, analysis and calculation of BD. BD calculation can no longer be like the calculation mode of small sample data set, which can no longer carry out statistical analysis and iterative calculation on global data [11]. Therefore, we must break through the independent and identically distributed data and the sufficiency of sampling in traditional computing. When solving BD problems, we need to re-examine and study the computability, computational complexity and solution algorithms of BD. We must change our view of the nature of data computing, which will provide a basic method for processing and analyzing BD. By supporting value driven applications in specific fields, we can establish a new computing paradigm for studying the computability of BD. Therefore, we need to break through the traditional data around machine computing based on the whole life cycle of BD, which needs to build a data centric push computing model. Through distributed and streaming computing algorithms, we can form a BD computing framework integrating and optimizing communication, storage and computing. Based on the local calculation and approximation methods of bootstrap and sampling, we can propose a new algorithm theoretical basis that does not depend on the total data [12].

2.3. System complexity
For big DP systems with different data types and applications, we can form a variety of platforms to support BD scientific research. In the face of BD with huge scale, complex structure and sparse value, we will face a variety of problems which poses new challenges to the system architecture, computing framework and processing methods of big DP system. Therefore, we put forward strict requirements for the operation efficiency of big DP system, which must have the characteristics of high efficiency. The big DP system needs to meet the basic criteria of design, implementation, testing and optimization, which is an important basis for building energy-efficient distributed storage [13]. The efficiency evaluation and optimization of big DP system is a great research challenge, which needs to comprehensively measure a variety of efficiency factors in the system. Therefore, people need to combine the value sparsity and weak locality of access of BD, which can effectively solve the problems of energy efficiency optimization and distributed storage architecture [14].

3. DP system algorithm
3.1. Introduction to k-means algorithm
Suppose dataset $D$ contains $n$ data objects, which are divided into $k$ clusters $C_1, \ldots, C_k$. The distance relationship between the data object $x \in C_j$ and the cluster center $C_i$ of the cluster is expressed by the $\text{dist}(x, C_i)$ distance function, as shown in Formula 1.

$$\text{dist}(x_i, x_j) = \sqrt{\sum_{m=1}^{m} (x_{im} - x_{jm})^2} \tag{1}$$

SSE is the threshold target to measure clustering completion, as shown in Formula 2.

$$\text{SSE} = \sum_{j=1}^{k} \sum_{x \in C_j} |x - C_j|^2 = \sum_{j=1}^{k} \sum_{x \in C_j} \text{dist}(x, C_j)^2 \tag{2}$$

3.2. Optimal feature selection method of cart classification tree algorithm

By using information gain ratio to select features, we can reduce the problem of large information gain caused by many eigenvalues. Assuming $k$ categories, the probability of the $k$ category is $p_k$, and the Gini coefficient expression of the probability distribution is shown in formula 3.

$$Gini(p) = \sum_{k=1}^{K} p_k (1 - p_k) = 1 - \sum_{k=1}^{K} (p_k)^2 \tag{3}$$

The Gini coefficient expression of sample $D$ is shown in formula 4.

$$Gini(D) = 1 - \sum_{k=1}^{K} \left( \frac{|C_k|}{|D|} \right)^2 \tag{4}$$

4. DP system

4.1. Overall architecture design based on Arion

Arion takes Hadoop as a compatible and improved object, which can separate data storage, task execution and scheduling management. Based on Hadoop pipe, we can expand a set of communication protocols between functional modules. This paper sets the overall framework diagram as Arion, as shown in Figure 1. Among them, the middle part is the task execution engine based on llvm and its adaptation layer. In the execution engine, the superoptimization instruction level can accelerate the engine to rely on platform technology and compile super optimization technology, which will mine optimized code fragments and improve the efficiency of the engine. The storage layer is the data management layer, which is the lock independent BD file system arionfs. By using lock independent mechanism, we can effectively solve the conflict in the process of data access, which will improve the concurrency of the system.
4.2. Task execution underlying optimization
The BD platform has application pertinence, and the calculation in the platform execution engine also has a certain mode. By compiling optimization technology, we can build an application specific optimization execution engine, as shown in Figure 2.

5. Conclusion
The traditional system has the problems of incomplete processing and low efficiency, which requires us to improve the operation efficiency of the system based on a variety of deep learning algorithms. Therefore, the DP system must have the characteristics of high precision, flexible use and strong adaptability, which can improve the overall application performance of various equipment. Facing the challenges brought by BD, the research on computer-based DP system is of great significance. The DP system designed in this paper can optimize the system processing speed, which can help some scholars for reference.

Acknowledgments
Design of data processing system based on computer.

References
[1] Cheng Xueqi, Jin Xiaolong, Wang yuanzhuo, Guo Jiafen, Zhang Tieming, Li Guojie. Overview of BD system and analysis technology [J]. Journal of software, 2014, 25 (09): 1889-1908.
[2] Fu Lin. Big DP system model and its application analysis [J]. Integrated circuit applications, 2018, 35 (12): 30-31.
[3] Gong Jia, Xie Quan, Wang Shengdong. Construction and application of naval combat big DP system [J]. Command information system and technology, 2015, 6 (02): 22-26.
[4] Li Fei, Liu Zhimin, Guo Jinyun, Li Yangyang. Comparative analysis of GNSS online DP system [J]. Global positioning system, 2016, 41 (06): 92-97 + 101.
[5] Li Xiaofei. Research on big DP system based on cloud computing technology [J]. Journal of Changchun Institute of Engineering (Natural Science Edition), 2014, 15 (01): 116-118 + 125.
[6] Liu Yongli, Liu Yunpeng. Research on autonomous closed-loop test and verification technology of real-time DP system [J]. Software engineering, 2021, 24 (03): 2-7.
[7] Tang Lin, He Tianyu. Design and implementation of natural language DP system based on Python [J]. Electronic technology and software engineering, 2018 (16): 160-162.
[8] Tang Yan, Li Faqin, Xiao Fengsheng, Zhuang Wei. Design of a DP system for automobile ride test [J]. Mechanical design and manufacturing, 2018 (05): 68-70.
[9] Wang Peng, Wang Xiaoliang, Zhang Zhe, Wu renbiao. Navigation flight DP system based on stored procedure [J]. Computer engineering and design, 2015, 36 (04): 1084-1089.
[10] Xiang Xiaojia. A design method and implementation of orthogonal decomposition big DP system [J]. Computer research and development, 2017, 54 (05): 1097-1108.
[11] Zhang Dong. UML modeling and implementation of telemetry DP system [J]. Electronic measurement technology, 2017, 40 (01): 52-56.
[12] Zhao Xiaoyuan, Wang Bo, Qiu Tao. Development and design of power marketing inspection system based on mobile terminal field operation [J]. Electronic design engineering, 2016, 24 (23): 62-64.
[13] Zhou Yaozong, Kuang Cuilin, Liu Ziping, He Ruiyao. Performance analysis of GNSS online PPP DP system [J]. Journal of navigation and positioning, 2018, 6 (03): 113-118.
[14] Zhu Jianzhang, Qin Qianqing. Research status and development trend of remote sensing BD [J]. Chinese Journal of image and graphics, 2016, 21 (11): 1425-1439.