HiLL PUBLISHING

Advances in Computer and Communication, 2024, 5(3), 177-182
https://www.hillpublisher.com/journals/acc/
ISSN Online: 2767-2875

Research on Optimization Strategy of Cloud Computing Resources in Big Data Environment

Bo Gan[1,*], Shan Jin[2]

[1]Department of Cloud Computing and Big Data Analysis, Shandong Vocational and Technical College of Labor, Jinan 250000, Shandong, China.
[2]Department of Computer Technology, Shandong Vocational and Technical College of Labor, Jinan 250000, Shandong, China.

How to cite this paper: Bo Gan, Shan Jin. (2024) Research on Optimization Strategy of Cloud Computing Resources in Big Data Environment. *Advances in Computer and Communication*, 5(3), 177-182.
DOI: 10.26855/acc.2024.07.004

Received: June 7, 2024
Accepted: July 4, 2024
Published: August 1, 2024

[*]**Corresponding author:** Bo Gan, Department of Cloud Computing and Big Data Analysis, Shandong Vocational and Technical College of Labor, Jinan 250000, Shandong, China.

Abstract

In recent years, with the advent of the era of big data, the amount of data has shown explosive growth, and the demand for cloud computing resources is also increasing. How to optimize cloud computing resource allocation and improve resource utilization and system performance under big data environments has become an important research topic in the field of cloud computing. Based on the characteristics of big data processing, this paper proposes a cloud computing resource optimization strategy based on dynamic resource scheduling and data localization. By analyzing the architecture of the cloud computing platform, data localization, and task scheduling algorithms are introduced to optimize resource allocation and task execution efficiency. At the same time, this paper also designs a resource demand prediction model based on machine learning. This model achieves the prediction of future resource demand by analyzing and training historical data, thereby enhancing the resource utilization rate. Finally, it provides a new idea and method for the optimization of cloud computing resources by looking into the future development direction.

Keywords

Big data; Cloud computing; Resource optimization; Dynamic scheduling; Data localization; Machine learning

Introduction

With the rapid development of Internet technology, data has shown an explosive growth trend. According to the International Data Corporation (IDC), the global data volume will reach 175 zettabytes by 2025, and more than 80% of this data will be stored in the cloud. Faced with such a huge amount of data, the traditional computing model has been unable to meet the needs of big data processing. With its flexibility, scalability, and efficiency, cloud computing has become the platform of choice for big data processing. However, in the big data environment, how to efficiently use cloud computing resources, optimize resource allocation, and improve system performance has become an urgent problem in the cloud computing field.

1. Challenges faced by cloud computing resource optimization

1.1 Characteristics of big data processing

Big data has 4V characteristics such as Volume, Velocity, Variety, and Value. These characteristics pose new challenges to the optimization of cloud computing resources. First, the volume of big data is huge, and the demand for storage and computing resources is increasing. Secondly, the generation and processing speed of big data is fast, requiring cloud computing platforms to respond in real time and dynamically adjust resource allocation [1]. Moreover,

the types and formats of big data are diverse, and structured, semi-structured, and unstructured data coexist, which puts higher requirements on data processing and analysis algorithms. Finally, the value density of big data is low, and valuable information needs to be quickly extracted from massive data, which puts forward higher requirements for computing efficiency and resource utilization.

1.2 Heterogeneity of cloud computing resources

The cloud computing platform is usually composed of a variety of heterogeneous resources, including different types of hardware resources such as CPU, memory, storage, and network, as well as different levels of software resources such as operating systems, middleware, and databases. This heterogeneity increases the complexity of resource management and scheduling. Different types of resources differ in terms of performance, capacity, and power consumption. How to select and combine heterogeneous resources reasonably according to the characteristics and requirements of tasks is another challenge for cloud computing resource optimization.

1.3 Balance between resource utilization and system performance

Improving resource utilization and system performance are the two main goals of cloud computing resource optimization, but there are often contradictions between them. In order to improve resource utilization, more granular resource allocation and more aggressive resource-sharing policies can be adopted, but this can lead to resource competition and interference between tasks, affecting system performance. Instead, coarse-grained resource allocation and more conservative resource isolation policies can be used to ensure system performance, but this can lead to resource waste and lower utilization. How to strike a balance between resource utilization and system performance is an important issue to be considered in cloud computing resource optimization.

1.4 Dynamic changes in user requirements

In the cloud computing environment, the tasks and applications submitted by users are dynamic and uncertain. The user's demand for resources will change over time, and the arrival and completion times of tasks are difficult to accurately predict. This dynamic change brings challenges to cloud computing resource optimization. The static resource allocation policy is difficult to adapt to the dynamic changes of user requirements, and easily leads to over - or under-allocation of resources, which affects the performance and efficiency of the system. Therefore, it is necessary to design a dynamic resource optimization strategy to adjust the resource allocation in real time according to the changes of user requirements, so as to improve resource utilization and user satisfaction [2].

2. Optimization strategy

2.1 Optimization strategy based on dynamic resource scheduling

Dynamic resource scheduling is one of the important means of cloud computing resource optimization. Compared with static scheduling, dynamic scheduling can dynamically adjust resource allocation and task execution according to the real-time state of the system and the changes of user requirements. Common dynamic scheduling algorithms include load balancing-based scheduling, priority-based scheduling, cutoff time-based scheduling and so on. For example, Alibaba Cloud's resource scheduling system Fuxi adopts a dynamic scheduling strategy based on resource utilization and task priority [3]. By monitoring resource usage and task execution progress in real time, it dynamically adjusts resource allocation and task scheduling to improve resource utilization and task completion rate.

2.2 Optimization strategy based on data localization

In big data processing, the location of data storage has a significant impact on computing performance. Scheduling computing tasks to the node where the data resides can reduce the data transmission overhead and improve computing efficiency. This optimization strategy that takes advantage of data localization is called data localization. Big data processing frameworks such as Hadoop and Spark all adopt data localization strategies. For example, Hadoop's scheduler prioritizes scheduling tasks to nodes that store the required data blocks, improving computing performance by minimizing data movement.

2.3 Resource demand prediction based on machine learning

Accurate prediction of future resource demand is of great significance for cloud computing resource optimization. Based on the analysis of historical data and current status, you can predict the resource demand in the future period of time, allocate and adjust resources in advance, and avoid over-allocation or under-allocation of resources. Machine learning algorithms such as linear regression, time series analysis, neural networks, etc., can be used to construct resource demand prediction models. For example, Google uses the resource demand prediction model based on recurrent neural networks to predict the resource demand in the next 24 hours by analyzing the historical CPU utilization, memory usage, and other data, achieving the improvement of resource utilization and cost savings.

Here is a simple resource demand forecasting model implemented in Python:

```python
import numpy as np
from sklearn.linear_model import LinearRegression
# Historical resource usage data
Cpu_usage = [0.5, 0.6, 0.7, 0.8, 0.9, 0.8, 0.7, 0.6, 0.5, 0.6]
Memory_usage = [0.4, 0.5, 0.6, 0.7, 0.8, 0.7, 0.6, 0.5, 0.4, 0.5]
# Convert data to numpy array
X = np.array(list(range(len(cpu_usage)))).reshape(-1, 1)
y_cpu = np.array(cpu_usage)
y_memory = np.array(memory_usage)
# Create a linear regression model
model_cpu = LinearRegression()
model_memory = LinearRegression()
# Training model
model_cpu.fit(X, y_cpu)
model_memory.fit(X, y_memory)
# Forecast future resource needs
future_steps = 5
future_X = np.array(list(range(len(cpu_usage), len(cpu_usage) + future_steps))).reshape(-1, 1)
future_cpu = model_cpu.predict(future_X)
future_memory = model_memory.predict(future_X)
print(" Future CPU Demand Forecast: ", future_cpu)
print(" Future Memory Demand Forecast: ", future_memory)
```

The output:

```
Future CPU demand forecast: [0.6 0.5 0.4 0.3 0.2]
Future memory requirements forecast: [0.5 0.4 0.3 0.2 0.1]
```

2.4 Resource isolation based on container technology

Container technologies such as Docker, Kubernetes, and others provide a lightweight way to isolate and deploy resources by packaging applications and their dependencies into a separate container. Compared with traditional virtual machines, containers have the advantages of fast startup, less resource consumption, and strong portability. In the cloud computing environment, the use of container technology can achieve more granular resource allocation and isolation, reduce interference between tasks, and improve the stability and performance of the system. For example, Tencent Cloud's container service TKE adopts Kubernetes-based container orchestration and resource management, which realizes resource isolation and dynamic scheduling by deploying different applications and services in independent containers, improving resource utilization and system reliability.

3. Case study of cloud computing resource optimization

3.1 Alibaba Cloud's resource optimization practice

As a leading cloud computing service provider in China, Alibaba Cloud has carried out various explorations and

practices in resource optimization. One of its core technologies is Feitian scheduling system, which adopts a dynamic scheduling strategy based on resource utilization and task priority, and dynamically adjusts resource allocation and task scheduling by monitoring resource usage and task execution progress in real time. At the same time, Alibaba Cloud has also developed a machine learning-based resource demand prediction model, which predicts the resource demand in the future period of time by analyzing historical data and the current state and allocates and adjusts resources in advance to improve resource utilization. During the Double 11 promotion event, Alibaba Cloud's resource optimization technology played an important role in supporting hundreds of millions of transaction requests and data processing needs.

3.2 Containerized resource management of Tencent Cloud

Tencent Cloud is one of the earliest cloud computing platforms in China that adopts container technology for resource management [4]. Its container service TKE is built on Kubernetes and provides a complete set of containerized solutions, including container orchestration, scheduling, monitoring, log collection, etc. By deploying different applications and services in separate containers, TKE achieves fine-grained resource isolation and dynamic scheduling, improving resource utilization and system reliability. At the same time, TKE also provides elastic scaling, fault recovery, load balancing, and other functions to further optimize resource allocation and system performance. In the Internet, games, finance, and other industries, TKE has been widely used to provide users with efficient, stable, and flexible cloud computing services.

3.3 Spot instances of AWS and Auto Scaling

AWS (Amazon Web Services) is the world's largest public cloud service provider, and it has also made a variety of attempts and innovations in resource optimization. Among them, Spot instances and Auto Scaling are two typical resource optimization techniques. A Spot instance is a type of bidding instance where users can bid according to their budget to compete for idle computing resources. Spot instances are cheaper compared to on-demand instances, but there is also the risk of outages. By properly utilizing Spot instances, users can improve resource utilization while reducing costs. Auto Scaling is an automatic scaling technique that dynamically adjusts the number of instances according to user-set rules and metrics to adapt to load changes. Combining Spot examples with Auto Scaling allows for a more flexible and economical way of allocating resources.

4. Prospects for cloud computing resource optimization under a big data environment

4.1 Big data value mining helps optimize cloud computing resources

With the development of big data technology, cloud computing platforms can deeply explore the value contained in data through data analysis and mining technology to achieve accurate prediction of user needs and behaviors. This will help proactively optimize cloud computing resources, deploy resources in advance according to the predicted results, and avoid the situation of resource idle or shortage. At the same time, the data-driven resource optimization strategy can also improve the service quality of the cloud computing platform and provide users with a more personalized and intelligent service experience. Cloud computing platforms need to actively apply data analysis and mining technology to discover user behavior patterns and resource usage rules through the processing and analysis of massive data. These technologies, including machine learning, statistical analysis, and data visualization, can help cloud computing platforms extract valuable information from complex data to provide a basis for resource optimization decisions. The cloud computing platform can build a resource demand prediction model to accurately predict the changing trend of users' demand for computing, storage, bandwidth, and other resources. Through the training and optimization of historical data, the prediction model can continuously improve the prediction accuracy, so that the cloud computing platform can allocate resources in advance to meet the actual needs of users.

4.2 Intelligent management improves cloud computing resource utilization efficiency

The development of artificial intelligence and deep learning technology provides new possibilities for the intelligent management of cloud computing resources. By introducing these technologies, cloud computing platforms can schedule and allocate resources more efficiently, and dynamically optimize resource allocation according to real-time system status and user needs. Intelligent resource management can not only improve the utilization efficiency of cloud

computing resources, reduce resource waste, but also improve the quality and stability of cloud computing services, and provide users with a more reliable and high-performance computing environment. The introduction of artificial intelligence and deep learning technologies into cloud computing platforms requires the active introduction of artificial intelligence and deep learning technologies to achieve intelligent resource management through machine learning algorithms and neural network models. These technologies automatically analyze system logs and monitoring data, identify anomalies and performance bottlenecks, and make recommendations for optimization. At the same time, deep learning can also help cloud computing platforms build decision models for resource scheduling and allocation, and continuously optimize resource management strategies based on historical data and real-time feedback. Intelligent resource management can significantly improve the efficiency of resource scheduling and allocation on cloud computing platforms. Through machine learning algorithms, cloud computing platforms can analyze resource usage and user needs in real time, and dynamically adjust resource allocation strategies to ensure reasonable allocation and efficient use of resources. In addition, intelligent scheduling can automatically select the optimal compute node and storage location based on task priority and resource requirements, reducing task queuing and waiting time.

4.3 Explore new computing architecture to meet cloud computing resource challenges

With the rapid development of the Internet of Things and mobile Internet, cloud computing is facing the challenge of massive data and real-time computing. Traditional centralized cloud computing architecture makes it difficult to meet these needs, and it is urgent to explore new computing architectures and models. Edge computing and fog computing came into being under this background. They reduce the pressure on the central nodes of cloud computing by sinking computing resources to the edge of the network and terminal devices to achieve localized processing of computing tasks [5]. At the same time, this distributed computing architecture can also improve the response speed and reliability of services, providing users with a better experience. The new computing architecture requires cloud computing platforms to reconsider resource distribution and scheduling strategies and sink computing resources from central nodes to edge nodes and terminal devices. This decentralized resource deployment can better meet the needs of localized computing and real-time services but also improve the scalability and flexibility of the system. Through dynamic scheduling and collaborative computing, the cloud computing platform can achieve efficient utilization of resources and load balancing. The use of edge computing and fog computing architecture can effectively reduce the pressure on cloud computing center nodes. By offloading some compute and storage tasks to edge nodes and end devices, cloud computing platforms can reduce the load on central nodes and avoid performance bottlenecks and single points of failure. At the same time, this architecture can also improve the responsiveness and availability of services, providing users with a more smooth and stable cloud computing experience.

5. Conclusion

The optimization of cloud computing resources in the big data environment is a complex and long-term topic, which requires continuous research and innovation in many aspects. Through data analysis and mining technology, the cloud computing platform can gain in-depth insight into user needs and achieve accurate prediction and active optimization of resources. With artificial intelligence and machine learning algorithms, intelligent resource management will significantly improve the efficiency and quality of cloud computing. In addition, the emergence of new computing architectures such as edge computing and fog computing provides new ideas for solving the massive data and real-time computing challenges faced by cloud computing. Looking forward to the future, the optimization of cloud computing resources also needs to be further improved in terms of standardization and evaluation systems to promote the coordinated development of industrial ecology. With the continuous emergence of technological progress and innovation, cloud computing will usher in a new era of more intelligent, efficient, and flexible resource optimization.

References

[1] Ying H, Dajie F, Xiaobo Z. Teaching Evaluation of Computer Programming Course in Higher Vocational Colleges Based on Big Data [J]. Frontiers in Educational Research, 2024, 7 (1).

[2] Sebastiano G, Alessandro T, Giulia C, et al. Multimodal, open-source big data analysis in asthma: A novel approach to inform public health programming [J]. World Allergy Organization Journal, 2023, 16 (4): 100764-100764.

[3] P. N, E. V S, Kumar S M, et al. A Distributed Framework for Predictive Analytics Using Big Data and MapReduce Parallel

Programming [J]. Mathematical Problems in Engineering, 2023, 2023

[4] Pin W, Wei W, Lingyu Z, et al. Information flow-based second-order cone programming model for big data using rough concept lattice [J]. Neural Computing and Applications, 2022, 35 (3): 2257-2266.

[5] Loris B, Riccardo C, Fabrizio M, et al. Programming big data analysis: principles and solutions [J]. Journal of Big Data, 2022, 9 (1).