Journal of Education, Teaching and Social Studies
ISSN 2642-2336 (Print) ISSN 2642-2328 (Online)
Vol. 6, No. 2, 2024
www.scholink.org/ojs/index.php/jetss

Original Paper

Research on the Application of Cloud Computing Technology in Computer Big Data Analysis

Hongzhi Huang

School of Computer and Software Engineering, Xihua University, Chengdu, Sichuan 610039, China

Received: May 3, 2024 Accepted: May 15, 2024 Online Published: May 22, 2024

doi:10.22158/jetss.v6n2p141 URL: http://dx.doi.org/10.22158/jetss.v6n2p141

Abstract

This paper discusses the application of cloud computing technology in big data analysis. With the explosive growth of data volume, traditional data processing methods have been unable to meet the demand. Cloud computing technology provides a new solution for big data service analysis with its powerful computing power, elasticity and cost-effectiveness. This paper will introduce the basic concepts and characteristics of cloud computing, explore its specific application in big data analysis, and analyze its challenges and future development direction.

Keywords

Cloud computing, Big data analysis, Distributed computing, Elastic computing, Data

1. Introduction

With the rapid development of information technology, the speed of data generation and collection has increased exponentially. The advent of the big data era has made data analysis more complex and difficult. Traditional computing and storage architectures can no longer meet the needs of such massive data processing. The rise of computing technology provides an ideal platform for big data analysis. Through cloud computing, computing resources can be flexibly adjusted to achieve data storage and processing, further promoting the rapid development of data analysis.

2. Basic Concepts and Characteristics of Cloud Computing

2.1 Definition of Cloud Computing

Cloud computing is an Internet-based computing model that provides computing resources (such as computing power, storage space, application software, etc.) to users in the form of services through the network. Users can use resources in layers as needed, thereby consuming the underlying resource infrastructure.

2.2 Characteristics of Cloud Computing

• Elasticity and scalability: Cloud computing can dynamically adjust resource configuration according to user needs to achieve iterative expansion of computing and storage capabilities.

• High availability and reliability: Cloud computing provides high-availability services, and ensures service continuity and data security through failure and failover mechanisms.

• Secondary service: Users only need to pay for the resources they actually use, reducing initial investment and operating costs.

• Virtualization technology: Cloud computing uses virtualization technology to achieve efficient utilization and management of resources, making the allocation of efficient computing resources more flexible.

3. Application of Cloud Computing in Big Data Analysis

3.1 Data Storage

Big data analysis storage requires a large amount of reconstructed and non-reconstructed data. Cloud computing provides elastic data storage solutions, such as Amazon's S3 (Simple Storage Service) and Google's Cloud Storage. These services not only provide massive space, but also support fast access and reading of data.

3.1.1 Challenges of Data Storage

In big data analysis, storing massive data is a huge challenge. Traditional storage systems often have difficulty coping with the explosive growth of data. Whether it is formatted data (such as database records) or unformatted data (such as text, images, videos), storage and management strategies are required. In addition, data security and efficient availability are also key issues.

3.1.2 Cloud Storage Solutions

Cloud computing provides flexible storage solutions. Taking Amazon S3 as an example, S3 is an object storage service that supports the storage and management of massive data. The storage capacity of S3 can be almost unlimited, and users can expand storage space at any time as needed. It also provides guarantees of high durability and high availability, ensuring the security and reliability of data through multi-region replication and important data mechanisms.

Google Cloud Storage is also a similar service that provides global data storage and access. It supports multiple storage types, including standard storage, nearline storage, and cold storage, to meet different data access needs. Google Cloud Storage also provides powerful security features, such as data encryption and access control, to protect the privacy and security of user data.

3.1.3 Data Read and Write Performance

Cloud storage services not only provide massive storage space, but also support fast data read and write. Taking Amazon S3 as an example, it achieves massive data access through optimized data sharding and load balancing mechanisms. Whether it is data read or write operations for large-scale data access, S3 can provide excellent performance.

Google Cloud Storage also has high-performance data access capabilities. Through intelligent data distribution and storage mechanisms, it can quickly respond to user data requests. Especially in big data analysis scenarios, fast data read and write are key factors to ensure analysis efficiency.

3.1.4 Future Development Trends

With the development of big data technology, cloud storage solutions are also developing. For example, Amazon's Glacier and Google's cold storage services are specially designed for long-term archiving and low-frequency access data, providing external storage costs. In addition, the combination of distributed file systems and object storage technologies will further enhance the performance and flexibility of cloud storage.

3.2 Data Processing

Cloud computing platforms achieve efficient batch processing of large-scale data through multiple computing frameworks (such as Hadoop and Spark). Hadoop's MapReduce programming model and HDFS (Hadoop Global File System) allow users to store and process data in the cloud. Spark further improves the speed of data processing through in-memory computing.

3.2.1 Hadoop's MapReduce and HDFS

Hadoop is an open source big data processing framework, mainly composed of the MapReduce programming model and HDFS. MapReduce is a distributed computing model that achieves batch processing of large-scale data by layering tasks into batch execution of multiple small tasks. HDFS is Hadoop's circular file system, which is responsible for distributing and storing data on multiple nodes, providing high reliability and high throughput data storage and access.

On cloud computing platforms, Hadoop's advantages are more obvious. Cloud services usually provide pre-configured Hadoop clusters that users can quickly deploy and use. Amazon's EMR (Elastic MapReduce) and Google's Dataproc are typical examples. Users can easily run Hadoop jobs in the cloud without worrying about the underlying infrastructure management.

3.2.2 Spark's In-memory Computing

Spark is a more modern big data processing framework. Compared with Hadoop's MapReduce, Spark uses in-memory computing, which greatly improves the data processing speed. The core of Spark is RDD (Resilient Distributed Dataset), a diversified data structure that supports efficient computing and fault tolerance of in-memory data.

In the cloud computing environment, Spark's performance advantage is more prominent. Cloud services usually provide optimized Spark clusters to support the rapid processing of massive data. For example, Amazon's EMR and Google's Dataproc both support Spark, and users can use the powerful computing power of the cloud platform to accelerate the execution of data processing tasks.

3.2.3 Data Processing Optimization Strategies

In order to further improve the efficiency of data processing, cloud computing platforms provide a variety of optimization strategies. For example, through automatic scaling, cloud platforms can dynamically adjust computing resources according to changes in workload mechanisms to ensure

efficient execution of processing tasks. Data localization is also an important optimization strategy. By arranging computing tasks on nodes that store data, the overhead of data transmission is reduced.

In addition, cloud computing platforms also provide a wealth of data processing tools and libraries to support a variety of data processing and analysis tasks. For example, Amazon's Glue and Google's Dataflow provide preconfigured data processing pipelines, allowing users to quickly build and run complex data processing workflows.

3.2.4 Future Development Trends

With the continuous development of big data processing technology, the data processing framework on cloud computing platforms is also evolving. For example, the architecture based on containers and microservices makes data processing more flexible and efficient. In the future, with the maturity of edge computing and stream processing technologies, the data processing capabilities of cloud computing platforms will be further enhanced, providing more flexible solutions for efficient analysis of big data.

3.3 Data Analysis

The cloud computing platform provides a wealth of data analysis tools and services. For example, Amazon's EMR (Elastic MapReduce) and Google's BigQuery, users can use these services to perform complex data queries and analyses. Through the cloud computing platform, data scientists and analysts can quickly perform data mining, machine learning model training, and result visualization.

3.3.1 Data Query and Analysis Tools

The cloud computing platform provides a variety of data query and analysis tools to meet different data analysis needs. Taking Amazon's EMR as an example, it not only supports Hadoop and Spark, but also integrates a variety of data query tools such as Hive, Pig, and Presto. Users can use these tools to query and analyze massive amounts of data and discover patterns and trends that are quickly hidden in the data.

Google's BigQuery is a high-performance data warehouse service designed for large-scale user data analysis. Through SQL syntax, PB-level data can be quickly queried and analyzed. The advantage of BigQuery is its serverless architecture, which does not require users to care about computing resource configuration, and mainly focuses on data analysis itself.

3.3.2 Data Mining and Machine Learning

The cloud computing platform also provides a wealth of data mining and machine learning tools to help users extract valuable information from data. Amazon's SageMaker and Google's AI platform are major machine learning platforms that provide full-process support from model training to model deployment to obtain valuable information from data.

SageMaker supports a variety of machine learning algorithms and frameworks, such as TensorFlow, PyTorch, scikit-learn, etc., which users can flexibly choose and use. SageMaker also provides automated model training and tuning functions, reducing the technical resources of machine learning.

Google's AI platform also supports a variety of machine learning frameworks and provides powerful

model management and deployment functions. Through the cloud computing platform, data scientists and engineers can quickly train and deploy machine learning models to achieve intelligent analysis and application of large-scale data.

3.3.3 Result Visualization

The ultimate goal of data analysis is to convert analysis results into visual information to help decision makers better understand and use data. The cloud computing platform provides a variety of data visualization tools, such as Amazon's QuickSight and Google's Data Studio, which allow users to easily create and share data reports and dashboards.

QuickSight is a cloud-based business intelligence service provided by Amazon that supports a variety of data sources and visualization types. Users can quickly create data visualization charts and dashboards through drag-and-drop operations and share analysis results with team members.

Google's Data Studio provides a wealth of visualization components and templates, supporting real-time data updates and interactive analysis. Through Data Studio, users can present analysis results to different audiences, helping them better understand the stories behind the data.

3.3.4 Future Development Trends

With the continuous advancement of data analysis technology, data analysis tools on cloud computing platforms are also constantly being equipped. For example, automated data analysis (AutoML) technology based on artificial intelligence makes data analysis more intelligent and efficient. In the future, as the demand for data development and analysis continues to grow, cloud computing platforms will provide more comprehensive and flexible data analysis solutions to promote digital transformation and innovative development in various industries.

4. Advantages of Cloud Computing Technology

4.1 Cost-effectiveness

The auxiliary use and payment model of cloud computing greatly reduces the cost of data analysis. Traditional IT infrastructure construction requires enterprises to invest a lot of money to purchase and maintain expensive hardware equipment, including servers, storage devices and network equipment. This not only increases the initial investment, but also brings long-term maintenance and management costs. The emergence of cloud computing has changed this situation.

4.1.1 Reduce Initial Investment

The subsequent service model of cloud computing makes enterprises spend a lot of money to purchase hardware equipment. Cloud services such as Amazon AWS, Microsoft Azure and Google Cloud provide rich computing, storage and network resources, and enterprises can flexibly use and rent these resources according to their needs. This model not only reduces venture capital, but also enables enterprises to start data analysis projects more quickly.

4.1.2 Reduce Operating Costs

In traditional IT environments, enterprises need dedicated IT teams to maintain and manage hardware

equipment, including regular hardware upgrades, fault repairs and performance optimization. Cloud computing migrates these tasks to cloud services, and enterprises need to worry about the maintenance and management of the hardware they serve again. In addition, the automated operation and maintenance tools provided by cloud computing platforms, such as automatic scaling, load balancing and monitoring services, make the operation and maintenance of the system more reliable.

4.1.3 Tiered Payment Model

The per-second payment model of cloud computing allows enterprises to pay only for the resources actually used. This model not only reduces resource waste, but also improves the efficiency of capital use. For example, enterprises can temporarily increase computing resources and pay the corresponding fees during business peaks, and reduce resource usage and save costs during business troughs. Cloud services initially also provide a variety of pricing strategies, such as hourly, monthly and annual payments. Enterprises can choose the best pricing plan based on their own budgets and needs.

4.1.4 Cost Forecasting and Optimization

The cloud computing platform provides detailed cost reports and analysis tools to help enterprises monitor and analyze resource usage in real time, thereby predicting and optimizing costs. For example, AWS's cost management tool (Cost Explorer) and Azure's cost management and controller (cost management and billing) provide a wealth of cost analysis and optimization suggestions. Enterprises can use these tools to identify cost hotspots, optimize resource usage strategies, and further reduce costs.

4.2 Resource Elasticity

The cloud computing platform can dynamically adjust computing resources according to changes in workloads to ensure the efficient execution of data analysis tasks. This elastic expansion capability enables enterprises to respond more flexibly to fluctuations in data processing needs.

4.2.1 Dynamic Resource Allocation

The cloud computing platform realizes dynamic allocation of computing resources through automated resource management tools. For example, AWS's Auto Scaling service can automatically adjust the number of computing resources according to preset policies to cope with changes in business load. When the workload increases, the Auto Scaling service will add more computing instances to ensure system performance and stability; when the workload decreases, the Auto Scaling service will release excess resources to save costs.

4.2.2 Efficient Use of Resources

The elastic expansion capability of cloud computing has greatly improved resource utilization. In traditional IT environments, in order to cope with sudden peak loads, enterprises often need to equip a large number of spare resources, resulting in a waste of resources. Cloud computing platforms can dynamically allocate and efficiently utilize resources through elastic expansion, thereby improving resource efficiency. For example, Google Cloud's Kubernetes Engine (GKE) realizes dynamic allocation and efficient utilization of resources through containerization technology, and users can

flexibly adjust resource configuration according to actual needs.

4.2.3 Coping with Sudden Loads

The elastic and rapid expansion capability of cloud computing platforms enables enterprises to cope with sudden load demands. For example, during e-commerce promotions, the number of visits and transactions often increases suddenly. Traditional IT infrastructure is difficult to cope with this in a short period of time. Cloud computing platforms can quickly increase computing and storage resources through automated expansion to ensure the stable operation of the system. In addition, cloud computing platforms also support global load balancing, distributing traffic to different geographical regions, improving the availability and reliability of the system.

4.2.4 Future Development Trends

With the continuous development of cloud computing technology, the application scenarios of resource elasticity will become more extensive. For example, the rise of edge computing and serverless computing (Serverless) technologies will further enhance the elasticity and flexibility of resources. Edge computing performs computing and storage near the data source, reducing the delay of data transmission and improving the efficiency of real-time data processing. Serverless computing further reduces resource waste and management costs by continuously executing code without pre-setting servers.

4.3 Efficient Collaboration

Cloud computing platforms support multi-user collaboration and improve the sharing of data and analysis tools among different teams and departments. Through the cloud platform, team members can access and analyze data in real time, improving work efficiency and decision-making quality.

4.3.1 Real-time Data Access

Cloud computing platforms provide a unified data storage and access interface, allowing team members to access data anytime, anywhere. For example, Amazon's S3 and Google Cloud Storage support storage access and data sharing for multiple users. Through permission management and access control, team members can securely access and use data according to their roles and permissions. This real-time data access capability allows team members to quickly obtain the latest data, improving work efficiency and response speed.

4.3.2 Collaborative Analysis Tools

The cloud computing platform provides a wealth of collaborative analysis tools to support team members to jointly conduct data analysis and decision-making. For example, Google's BigQuery and Amazon's Redshift support concurrent queries and analysis by multiple users. Team members can jointly conduct data analysis and discussion through shared data tables and query scripts. The cloud platform also provides data visualization tools, such as Google's Data Studio and Amazon's QuickSight, to support team members to jointly create and share data reports and dashboards, and carefully analyze and display results.

4.3.3 Project Management and Communication

The cloud computing platform also provides project management and communication tools to support collaboration and communication among team members. For example, Google's G Suite and Microsoft's Office 365 provide a wealth of collaborative documents and communication tools, such as Google Docs, Sheets, Slides and Microsoft Word, Excel, PowerPoint, etc. Team members can use these tools to jointly edit and share documents, communicate and discuss in real time, and improve the efficiency of team collaboration.

4.3.4 Security and Compliance

In the process of multi-user collaboration, data security and compliance are important issues. Cloud computing platforms ensure data security and privacy through strict permission management and access mechanism control. For example, AWS's IAM (Identity and Access Management) and Google Cloud's IAM provide sophisticated permission management and auditing functions to ensure that only authorized users can access and operate data. In addition, cloud platforms also support a variety of certifications, such as ISO 27001, SOC 2, GDPR, etc., to meet the compliance requirements of different industries and regions.

4.3.5 Future Development Trends

With the continuous growth of collaboration needs, collaboration tools on cloud computing platforms are also evolving. For example, AI-based intelligent collaboration tools will further improve the team's collaboration efficiency and decision-making quality. AI technology can help team members obtain and understand data more quickly through intelligent recommendations and automated analysis. In addition, the application of virtual reality (VR) and augmented reality (AR) technologies will enable team collaboration to provide a more immersive and interactive experience, improving the effectiveness of remote collaboration.

5. Challenges of Cloud Computing in Big Data Analysis

5.1 Data Security and Privacy

In a cloud computing environment, data storage and processing rely on third-party services, which makes data security and privacy protection the focus of enterprises and users. First, data needs to be encrypted during transmission to prevent interception and theft. Cloud services usually use TLS/SSL protocols to ensure the security of data transmission. Secondly, data also needs to be encrypted during storage. Many cloud services provide static data encryption services, such as Amazon's S3 encryption and Google Cloud's default encryption. In addition, access control and authentication mechanisms are also means to protect data security. Through strict permission management and multi-factor authentication, cloud services can effectively prevent unauthorized access to important data. However, even so, data leaks still occur from time to time, so continuous security monitoring and timely response mechanisms are also essential.

5.2 Performance Optimization

Although cloud computing provides powerful computing power, how to optimize the use of resources and improve the efficiency of data processing is still a problem that needs to be solved. Especially when dealing with scenarios such as real-time data and location transactions, performance optimization is extremely important. First of all, automation and elastic expansion of resource configuration are the key to improving performance. The cloud platform provides automatic expansion and load balancing functions, which can dynamically adjust resource configuration according to the workload to ensure that the system is always running in the best state. Secondly, data localization and caching strategies can also significantly improve performance. Store data in the location closest to the computing node to reduce the delay of data transmission. In addition, using memory computing technology, such as Spark's memory processing, can greatly improve the speed of data processing. Finally, through monitoring and analysis tools, continuous optimization of system performance and resource utilization is also an important means of performance optimization.

5.3 Regulatory Compliance

Different countries and regions have different legal and regulatory requirements for data storage and processing. Cloud computing services need to ensure that their services comply with local laws and regulations and reduce legal risks. For example, the EU's General Data Protection Regulation (GDPR) imposes strict requirements on the storage and processing of personal data, and cloud services must ensure that their services comply with these regulations. Similarly, the U.S. Health Insurance Portability and Accountability Act (HIPAA) also has clear requirements for the protection of medical data. These laws and regulations, cloud service construction must provide compliance certification and compliance reports to help enterprises ensure that their cloud computing use complies with local legal requirements. In addition, cloud service construction also needs to provide data sovereignty solutions, allowing customers to choose the location of data storage and face data claims in different regions. Current legal compliance audits and updates are also important means of ensuring cloud service compliance.

6. Future Development Direction

6.1 Edge Computing

With the popularity of Internet of Things (IoT) devices, edge computing has become an important supplement to cloud computing. Edge computing reduces the delay of data transmission and improves the efficiency of data processing by performing calculations and storage near the data source. For example, in intelligent real-time traffic systems, edge computing can process data near traffic lights or cameras, respond quickly, and optimize traffic flow. In the future, the combination of edge computing and cloud computing will further promote the development of big data analysis and achieve more efficient real-time data processing. By delegating some computing tasks to edge devices, cloud computing centers can centrally handle more complex data analysis and prediction tasks, thereby

improving the performance and reliability of the overall system.

6.2 Artificial Intelligence and Machine Learning

Cloud computing platforms provide an ideal infrastructure for artificial intelligence and machine learning. Through cloud computing, users can easily access and use various AI and ML tools for large-scale model training and reasoning. For example, Amazon SageMaker and Google AI platform provide comprehensive machine learning services that support the entire process from data model to model deployment. In the future, the deep integration of cloud computing and artificial intelligence will bring more automated data analysis solutions. The development of automated machine learning (AutoML) technology enables non-professional users to create machine learning models. Combined with the efficient and powerful computing power of AI, enterprises can quickly gain insights and make smarter business decisions.

6.3 Multi-cloud and Hybrid Cloud Architecture

In order to avoid dependence on a single cloud service, more and more enterprises are beginning to adopt multi-cloud and hybrid cloud architecture. Multi-cloud architecture allows enterprises to use resources built by multiple cloud services at the same time, optimize performance and cost, and reduce service interruptions. Hybrid cloud combines the advantages of public cloud and tree cloud, providing higher flexibility and security. For example, enterprises can store sensitive data in tree cloud while using the basic computing resources of public cloud for data analysis. In the future, multi-cloud and hybrid cloud architecture will become an important trend in big data analysis, providing more flexible resource management and more efficient data processing capabilities. Enterprises can dynamically adjust cloud resource configuration according to business needs to achieve the best performance and cost.

7. Conclusion

Cloud computing technology provides powerful computing power and flexible and efficient resource management methods for big data analysis. Through cloud computing, enterprises can store and process massive data and achieve in-depth data analysis and insights. However, cloud computing still faces severe challenges in data security, performance optimization and regulatory compliance. With the continuous advancement and innovation of technology, the application of cloud computing in big data analysis will be more extensive and in-depth, promoting the digital transformation and development of all walks of life.

References

Chen, Z. K. (2021). Analysis on the application of cloud computing technology in computer big data analysis. *China New Communications*, *23*(14), 103-104.

Gao, J. (2024). Research on the application of cloud computing technology in computer big data analysis. *Shanxi Science and Technology News*, *2024-04-19* (B08).

Gao, S. L. (2021). A brief discussion on the application of cloud computing technology in computer big data analysis. *Network Security Technology and Application*, *2021*(07), 82-83.

Huang, F. (2020). Application countermeasures of cloud computing technology in computer big data analysis. *Information and Computer (Theoretical Edition)*, *32*(16), 20-22.

Huang, S. J., & Gu, J. X. (2022). Application of cloud computing technology in computer big data analysis. *Electronic Technology and Software Engineering*, *2022*(08), 235-238.

Jia, J. Y. (2022). Application of computer big data and cloud computing technology. *Electronic Technology*, *51*(12), 172-173.

Jia, Y. G. (2020). Research on the application of cloud computing technology in computer big data analysis. *Computer Products and Circulation*, *2020*(06), 163.

Li, Z. W. (2023). Application of cloud computing technology in computer big data analysis. *Information and Computer (Theoretical Edition)*, *35*(15), 1-3.

Liang, H. (2020). Application of cloud computing technology in computer big data analysis - Review of "Cloud Computing and Big Data". *Science and Technology Management Research*, *40*(16), 267.

Liu, N. (2023). Exploration of the application of cloud computing technology in computer big data analysis. *Digital Communication World*, *2023*(04), 128-130.

Lv, G. Q. (2022). Analysis on the application of cloud computing technology in computer big data analysis. *Smart China*, *2022*(07), 92-93.

Mi, J. (2022). Research on the application of cloud computing technology in computer big data analysis. *Henan Science and Technology*, *41*(06), 16-19. http://doi.org/10.19968/j.cnki.hnkj.1003-5168.2022.06.003

Wang, K. X. (2023). Exploration of the application of cloud computing technology in computer big data analysis. Guangdong Teachers Continuing Education Association. *Proceedings of the Guangdong Teachers Continuing Education Association Seminar on "Education and Innovation Integration"* (III). Qilu University of Technology (Shandong Academy of Sciences), 5. http://doi.org/10.26914/c.cnkihy.2023.062243

Wei, J. L. (2022). Thinking on the application of cloud computing technology in computer big data analysis. *Modern Industrial Economy and Informatization*, *12*(09), 76-78. http://doi.org/10.16525/j.cnki.14-1362/n.2022.09.031

Xia, D. H. (2022). Application of cloud computing technology in computer big data analysis. *Wireless Internet Technology*, *19*(03), 111-112.

Zhang, R. (2021). Application countermeasures of cloud computing technology in computer big data analysis. *Information and Computer (Theoretical Edition)*, *33*(09), 40-42.

Zhang, X. L. (2022). Application of cloud computing technology in computer big data analysis. *Computer Knowledge and Technology*, *18*(19), 25-27. http://doi.org/10.14004/j.cnki.ckt.2022.1290

Zhao, L. (2023). Analysis on the application of cloud computing technology in computer big data analysis. *Modern Industrial Economy and Informatization*, *13*(02), 175-177. http://doi.org/10.16525/j.cnki.14-1362/n.2023.02.064