Multi-dimensional data analysis technology of business application system based on Spark framework

Changchao Dong1*, Yanbin Jiao2, Youyong Chen1, Lanxian Feng1

1Fujian Yirong Information Technology Co., Ltd, Fuzhou, Fujian, 350001, China
2State Grid Information & Telecommunication Group Co., Ltd, Beijing, 102211, China
*E-mail: dongchangchao@sgitg.sgcc.com.cn

Abstract. The current multidimensional data analysis technology uses the data approximation to obtain deep data information in the cube, which requires repeated traversal of the full link data, resulting in low processing efficiency and unable to analyze a large amount of data in parallel. Therefore, a multi-dimensional data analysis technology based on Spark framework for business application system is proposed to optimize the defects of traditional technology. Spark technology is used to build a full-link data analysis framework, and the data warehouse logic is designed to build a full-link data warehouse. Genetic algorithm is used to mine the full-link data of the business system, and the data are stored in the data warehouse, and the data warehouse and the sequential structure neural network are established. In the sequential structure neural network, multi-dimensional data analysis is realized from multiple angles and aspects. The technical feasibility verification results show that the research technology shortens about 14.7% of the time on average, significantly improves the analysis efficiency and can conduct concurrent analysis on a large number of data, which is better than the application effect of traditional analysis technology.

1. Introduction

Business application system is a kind of management information system that provides information service support to relevant users for special application needs. Business applications need to control, record and manage the process and results of changes in the business requirements of the whole organization, so as to load and realize the handling and processing of various types of business. The core of the business application system is to realize the management of business changes. The data application processing of the business application system involves the whole link data including the physical layer, service layer, application layer and other levels. Therefore, when analyzing the full-link data of the business application system, efficient data information analysis is required while ensuring data integrity. When discovering the hidden value of data in business application systems, it is often necessary to observe and analyze the data from multiple specific perspectives. The multidimensional analysis technology helps analysts to flexibly and comprehensively analyze the data from multiple dimensions, sides, and levels to deeply explore the potential values hidden in the data [1].

Traditional relational data management techniques increase the computing power of individual nodes when dealing with very large data sets, mainly by means of vertical scaling (improving the computer hardware environment, such as adding CPUs and increasing memory). This approach requires high hardware equipment, is expensive, and has poor scalability, which is insufficient to meet the requirements of today's distributed business applications for multidimensional data analysis [2]. The data analysis in the literature [3] using the principle of aggregated approximate query processing exploits...
the correlation between data and provides an in-depth analysis of the association between different data dimensional features. This technique does not require prior annotation and is able to obtain deep information in the data through the implied similarities between the data. However, the technique is easy to fall into clustering loops in the clustering process, which leads to loss of data analysis results. The literature [4] uses two-dimensional matrix form for mathematical representation of data, which is often transformed into two-dimensional matrix processing even when multidimensional matrix problems are encountered. While the data dimensionality in the business application system is high and the data volume is large, the computational volume of reducing the full-link multidimensional data for dimensional processing is too large.

Spark is a big data processing framework with great versatility, fast computation speed, and also fast query and analysis capability, and the core of design is high-speed in-memory computation [5,6]. Spark can cache intermediate computation results in memory, and the computation process does not need to repeatedly read and write disk peripherals, thus avoiding a large number of I/O operations. With regard to the above discussion, in order to improve the processing efficiency of business application system on data and enhance the application effect of the system on data, this paper will be based on the multidimensional analysis technology of full-link data of business application system based on Spark framework.

2. Research on Multi-dimensional Data Analysis Technology of Full Link in Business Application System Based on Spark Framework

2.1. Build a Spark framework for full link data analysis

In a distributed environment, Spark cluster adopts a master-slave structure, where a node called Driver is responsible for central coordination, running the Driver program code written by the user and scheduling each Worker work node, which is also called Executor node, used to execute Executor code [7]. When the application runs, the Driver program executes the main method of the user program to create the SparkContext runtime environment and connects to the Cluster Manager such as YARN, Mesos. It assigns Worker Node to start the Executor according to the cluster environment, and then Spark gets the Executor processes on the nodes in the cluster, these processes can run computations and can read or store computational data in the local file system. Finally, the code of the computing task Task is sent to the Executor to run through SparkContext. The running architecture of Spark is shown in Figure 1.

Figure 1 Spark runtime architecture diagram

Spark eliminates the complexity of data storage and communication between nodes. It manages and allocates resources globally through Master. Work: does the real computation, while DAG orchestrates the overall process of data computation. The basic data processing process can first read the original data from the Hadoop file system HDFS to generate the initial RDD through the TextFileu function or other means of reading the file. In the process of reading, RDD is automatically divided into several partitions part1...parti (i is the total number of partitions) according to the access partitions of HDFS to realize parallelization. After that, feature engineering processing was carried out on the original data
to remove the feature variables that had no influence on the final results, and features of all dimensions were standardized. ZipWithIndex () function was used to dualize some feature vectors, and finally input feature vectors and input targets were formed to support the stacked model for integrated operation. Full-link data flow of business application system in Spark framework is shown in figure 2.

Figure 2 Data flow in Spark

2.2. Design of full link data warehouse for business application system
In the Spark framework for full-link data analysis built above, a full-link multidimensional data warehouse is designed to achieve the goal of improving the efficiency of data analysis in business applications while preventing the loss of raw data. The data in the data warehouse designed in this paper includes the original layer data, the offline intermediate layer data and the result layer data. The data in the raw layer is the log data parsed by Spark, the data in the offline intermediate layer is the data required by the recommendation system, and the data in the result layer is obtained by the core subsystem of the system after calculation. The data size and the computation size of these three layers are reduced in order [8].

The data warehouse design is carried out by determining the data warehouse logic according to the structure of each data table shown in the following content.

(1) The fact information table is used to maintain the basic information of business application system metadata in the data warehouse, which mainly includes: fact table ID, business line, fact table name, whether it is valid, creator, creation time, modification time and other important attributes. (2) The indicator information table is used to maintain the basic information of the indicators, which mainly includes: self-incrementing ID, fact table ID corresponding to the indicator field, whether it is published or not, creator, creation time, modification time and other important attributes. (3) The dimension information table is used to maintain the basic information of the dimension, mainly including: dimension ID, business line ID, dimension status, creator, creation time, modification time and other important attributes.

2.3. Multidimensional analysis of the whole link data of the system
Before establishing the mapping relationship between the data in the data warehouse and the sequence structure neural network, the data mining is performed using genetic algorithm and the mined data are stored in the constructed data warehouse. During the iterative process of the genetic algorithm, multi-fold cross-validation was introduced to mitigate the overfitting of individuals to the target and to ensure the robustness of individual fitness values; the results of linear and tree models were fused to enhance the robustness of feature combinations[9]. Then the fitness function of the genetic operator is as follows.
\[ F(x_1, x_2, \cdots, x_n) = \frac{\sum_{i=1}^{M} L(x_1, x_2, \cdots, x_n)}{cv} + \frac{\sum_{i=1}^{M} G(x_1, x_2, \cdots, x_n)}{cv}, j = 1, \cdots, M, n = 1, \cdots, N \] (1)

In the above formula, \( M \) represents the number of individuals in the population; \( N \) is the gene length of each individual; \( cv \) represents the number of cross-validation for each model; \( L(x_1, x_2, \cdots, x_n) \) represents the linear model. In this paper, its output value is derived from linear regression; \( G(x_1, x_2, \cdots, x_n) \) represents the tree model, whose output value in this paper is derived from the gradient iterative tree[10-12]. The selection operator of the genetic algorithm adopts the roulette algorithm, which directly represents the proportional relationship between the probability of an individual being selected and the individual fitness value, and is formulated as follows:

\[ P(x_1, \cdots, x_n) = \frac{F(x_1, \cdots, x_n)}{\sum_{i=1}^{M} F(x_1, \cdots, x_n)}, i = 1, \cdots, M \] (2)

In the iterative process of genetic algorithm mining, the mean square error (MSE) is used to measure the accuracy of data mining results. In order to ensure that the roulette selection operator is consistent with the evaluation index when selecting the optimal individual, the reciprocal operation of the adaptive value of each round of evolution of the individual is made as follows:

\[ F'(x_1, \cdots, x_n) = \frac{1}{F(x_1, \cdots, x_n)} \] (3)

With equivalence classes that use neurons to partition a rule profile, some operations can be defined to obtain the corresponding knowledge rules. The mapping support of the sequential structure neural network is defined as:

\[ \zeta = \frac{MAP(\psi_i)}{Card(X)} \] (4)

In the above equation, \( MAP(\psi_i) \) is the number of samples in the original domain of neuron mapping; \( Card(X) \) is the number of points in all of \( X \). The statistical index and grouping index between different business data are queried in the data multidimensional cube, and the information in the data is obtained from multiple angles and multiple sides through data slicing, slicing, rotating, drilling and other processing in the neural network. So far, the research on multi-dimensional data analysis technology of full link of business application system based on Spark framework has been completed.

3. Technical feasibility study

3.1. Verify content and verify environment settings

The method of literature [3], the method of literature [4] and the multi-dimensional analysis technology of business applications based on the Spark framework proposed in this paper are used as experimental methods to compare the response rate of data analysis and server response technology. Request data analysis technology compares the response rate of data analysis technology with the concurrent throughput of the server in response to requests. Select the business application system as the data analysis technology application object, and the database of the business application system is the MySql database.

Based on the Proxmox platform, a Hadoop cluster is built on CentOS7 operating system, including 1 master node (running namenode and jobtracker services) and 4 slave nodes (running datanode and tasktracker: services). The total memory size of the cluster is 40G. Build a business application system
on this cluster for experimental research. The artificial data includes the weighted network data set automatically generated by the program and the multi-dimensional attribute data set collected by the network operation node using the Perfmon tool.

3.2. Results and Discussion

Three multidimensional data analysis techniques are used to analyze and process the full-link data in the business application system, and the response efficiency and request response throughput of the three analysis techniques at different data scales are compared in figure 3 below.

By analyzing (a) in Figure 3, it can be seen that when the method in this paper analyzes the full link data in the application system, the response time of the technology does not increase significantly with the increase in the amount of data, and the technology is responding. The throughput of data being processed is rising. According to analysis (b), as the amount of data in the business application system increases, the technical response time of the method in [3] increases significantly, while the data throughput of the analysis technology remains relatively stable when analyzing data. (c) It can be seen that when the method of literature [4] is adopted, the larger the amount of data that needs to be analyzed, the longer the time for technical data analysis and processing. The higher the throughput. By comparing (a), (b), (c) three pictures, when using data analysis technology to process data in business application system, the time of this method is much shorter than that of the other two technologies. In addition, the data throughput of the method in this paper is much higher than that of the two comparison methods, indicating that the method in this paper can process a larger amount of data concurrently and has better performance.

Figure 3 Comparison of test results of data analysis technology

By analyzing (a) in Figure 3, it can be seen that when the method in this paper analyzes the full link data in the application system, the response time of the technology does not increase significantly with the increase in the amount of data, and the technology is responding. The throughput of data being processed is rising. According to analysis (b), as the amount of data in the business application system increases, the technical response time of the method in [3] increases significantly, while the data throughput of the analysis technology remains relatively stable when analyzing data. (c) It can be seen that when the method of literature [4] is adopted, the larger the amount of data that needs to be analyzed, the longer the time for technical data analysis and processing. The higher the throughput. By comparing (a), (b), (c) three pictures, when using data analysis technology to process data in business application system, the time of this method is much shorter than that of the other two technologies. In addition, the data throughput of the method in this paper is much higher than that of the two comparison methods, indicating that the method in this paper can process a larger amount of data concurrently and has better performance.
4. Conclusion
In order to improve the efficiency of data analysis and processing, this paper proposes a multi-dimensional data analysis technology based on the Spark framework for the entire link of the business application system, and verifies the feasibility of the technology in practical applications. In future research, the analysis and processing efficiency of this technology should be further improved based on the characteristics of the full-link data flow of the business application system.

Acknowledgment
This paper is funded by the Innovative Seed Fund of the overall project of the headquarters of the State Grid Information & Telecommunication Group Co, Ltd, China in 2020(Grant No. SGIT0000XMJS2000658).

References
[1] Rychtáříková Renata, Korbel Jan, Macháček Petr, et al. (2018) Point Divergence Gain and Multidimensional Data Sequences Analysis.[J]. Entropy (Basel, Switzerland), 20(2): 106.
[2] Xiaohui Lin, Yanhui Zhang, Chao Li, et al. (2019) A new data analysis method based on feature linear combination[J]. Journal of Biomedical Informatics, 94: 103173.
[3] Sadineni P K . (2020) SAMPLING BASED JOIN-AGGREGATE QUERY PROCESSING TECHNIQUE FOR BIG DATA[J]. Indian Journal of Computer Science and Engineering, Vol. 11(5):532-546.
[4] Wu Gang, Li Fei. (2020) A randomized exponential canonical correlation analysis method for data analysis and dimensionality reduction[J]. Applied Numerical Mathematics, 164:101-124.
[5] Ferdaous Jenhani, Mohamed Salah Gouider, Lamjed Ben Said. (2019) Streaming Social Media Data Analysis for Events Extraction and Warehousing using Hadoop and Storm: Drug Abuse Case Study[J]. Procedia Computer Science, 159: 1459-1467.
[6] Suan Lee, Seok Kang, Jinho Kim, et al. (2019) Scalable distributed data cube computation for large-scale multidimensional data analysis on a Spark cluster[J]. Cluster Computing, 2(1): 2063-2087.
[7] D. Kim, N. Park. (2018) Spark framework based on a heterogenous pipeline computing with open CL[J]. The transactions of The Korean Institute of Electrical Engineers, 67(2):270-276.
[8] Mustapha Belouch, Salah El Hadaj, Mohamed Idhammad. (2018) Performance evaluation of intrusion detection based on machine learning using Apache Spark[J]. Procedia Computer Science, 127: 1-6.
[9] Dario Cottafava, Giulia Sonetti, Paolo Gambino, et al. (2018) Explorative Multidimensional Analysis for Energy Efficiency: DataViz versus Clustering Algorithms[J]. Energies, 2018, 11(5):1312.
[10] Park K , Peng L . (2018) A development of LDA topic association systems based on spark-hadoop framework[J]. Journal of Information Processing Systems, 14(1):140-149.
[11] Wang Zhibin, Zhang Shouzhen, Yan Jian, et al. (2021) Technical Route and Application Data Analysis of New Energy Vehicle[J]. Journal of Physics: Conference Series, 1813(1): 012049.
[12] Sergio Villamayor-Tomas, Christoph Oberlack,Graham Epstein, et al. (2020) Using case study data to understand SES interactions: a model-centered meta-analysis of SES framework applications[J]. Current Opinion in Environmental Sustainability, 44 : 48-57.