Research on Key Technologies of Distributed File System for Large Data Analysis Based on Supply Chain Management

Zhu Jia

Shandong Institute of Commerce & Technology, Jinan, 250013, China

Keywords: Supply Chain; Management; Big Data; Distributed Files; Technology

Abstract: With the advent of the era of big data, data analysis and processing capabilities have become increasingly important technical means for data centers and Internet companies. With the expansion of information scale and the diversification of data structures, massive data storage has become a hot research topic in large data analysis. Supply chain will generate a lot of information in the process of operation. Through large data analysis, this information will be transformed into business intelligence. When collaborative management of supply chain system information is carried out, the system is improved by using the functions of data and information fast reflection, high-speed transmission and feedback regulation of the system itself under large data. The emergence of the distributed file system perfectly matches the requirements of massive data applications, making it the only choice for solving the huge data dilemma in the contemporary era. As an entry point for users to use and manage distributed file systems, the distributed file system provides users with an abstraction of the entire file system, shielding users from the complexities between the client, the metadata server, and the data server when using the distributed file system. Interactive communication and processing logic.

1. Introduction

In the process of China's rapid economic reform and innovation, more and more enterprises have entered the stage of rapid development. In the era of economic globalization, the economy is developing towards integration and integration. The development of enterprises has unprecedented opportunities in the fierce competition [1]. The practical application of big data analysis in supply chain management, the main types of data directly determine the form of data presentation and specific collection channels [2]. The development of the Internet ushered in the era of big data, and the processing and analysis technology of data information has become a hot topic of discussion among people from all walks of life. Increasing data and rapidly growing number of files have become the main characteristics of data storage in the current new era [3]. As far as the current problems are concerned, relevant scholars should optimize the technical means according to the characteristics of Internet data and distributed files to reduce the cost of data migration, realize the expansion and optimization of system space, and improve the efficiency and quality of data processing. In the actual operation of distributed technology, the electronic computer system is processed in various forms, such as mapping the dimensions and metrics of information data. This is done through a multidimensional model. All operations that interact with the metadata server and data server are done in the DFSClient class [4]. This instantiates the client protocol when the client instantiates, creates a remote procedure call connection to the metadata server, and then implements specific functional big data analysis directly through remote procedure calls based on the user's intent to better understand what we are not aware of. The world competes with business and gains a competitive advantage based on technology. Therefore, big data analysis will play an increasingly important role for business leaders and industry elites in all walks of life [5].

Traditional distributed file systems are difficult to meet the requirements of the new situation in terms of scalability, reliability and data access performance. A distributed file system for large-scale data analysis and cluster applications is designed and implemented [6]. In most file systems, all control and logic are implemented in the code itself of the kernel file system. The logical implementation of distributed files is distributed in all nodes, which simplifies the client and brings

DOI: 10.25236/isaicn.2019.022

huge (dynamic) scalability [7]. Information quality forms a transition from the lack of information to the existence of complete information. Since uncertainty always exists, gray analysis can lead to a series of clear statements about the solution. In one extreme case, this solution has no solution, and in the other extreme, a system with perfect information has a unique solution [8]. Data is extremely important for supply chain management. Through the analysis and processing of such data, it can effectively guide the development of new projects in the specific supply chain management, the participation of relevant stakeholders, and the supervision of supply chain risks in the actual operation and development process. And the market in-depth investigation and other aspects of scientific planning and design [9]. Whether it is economic activities or social activities, it will generate a variety of trading activities on the Internet at an extremely fast rate, and will also generate real-time data. The ability to capture and understand these data and information is the foundation of big data analysis. Therefore, this paper studies the key technologies of distributed data system for big data analysis of supply chain management [10].

2. Materials and Methods

2.1 Application of Big Data Analysis in Supply Chain Management

High integration and information sharing of supply chain system can improve the quick response ability of supply chain to market and improve customer satisfaction. A reliable system form is established for storage and data verification at block level. On this basis, different block information is transmitted to the corresponding nodes to minimize redundant overhead and improve the usability of the file system in all aspects. Finally, the specific content is effectively combined with metadata information. Because massive data is accumulating continuously, in the process of accumulation, large storage space is needed, and its performance needs to be expanded, which requires the establishment of a matching storage organization model and index mechanism. The dimension information and factual information are stored separately, and the mapping between them is realized by using foreign keys. However, since the actual operation process involves a connection operation, this also makes the operation of the actual operation process less efficient. The data is scientifically configured, and various types of data are dig deeper. The reliability of the sample data (reliability) and the validity of the sample data ensure a good fit of the hypothesis. Factor analysis and reliability analysis are shown in Table 1. Show. To obtain valuable data to promote the operation of all stages of enterprise supply chain management, fundamentally improve the corresponding speed of enterprise supply chain management, and effectively upgrade the modern enterprise supply chain management process to ensure the maximum economic benefits.

Table 1 Factor analysis and reliability analysis

	Fitness value	Load volume
Information synergy	0.731	0.819
Big Data Applications	0.826	0.720
Supply chain optimization	0.853	0.821

2.2 Relevant Technology of Distributed File System Client

Object storage and separate management of metadata and data are adopted. Metadata server cluster manages the namespace of file system by caching and distributed metadata server. The actual file data transmission is carried out between client and object server. Converting the storage space of data storage into the storage space of common application host greatly improves the performance-price ratio. Through large data analysis, the analysis of logs can be evenly distributed on different PC hosts. More logs can be collected and processed calmly, and the system can be improved for monitoring and early warning. Mapping placement groups to storage devices relies on a pseudo-random mapping rather than relying on any metadata mapping, minimizing storage overhead and simplifying data distribution and lookup. Accelerate the transformation of the traditional storage and transportation industry into the modern supply chain industry, implement specialized, large-scale operation, share related supply chain facilities, thereby reducing operating costs, giving full play to the overall advantages of supply chain enterprises, and improving the efficiency of the supply chain, which will play an important role. Modern enterprise supply chain financial management brings a major opportunity. The new data covers map data, video data and audio data, etc. It can be effectively applied to visualization work with strong real-time and precision.

Large data distributed storage system can store large-scale structured data, through large-scale storage and management technology, to achieve efficient data processing. It is possible to share information among supply chain nodes by collecting, screening and sorting out data. Through the establishment of a new data management platform with new data processing and sorting algorithms, the data of various types of structure can be screened, cleaned and transmitted. In the whole distributed file system, the name space of the file system is split by the name of the file path, which becomes the main principle of dividing the name space. This principle can reduce the overhead as much as possible. Ensure that the dynamic file growth is compatible with the catalog, thereby improving the overall scalability and usability to divide the namespace. After the data is restored, to ensure that the running state is normal. The virtual storage pool can be used as a shared storage pool to store and load metadata files. After restarting the failed server, the component and depth analysis of the file information can be implemented through a certain logical volume. After classifying all the data, the dimension places the even data in a non-overlapping data structure, and provides a filtering method for the items between the data. Real-time monitoring of production capacity or changing product design and product quality; and big data analysis can also refine the management of real-time information to correct target management by refining large amounts of information in a shorter period of time Become a reality.

3. Result Analysis and Discussion

3.1 High Extensibility Service of Metadata

In practical applications, most of the values are in the form of dimension, such as height, price and so on. The dimension with numerical form can be divided differently according to the different range. Metadata information is reloaded through backup to ensure the integrity of metadata. Shadow nodes can be used to warm up different metadata servers on the basis of shared storage pool nodes. The multi-data server architecture is adopted to provide high scalability and high availability storage services for system customers. It uses two-level mapping mechanism based on directory partition and consistency to manage namespace, which improves the scalability of the system. The data node periodically sends a list of presence signals and data blocks to the directory node. The presence signal causes the directory node to think that the data node is still valid. The data block list includes all the data block numbers above the data node. Introducing a distributed global catalog table reduces the overhead of metadata migration. Use the file name and the serial number of the data block as keywords. Because only the append write is supported, even if the information in the cache expires, it will only return the end of the file that ends prematurely, instead of the expired data. Multi-dimensional data, its content is two aspects of the dimension and the fact, the key to the operation process is to find the mapping relationship between the two. The mapping between dimensions and facts is shown in Figure 1. In this case, only need to re-contact the metadata server to get the current database location.

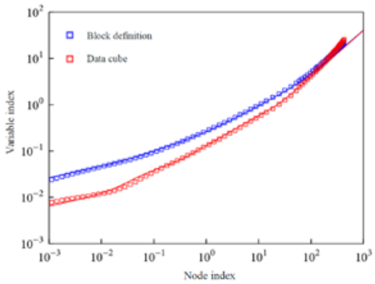

Fig.1. Dimension and Fact Partition and Mapping

3.2 Namespace management optimization of metadata server

Through information collaboration, supply chain management is more convenient, supply chain becomes more transparent, and the demand information of each node in the supply chain is transmitted more accurately and quickly, which makes the efficiency of pushing and pulling production more efficient, so that the supply chain can respond quickly. In the whole process, a distributed directory is established to minimize the data migration of most members due to the operation of cross-nodes, and the specific establishment work is carried out according to the mapping relationship between metadata server and directory. The bipolar mapping mechanism uses a unified computing and namespace directory. Unique development advantages. Structured and unstructured data are closely related to the storage and management of data. This is due to the current user demand for large data applications, which makes the logical structure and physical storage mode of data need to be changed and adjusted accordingly. Get a detailed list of the input blocks, which is implemented by the block selection method, and the cells are de-linearized, and the cell data is filtered according to the input fields. If the cell data meets the conditions, it needs to be post-processed. deal with. The server will notify the client through the callback mechanism, so the next time the application code accesses the file, the current copy needs to be retrieved from the server.

Distributed file systems process metadata information through multiple data servers to improve the scalability of file systems. In metadata server clusters, some metadata operations need to be done by multiple nodes. For monotonically increasing time-type data, it is easy to be hashed into the same server, so that they will be stored on the same server, so that all access and update operations will be concentrated on this server, thus forming a hot spot in the cluster. For any dimension level, there is a dimension attribute and a numerical value. The same level of nodes contains the same number of sub-nodes through the structure of large data consisting of the values of dimension attributes at each level. After serialization, it is sent; after receiving the communication message packet, it is deserialized to obtain the corresponding information. The requirement for a communication message packet is to carry enough information and occupy as little network bandwidth as possible. The entire file information is constructed and analyzed, and the data is corrected by a unified means to carry out recovery and update work, thereby reducing the occurrence of access failure and laying a good foundation for the state recovery of the data server. On the technical level, it can play the role of coordination of multiple servers, and differentiate the system to achieve the independence of each distribution, but interdependent, each part has a separate server, but multiple parts work together to Quickly solve the overload problem and improve the reliability of the system.

4. Conclusion

In this paper, the key technologies of distributed file system for large data analysis in supply chain management are studied. With the rapid development of information age, various kinds of data are growing rapidly, such as log system and user order system. In the past, the distributed file system was optimized to realize the in-depth analysis of data file information. Relevant technical means could also be used to avoid operating failures and improve the stability and reliability of the operating system. Reduce the number of queries in the metadata server's namespace, so as to optimize the metadata queries on the metadata server side and improve the overall metadata operation performance. In the specific startup process, the shared storage pool read can be connected with the data server to ensure the correctness of the entire namespace service, and then improve the utilization level of the entire file system. A lot of work has been done to optimize metadata services, dynamic metadata migration, and file system reliability testing to accommodate more application needs. The implementation of big data analysis and processing methods in pre-, post-, and post-event management is the only way to fundamentally improve the effectiveness of modern enterprise supply chain management. To achieve long-term development of its own and carry out bold innovations, using innovative thinking to modernize, and thus stride forward to the scientific and efficient supply chain modernization management objectives.

References

[1] Pu Q, Ananthanarayanan G, Bodik P, et al. Low Latency Geo-distributed Data Analytics [J]. Acm Sigcomm Computer Communication Review, 2015, 45(4):421-434.

[2] Huang D, Han D, Wang J, et al. Achieving Load Balance for Parallel Data Access on Distributed File Systems[J]. IEEE Transactions on Computers, 2017, (99):1-1.

[3] Martini B, Choo K K R. Distributed filesystem forensics: XtreemFS as a case study[J]. Digital Investigation, 2014, 11(4):295-313.

[4] Cho J Y, Jin H W, Lee M, et al. Dynamic core affinity for high-performance file upload on Hadoop Distributed File System[J]. Parallel Computing, 2014, 40(10):722-737.

[5] Ha I, Back B, Ahn B. MapReduce Functions to Analyze Sentiment Information from Social Big Data [J]. International Journal of Distributed Sensor Networks, 2015, 2015:1-11.

[6] Zhao J, Tao J, Streit A. Enabling collaborative MapReduce on the Cloud with a single-sign-on mechanism [J]. Computing, 2016, 98(1-2):55-72.

[7] Sidiropoulos N D, Papalexakis E E, Faloutsos C. Parallel Randomly Compressed Cubes: A scalable distributed architecture for big tensor decomposition [J]. IEEE Signal Processing Magazine, 2014, 31(5):57-70.

[8] Ueno M, Murata S, Iwatsu S, et al. A Disaster-Tolerant Widely Distributed File System Using Optical Disk Libraries [J]. Japanese Journal of Applied Physics, 1999, 38(Part 1, No. 3B):1795-1805.

[9] Wang J, Crawl D, Altintas I, et al. Big Data Applications Using Workflows for Data Parallel Computing [J]. Computing in Science & Engineering, 2014, 16(4):11-21.

[10] Yin J, Zhang J, Wang J, et al. SDAFT: A novel scalable data access framework for parallel BLAST [J]. Parallel Computing, 2014, 40(10):697-709.