HiLL PUBLISHING

Engineering Advances, 2023, 3(3), 255-258
https://www.hillpublisher.com/journals/ea/
ISSN Online: 2768-7961

Research on Distributed Storage Technology for Big Data

Zhenyao Duan

Chongqing University of Science and Technology, Chongqing, China.

How to cite this paper: Zhenyao Duan. (2023) Research on Distributed Storage Technology for Big Data. *Engineering Advances*, 3(3), 255-258.
DOI: 10.26855/ea.2023.06.020

Received: May 31, 2023
Accepted: June 29, 2023
Published: July 27, 2023

***Corresponding author:** Zhenyao Duan, Chongqing University of Science and Technology, Chongqing, China.

Abstract

With the widespread popularity of mobile Internet and the rapid development of Internet technology, a large number of new applications have emerged in the network, and the amount of data generated by these applications has shown an amazing explosive growth, and people are deeply aware of the huge value contained in these huge data. The term "big data" is widely used in various industries." With the continuous development of emerging fields such as artificial intelligence, machine learning, and user profiling, the term "big data" has been widely used to describe the amount of data and has become an indispensable support force in these fields. Cloud storage, as an emerging data service approach, has been widely used among various business systems in recent years. With the booming development of cloud storage technology, its performance is also improving by leaps and bounds, along with the rapid progress of cloud computing technology. At present, cloud storage technologies are mainly divided into two types of distributed computing models based on virtualization platforms and centralized architectures based on file systems. The purpose of this paper is to study the application of distributed database technology in the big data environment in order to achieve its continuous optimization in the real environment, so as to better serve the daily life of human beings.

Keywords

Big data, distributed cloud storage, artificial intelligence

Introduction

With the continuous progress of society, information technology has penetrated into people's daily life, and with the increasing requirements of operators for data management and practical applications, the amount and type of data available in distributed database technology has been greatly expanded, so that traditional data processing methods can no longer meet the needs of people today. The term "Big Data" is widely used in various industries. In this context, the rapid development of distributed database technology is necessary to meet the processing needs of modern big data management, analysis and storage, so as to improve the efficiency of data applications and make an important contribution to the construction of society. At present, government agencies and enterprise units in most regions of China have established their own data centers, and distributed database system is one of the more advanced database technologies. The traditional cloud storage technology has been based on clusters, and this approach has been widely used. Using distributed cloud storage technology based on distributed files and grids, we can connect multiple storage devices together through software and store users' data independently in each device, thus greatly improving the efficiency and convenience of users' access to data.

1. Overview

1.1 Overview of distributed storage technology

In fact, distributed storage technology is a concept related to centralized storage technology and is unique in that it

stores data in a virtual network space rather than being fixed to a specific node. In order to meet the trend of the times, traditional data processing systems must be improved and innovated, and distributed database systems are one of the effective ways to do so. Specifically, distributed storage technology takes full advantage of the network and virtualizes the scattered storage space into a whole and uses it as the core of data storage. In this process, what we usually call "distributed" means that this storage technology can flexibly allocate storage resources among nodes according to users' needs. However, the data is actually stored independently in different storage spaces, not in any one node in the traditional sense. Therefore, distributed storage technology is to improve data utilization and reduce network traffic by building a unified cluster of storage devices, thus achieving the purpose of improving system performance. With the continuous development of distributed storage technology, distributed storage management system has come into being, which can efficiently integrate the fragmented network storage space and realize the decentralized storage of load through multiple servers, thus providing reliable guarantee for the security, reliability and practicality of the system. Therefore, distributed storage technology is considered to be one of the future trends of network storage technology. Decentralized storage and centralized management are the most significant features of distributed storage technology and the biggest advantage of its wide application [1].

1.2 Characteristics of Big Data and Storage Requirements

A large amount of uncertain heterogeneous data, the causes of which are diverse, also has new characteristics in its application: along with the continuous expansion of various application scales and fields, the data volume will show the trend and characteristics of explosive growth as well as massive data storage: in the typical applications of uncertain data, the number of data sources is large, the types of data are also diverse, and the data resources have heterogeneous characteristics. The data has uncertain features such as data block size, data type, and data access method; with the increasing abundance of applications such as cloud computing and Internet of Things, the generation, use, and access of data have become very complex, and also make the data present the characteristics of time, space, high access frequency, and high access frequency. In uncertain data applications, due to its huge data scale, it poses a huge challenge to the existing data storage architecture. First, in the face of massive data, decentralized organization and management strategies must be adopted, which requires (meta)data, (meta)data organization methods to adapt to uncertain data applications; second, in the face of growing massive data, which requires long-term accumulation of (meta)data to meet both the scale and performance requirements, as well as storage requirements, so the organization methods of (meta)data and indexing The method of organizing and indexing (meta)data is an urgent problem. In the face of large-scale uncertain data, the existing metadata storage methods and object-oriented query methods suffer from inefficiency during processing, and it is urgent to study new metadata organization methods and query methods to meet the user's demand for efficient query of multiple concurrent data. In this distributed environment, data sources are scattered on various nodes, which leads to inefficient network transmission. However, different data sources are highly autonomous and can modify their own structures and perform data updates autonomously, making it difficult to guarantee the consistency of the data fusion system. The uncertainty of data makes the fusion of large-scale uncertain heterogeneous data even more difficult, and the use of distributed parallel processing techniques enables global optimization of computation and storage. The characteristics of massive, uncertain and heterogeneous data bring new challenges to traditional data mining methods. For the characteristics of data heterogeneity, massive, distribution, and real-time, the layout of data mining engines and the collaborative scheduling of multiple engines have raised higher requirements. Both structured and unstructured data are related to the storage and management of data (indexing, concurrency, consistency, query, etc.), which is due to the needs of users in terms of big data applications (queries on massive unstructured data still have to be accurate and fast), thus putting new requirements on the logical structure and physical storage methods of data.

2. Distributed database technology in big data

2.1 Storage Process

In the unstructured data saving process, the user must first select the desired file from the client and package it as an "upload request", which is then sent to the data service layer, which then decides whether to send parameters to the server agent, depending on the server situation. The verdict is obtained by dividing and distributing said files according to said load balancing and sent to the data access layer. The data access layer reworks the document to form a document slice based on the decision information combined with the document it needs to upload. The documents are encapsulated and transferred to the data storage layer finally returning the result information to complete the storage process [2].

2.2 Low Cost Distributed Storage Technology

With the advent of the Big Data era, the rapid expansion of data scale has led to a dramatic increase in storage overhead. The proposed low-cost distributed storage technology has led to a significant reduction in storage costs, thus achieving the goal of saving storage costs. Among many low-cost distributed storage technologies, the most

cost-effective one is the object-oriented distributed storage technology, which separates the data transmission channel and data control channel mainly through the Web protocol to improve data transmission efficiency. In the object-oriented distributed storage, users can get two IDs, one is the storage object ID, the other is the storage space ID, these two IDs are corresponding to each other, so as to achieve the purpose of directed storage. At the same time, the memory between the two systems are at the same level and do not interfere with each other, so there is no cascading failure problem and effectively avoid data loss. Swift is a typical low-cost distributed storage system, which includes two parts: data access and data management. Its ability to query information through multiple nodes saves query time.

2.3 Optimization techniques within the storage hierarchy

Based on the dual consideration of cost and capacity, memories with different performance and price are generally used to reduce the memory overhead, thus constituting a multi-layer memory. Through the optimal configuration within the storage layer, the hierarchical, hierarchical and efficient hierarchical management of massive data can be realized, and the construction overhead of the corresponding system can be effectively reduced while ensuring sufficient data capacity. The memory hierarchy is optimized using a combination of "hot" and "cold" structures, thus achieving efficient optimization within the memory layer. In terms of cost savings, an information lifecycle management approach that transfers data to low-cost, high-rate storage devices with low latency can significantly reduce system manufacturing costs and energy consumption without significantly degrading overall system performance.

2.4 Application in Big Data Load Support

In the big data environment, data processing and analysis face the challenges of large data loads that are difficult to support, and these challenges seriously limit the efficient operation of big data. Traditional centralized databases are often system-wide driven when processing data, and big data is loaded when there is a large amount of redundancy in the data center. In the management of big data, the processing technology of distributed database can realize the real-time hierarchical management of data. The advantage of information classification processing is that the system can automatically process the information data and filter the duplicated data, and then later analyze and process the data accurately according to the information processing procedure, which can greatly improve the efficiency of database data processing and also improve the accuracy of data analysis and processing. In this way, even in the big data environment, distributed database technology can still provide high speed processing status for big data. The introduction of distributed database technology into the big data load support system can meet the analysis and management needs of different service types of data and improve the processing efficiency and accuracy of multiple types of data [3].

2.5 Improve the big data space management system

First is the data resource layer. Its role is to store and manage various data, which is equivalent to a relative data source. This technology can realize the fast reading and effective calculation of spatial big data, and can meet a variety of needs such as effective analysis and application of spatial big data for users. Second is the basic construction level. This layer is the basis to support the overall operation of the system, storage, computing, storage of all kinds of resources, but also contains security equipment, network and other resources, the resources are very rich. Meanwhile, at the application level, this layer is based on the business interface and functional building blocks provided by the component service layer, and realizes the construction of various forms of applications such as Web side and desktop side. The web application and desktop application contain many data types, where the web site contains the query of information and the display of results, etc.; while the desktop application contains the update of data and the security management of the system. Based on this, an object-oriented support method is proposed and object-oriented business processes are implemented on object-oriented objects.

2.6 System Resilient Scaling Technology

In the context of today's big data era, the scale of data is getting larger and larger, and the complexity and diversity of data is also getting stronger, which brings greater challenges to the storage of distributed data. System scalability is the core technology of distributed storage, which can effectively improve the scalability of storage system and can effectively solve the problems of metadata dispersion and data transparent migration in distributed storage. The system needs to estimate the number of nodes needed based on data volume and load, as well as dynamic data migration between nodes. The system should be able to perform dynamic data exchange between two nodes to achieve simultaneous load balancing. When errors occur in the nodes, the data can be recovered by methods such as copying, without affecting the high-level applications [4].

2.7 Big Data Business Management

In today's information-based society, all industries have begun to use the Internet for industrial change, and the In-

ternet is a major means of information management for enterprises. In this context, business between different industries is also able to be transmitted through network information, thus achieving the purpose of business docking. In big data management applications, enterprises are also confronted with multiple data types in huge quantities, which bring more practical difficulties to users' information processing. However, with the emergence of distributed database technology, it is able to meet the demand for classification of information and, in this technology, it is also able to achieve classification processing of customer registration, thus helping enterprises to screen out star and premium customers, which are usually targeted to the specific situation of the enterprise.

2.8 Big Data Loading

Distributed database technology applied to big data management can improve the efficiency of acquiring information and can automatically backup information to meet the automatic loading function of big data in the actual data processing. This technology has been widely used in various types of big data analysis. Big data usually has thousands of nodes, and the distributed database is also built on top of these nodes, which can significantly improve the overall data loading efficiency when combined with big data loading processing. For example, the information acquired every day is 3 billion, the actual efficiency of its information loading is 15000 items/s, and 30 billion data loading can be completed in a month, which is undoubtedly one of the most significant advantages of distributed database technology for the speed of massive data growth and the proliferation of data users for big data information loading [5].

Conclusion

With the development of Internet, mobile Internet and Internet of Things, "big data" has gradually become the trend of development, and "big data" is gradually formed, the reasons for its formation are intricate and complex, and its application also presents some new features. With the continuous expansion of various applications, the number of data will show an explosive growth, and accompanied by a large number of data storage characteristics. The new service environment and new application scenarios have put forward higher requirements for large-scale cloud storage. In the big data environment, the application of distributed storage technology to data storage can provide a new way of thinking for storing data efficiently and at low cost. Based on this, an object-oriented data storage method is proposed. In the big data environment, the use of distributed storage technology can effectively reduce the cost of data storage management system and expand its scope of application, break through the limitations of traditional data storage technology, achieve the maximum optimization of storage space, and reduce the waste of space resources. Based on a brief introduction of distributed storage technologies, two representative types of distributed storage technologies in the big data environment are analyzed, hoping to provide reference for the development of data storage technologies in the domestic big data environment.

References

[1] Wang Zhihui. Application of distributed database technology in big data [J]. Information Systems Engineering, 2019(12):21-22.

[2] Jiang N. Research on the application of distributed database technology in big data [J]. Electronic World, 2020(11):109.

[3] Deng B, Chen Huiping. Application of distributed database technology in big data [J]. Information Recording Materials, 2020, 21(06):150-151.

[4] Park Chengzhe. Distributed storage method for large-scale unstructured data [J]. Journal of Taiyuan Normal University (Natural Science Edition), 2021, 20 (03):57-61

[5] Zhao Yuanyuan, Wang Ke, Zhou Yao. Research on Data Layout Strategies Based on Distributed Storage Systems [J]. Modern computers, 2019 (21):18-22.