World Journal of Innovation and Modern Technology, Vol. 8, Issue 4, (Apr) 2025
ISSN 2682-5910
DOI: 10.53469/wjimt.2025.08(04).09

On Data Storage Technology Based on Big Data

Shanshan Rao

School of Computer and Software, Jincheng College, Sichuan University, Chengdu, 611731

Abstract: ***Big data storage process is an important link before data processing, with its efficient and accurate function of data storage. With the step of information science into, data from all walks of life all the time, exploding state, then more and more complex data types, the traditional relational database's ability to deal with the data already can not adapt to the development of the era of big data, a new storage technology is gradually replacing traditional database, this paper reviewed on the basis of the characteristics of big data the discussion of the mass data storage technology, the development trend of data storage are analyzed.***

Keywords: Data Storage, Data Analysis, Cloud Computing.

1. INTRODUCTION

With the increase and diversity of information data, big data has begun to enter people's sight in recent years. With the emergence of cloud computing, big data has developed rapidly and is now well-known to people. On the other hand, the development of big data has accelerated the technological innovation of cloud computing. Why do we say this? As we all know, the most obvious feature of big data is the huge amount of data. To efficiently store such a large amount of data, it is impossible to achieve it with traditional processing capabilities. Cloud based databases can better solve this problem; Both domestically and internationally, the issue of data storage has always been a concern, and people are also improving traditional databases. The technologies used for data storage and processing mainly include distributed, virtualization, and other technologies to overcome the problem of data storage People are constantly exploring. Lyu et al. [1] proposed optimized convolutional neural networks (CNNs) to enhance the efficiency of 3D point cloud object recognition, addressing computational bottlenecks in real-time applications. For urban energy management, Zheng et al. [2] developed a hybrid forecasting model (GWO-SARIMA-LSTM) leveraging the TRIZ method to optimize building energy consumption. In healthcare, Pang et al. [3] employed electronic health records (EHRs) to construct a data-driven framework for diabetes risk prognosis, highlighting the potential of AI in chronic disease management. Meanwhile, Liu et al. [4] introduced a privacy-preserving hybrid ensemble model for network anomaly detection, balancing security and data protection in cybersecurity. The cybersecurity domain is further explored by Xu et al. [5], who analyzed adversarial machine learning attacks and defense mechanisms, emphasizing the vulnerabilities of AI systems. Autonomous systems have also benefited from AI innovations, as seen in Wang et al. [6]'s end-to-end autonomous driving framework and Chen [11]'s work on scalable cloud infrastructure for autonomous vehicle data processing. In clinical settings, Shen et al. [7] integrated LSTM-based AI systems to optimize anesthetic dosing in cancer surgeries, demonstrating improved precision. AI's impact extends to behavioral health, where Lin et al. [8] investigated intelligent monitoring systems to enhance executive function in children with ADHD, while Peng et al. [9] studied the interplay between aerobic exercise intensity, executive function, and sleep. Lastly, in logistics, Luo et al. [10] designed a path-planning algorithm combining Transformer and GCN networks to improve robotic efficiency.

2. CHARACTERISTICS OF BIG DATA

When big data arrives, information is the result of the information age. What is big data? Simply put, big data refers to data that cannot be processed by ordinary software and technology. So, what are the characteristics of big data? Below we will briefly introduce the basic characteristics of big data, namely the 4V characteristic:

2.1 Large Volume of Data

The most obvious feature of big data is its massive amount of data, from individuals to enterprises and even to countries. According to statistics, the amount of data generated today has developed from TP level to PB level. The storage problem caused by the massive amount of data cannot be underestimated. Therefore, we need to filter the huge amount of data according to different standards, but more data is not better. We need to choose valuable data,

structured data, and clean data that does not meet people's expectations. This is also the most important step in processing and analyzing big data.

2.2 Complex Data Types (Variety)

Data types are not just numerical types that people consider, including traditional structured data. In addition, there are also unstructured data, including numbers, graphics, and so on. To be precise, any symbol that can describe material characteristics can be called data. Data analysis is about finding internal connections from these data. On the other hand, how to better store data is also a difficult problem to solve. So, with such complex data types, data storage has become one of the reasons restricting the development of the big data industry.

2.3 Low Value Density (Velocity)

The value of data presentation can be priceless or meaningless, such as the famous Haiyan system, which is currently the strictest system for detecting traffic violations. It monitors a large amount of traffic information every day, but the results of rapid data processing and analysis may not bring effective value to the police; Another example that is close to our daily lives is that the ticket sales records of airlines may only appear as numbers on the surface, but through certain data analysis, internal correlations can be found, providing airlines with flight plans and ticket policies. So the value of data needs to be analyzed through certain methods and presented to people in a visual way.

2.4 Fast processing speed

The 1-second rule states that due to the rapid development of information, humans will generate approximately 3 trillion bytes of data per day. With the development of information networks, this amount of data will continue to increase. Whether in daily life or on the internet, the amount of data is constantly expanding rapidly, and how to cope with such an increase has become a major challenge for people. However, it cannot be denied that this era has higher requirements for hardware and software performance in order to adapt to the rapid development of the information age.

3. DATA STORAGE

With the explosive growth of digital book data, multimedia, e-commerce enterprises, and other data, the capacity unit of data has gradually evolved from TB to PB level, and the corresponding data storage capacity has become a major challenge. In addition to preserving the past and present data volume, data storage also requires updating of data and preventing rapid data expansion from causing server overload and collapse. Therefore, efficient storage of massive data is also crucial.

Data storage refers to the data that a data stream needs to query during certain activities or transactions. Data is not temporary, and both enterprises and individuals often hope to save it. For individuals, historical data can not only provide effective value at that time but also help solve problems in the future. There are many ways to store data. For individuals, the amount of data is small, and some basic storage devices can meet user needs and are easy to implement and maintain in the later stage, such as USB drives, computers, etc. However, for enterprises, the amount of data is huge and requires confidentiality, which cannot be met by storage devices such as USB drives.

Generally speaking, data is stored inside a computer in certain specific formats or rules, such as disks; In addition, there are some external storage devices, and the storage of big data is obviously not that simple. The scale and complexity of data in the big data environment are increasing rapidly, and traditional database deployment can meet human needs. We need to develop new technologies to promote research on storage technology in the field of big data.

4. BIG DATA STORAGE TECHNOLOGY

The key to the era of big data is not only to help people analyze valuable information, but more importantly, how to store this valuable information knowledge and provide effective information for the future or present. The emergence of big data is accompanied by the development of the information industry, promoting the innovation of storage technology. In the face of the current situation of the information industry with a large amount of data and complex structure, what kind of storage methods should be adopted is also something that the information

industry has been striving to explore. The storage models currently used include: NoSQL, Cloud computing storage MPP, Distributed, we will focus on discussing NoSQL database models, and there are significant differences between distributed and traditional models in terms of stability, efficiency, and independence.

4.1 NoSQL

Traditional relational databases have slow data storage and processing speeds, low scalability and elasticity when facing big data, which also means they cannot become the primary choice for big data storage. Therefore, NoSQL is a data management technology that has emerged to meet the growing needs of the information industry. NoSQL refers to a series of non relational databases, which can be said to be born for big data and breaks the limitations of traditional database models.

NoSQL is used to handle unstructured data types and accommodate complex data models (as shown in Figure 1). Compared to traditional databases that store fixed data structures, NoSQL has stronger scalability and faster processing speed. For example, in traditional databases, each tuple has the same structure, and even if an instance does not have a certain field type, it will be assigned defined fields; The NoSQL model stores data using a key value pair standard, without first fixing the formatting structure of tuples. Instead, it defines data based on different requirements for each tuple, while achieving no connection between data. This greatly reduces the cost of space resources and time overhead.

NoSQL type databases can be built on low-cost hardware devices, and they also support distributed storage. Distributed storage is a network environment where data is distributed and stored on different nodes (servers), which is invisible to users. This makes NoSQL highly scalable and has low maintenance costs. Below, we will introduce two members of NoSQL database technology - MongoDB and HBase.

MongoDB is a document oriented storage model and an important member of NoSQL. It is essentially an intermediate product between relational and non relational, supporting the storage of multiple complex data types. MongoDB documents are generally stored in the format of BSO.

HBase is a storage architecture model for columns, which works by storing different columns on top of each other to form column clusters. Column clusters refer to collections of columns, which can update and read data columns in real time more quickly.

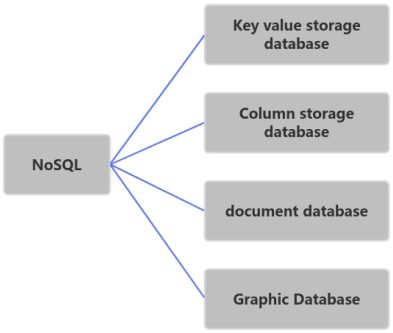

Figure 1

4.2 Distributed System Storage

4.2.1 Distributed Principle

The data types are diverse and the structures are mixed, making it very complex to process. Traditional databases tend to have slower response times when data grows to a certain level, such as processing tens of fields from a data table to hundreds of data tables. This disadvantage is actually related to their data processing mode, which is centralized storage and non distributed processing.

Data is stored on centralized servers, and if one of them is overloaded and crashes, data loss can easily occur. Distributed system storage refers to dividing data into different parts and allowing multiple processors to process

their data in parallel. This is similar to the model of cloud storage, where user data is stored on various servers with corresponding monitoring mechanisms and alarms. A node crash will not cause other nodes to crash at the same time. Distributed technology allows multiple legitimate users to access stored data and directories at a single point in time. On the other hand, distributed file systems can allow two or more nodes to simultaneously execute related database transactions.

4.2.2 Comparison between Distributed and Traditional Modes

(1) High access efficiency

Due to the distributed processing mode using data distributed across multiple servers on different processors, each server's transaction processing does not have to wait for the completion of the previous transaction. The data is processed in parallel between processors. For example, Hadoop is a distributed infrastructure that can process a large amount of data in a distributed manner. If the processor running on it produces 1000MB of data, the total storage time is determined by the positioning time plus the transmission time. If the transmission time is 1s, the scheduling is 2s, and the speed is 100MB/s, it takes 7.4 seconds to complete the transmission by 23 machines simultaneously. In contrast, the traditional processing mode requires waiting, which is equivalent to serial transmission of data. Compared to this, distributed data has less storage time and higher efficiency.

(2) Strong independence and scalability

Multiple servers work simultaneously, and each part is independent of each other. When a problem occurs on one server, it does not affect the process tasks of the other servers, greatly improving the stability of data and the system; Distributed scalability is strong, and the various data processed in a distributed manner are placed in different places, independent of each other. When new nodes are added, it will not affect the loss of data. If a node has excessive storage capacity, it can achieve load balancing and strong horizontal scalability. However, the traditional mode is the opposite, and its processing mode determines its poor elasticity.

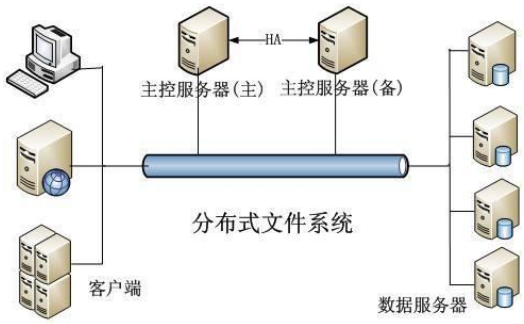

Figure 2

4.3 MPP architecture

MPP architecture is a real-time processing architecture for massive data, mainly used in industry big data. As an exclusive architecture, nodes run their own operating systems and data. This MPP architecture is widely used by some parallel relational databases. MPP products can support massive data levels of TB, which is unimaginable for traditional data databases. Therefore, for data warehousing and structured data analysis and storage processing, MPP database is the best choice. Nowadays, query systems such as EMC and Google Dremel are using this technology.

4.4 Cloud based database storage

Cloud database is based on cloud computing, where databases are built in a cloud computing environment. Enterprise users can rent database resources and store data on cloud databases. Compared with traditional databases, cloud databases use virtualized storage. The key to virtualized storage is to map physical facilities to logical resource pools. By providing virtualized storage space, such as virtual disks and virtual memory, for customers and enterprises, users can combine resources according to their needs. Nowadays, with the development of cloud computing, the three forms of IAAS (Infrastructure as a Service) as the foundation of cloud computing

can provide storage space and store corresponding data for users. Of course, this is only to meet the needs of ordinary users. User data is stored on different servers, if a larger enterprise has a massive amount of data, it can build its own private cloud platform and store enterprise data in the private cloud. This not only saves storage space, optimizes storage virtualization efficiency, but also reduces storage costs. Up to now, cloud technology has rapidly developed in recent years, especially in the field of cloud virtual storage, such as network cloud disks, virtual machines, and other applications.

5. CONCLUSION

The 21st century is the era of information and data. From life to work, people are constantly generating data, and big data brings not only social reform, but also changes in thinking and technology. This article provides a simple analysis of the characteristics of big data and data storage in the big data environment. We can say that in order to adapt to the development of the big data era and meet different information needs, corresponding storage technologies have emerged. Therefore, big data not only brings us information knowledge, but also guides the direction of storage technology reform. The big data era has begun, and data storage technology needs to be constantly updated to meet the requirements of the data era. Humans need to constantly explore in order to go further in data storage technology.

REFERENCES

[1] Lyu, T., Gu, D., Chen, P., Jiang, Y., Zhang, Z., & Pang, H. & Dong, Y.(2024). Optimized CNNs for Rapid 3D Point Cloud Object Recognition. arXiv preprint arXiv:2412.02855.
[2] Zheng, S., Liu, S., Zhang, Z., Gu, D., Xia, C., Pang, H., & Ampaw, E. M. (2024). Triz method for urban building energy optimization: Gwo-sarima-lstm forecasting model. arXiv preprint arXiv:2410.15283.
[3] Pang, H., Zhou, L., Dong, Y., Chen, P., Gu, D., Lyu, T., & Zhang, H. (2024). Electronic Health Records-Based Data-Driven Diabetes Knowledge Unveiling and Risk Prognosis. arXiv preprint arXiv:2412.03961.
[4] Liu, S., Zhao, Z., He, W., Wang, J., Peng, J., & Ma, H. (2025). Privacy-Preserving Hybrid Ensemble Model for Network Anomaly Detection: Balancing Security and Data Protection. arXiv preprint arXiv:2502.09001.
[5] Xu, J., Wang, Y., Chen, H., & Shen, Z. (2025). Adversarial Machine Learning in Cybersecurity: Attacks and Defenses. International Journal of Management Science Research, 8(2), 26-33.
[6] Wang, Y., Shen, Z., Hu, K., Yang, J., & Li, C. (2025). AI End-to-End Autonomous Driving.
[7] Shen, Z., Wang, Y., Hu, K., Wang, Z., & Lin, S. (2025). Exploration of Clinical Application of AI System Incorporating LSTM Algorithm for Management of Anesthetic Dose in Cancer Surgery. Journal of Theory and Practice in Clinical Sciences, 2, 17-28.
[8] Lin, L., Li, N., & Zhao, S. (2025). The effect of intelligent monitoring of physical exercise on executive function in children with ADHD. Alexandria Engineering Journal, 122, 355-363.
[9] Peng, Y., Zhang, G., & Pang, H. (2025). Impact of Short-Duration Aerobic Exercise Intensity on Executive Function and Sleep. arXiv preprint arXiv:2503.09077.
[10] Luo, H., Wei, J., Zhao, S., Liang, A., Xu, Z., & Jiang, R. (2024). Intelligent logistics management robot path planning algorithm integrating transformer and gcn network. IECE Transactions on Internet of Things, 2(4), 95-112.
[11] Chen, J. (2025). Leveraging Scalable Cloud Infrastructure for Autonomous Driving Data Lakes and Real-Time Decision Making.

Author Profile

Shanshan Rao (1997-), female, Han, Chengdu, Sichuan Province, China, undergraduate, Jincheng College, Sichuan University, research direction: e-commerce.