New research on Key Technologies of unstructured data cloud storage

Songqi Peng, Rengkui Liu[a, *], Futian Wang

State Key Lab Of Rail Traffic Control & Safety ,Beijing Jiaotong University, Beijing 100044, China

[a] rkliu@bjtu.edu.cn

* Corresponding author

Keyword: Unstructured Data; Cloud Storage; Database

Abstract: From the traditional to today's data network text files, pictures, mainstream audio and video, the Internet is gradually changing the data structure from unstructured data, which is unstructured data growth and a variety of network data storage management has brought new challenges. In this paper, various solutions of massive non structured data storage problems, summarize the key problems to realize unified storage of unstructured data, the design and implementation of an unstructured data using the data storage function of unified batch processing framework, solve all kinds of problems of the non uniform node data processing.

Introduction

With the rapid development of Internet, the relationship between enterprises and the Internet is more and more close. Many information flows through the Internet, which makes the data on the Internet now reach an unpredictable level. Maintenance management information needs a lot of manpower, technology and other valuable resources. These data are filled on the Internet, the vast majority of them have their own different formats of documents, pictures and videos and other unstructured data [1-2]. The management of unstructured data is considered to be a major problem in today's Internet technology, because the past can effectively structure data management tools and techniques for unstructured data and therefore not applicable. Many commercial applications have proved that the traditional relational database can manage structured data, but in recent years many rely on unstructured data network applications, network media development spawned in non relational database management structure, exposed more and more obvious limitations of the data, in particular the performance and reliability of the rapid expansion of the problem unstructured data show 3].

This paper studies the solution of various kinds of massive unstructured data storage, analyzes all the problems existing in the storage system, and summarizes the key issues to achieve the unified storage of unstructured data. Then, with a massive, heterogeneous and unstructured data association and other features for the storage problem, put forward through the unified storage management platform to solve the metadata management of unstructured data, unified data interface, consistency and key issues of heterogeneous storage and data availability, high integrated, and other types of storage facilities. And a mixture of various types of data storage problem selection mechanism to effectively through heterogeneous storage devices. At the same time, based on the unified storage platform, an unstructured data batch framework with unified data storage function is designed and implemented, which solves the problem of unified processing of heterogeneous data types.

Cloud storage technology

Cloud storage is mainly used to store large amounts of data to actively solve problems. It can not only provide specialized storage solutions, but also publish storage business separately. Cloud storage is an application model based on Web, which has the characteristics of low cost and extensibility. It is a service concept, not real memory nor specific device. Using connectivity to the Internet, users enjoy the ability to share the storage pool with shared cloud storage. Users do not need to know the contents of the system, do not need to know how to store, it is transparent to all equipment users, at any time and space authorized users can use the network connection to use cloud storage, cloud services 4-5. The cloud storage data architecture model shown in Figure 1.

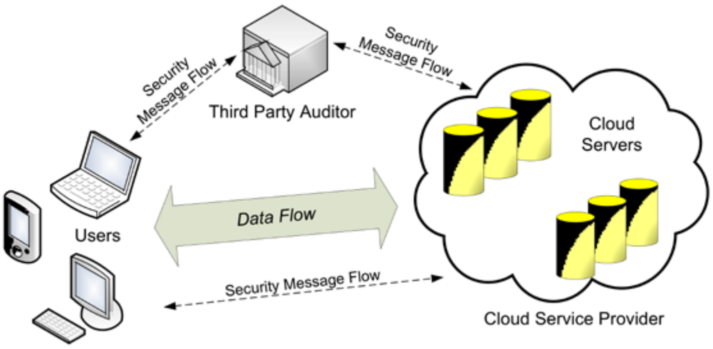

Figure 1. The architecture of cloud data storage service.

With the rapid development of modern network information technology, the information data has increased exponentially, formed in the era of big data, user generated data stored in the user data stored in the cloud environment has put forward higher requirements that need to be addressed: (1) efficient mass data storage and access requirements, users appear to hundreds of millions of monthly the dynamic query SQL data record, billions of dollars in relational database is inefficient, in the era of big data, the urgent need to solve the problem of data storage and efficient access to large amounts of data; positive development; (2) high concurrent read and write the database, the Internet network, the key to the user as the center, according to the personalized information the user needs to generate dynamic pages and information, such as the current micro Bo, this form of high concurrent access load data, usually form each Tens of thousands of seconds to read and write requirements; (3) high availability and scalability of the database requirements, system structure based on Web, it is very difficult to extend the database, when the user access to the database server is increasing rapidly, not simply the use of hardware and service node scalability and load balancing. Provide maintenance, upgrade and migration form stop uninterrupted data for web service requirements, will reduce the user experience: C4) support for unstructured data processing needs of the port, the relational database greatly limits the data processing and data types, various types of data can not be achieved in the future in the user requirements.

Unstructured data cloud storage hierarchy

The storage and use of unstructured data is very common. Many systems have to upload attachments, pictures, press releases and document management functions. At present, however, most implementations are stored by creating a writable directory on the server. Unstructured data is often larger, requiring more bandwidth and a certain computing power of the server, which has some impact on some of the server's high performance requirements. Server cluster synchronization. As applications require large-scale cluster support, the traditional approach will face more challenges. In order to synchronize data between nodes in each server, we need some similar network storage techniques to solve them. Many servers are uploaded to the server side of the Troy Trojan program invasion, most of these implementations because of vulnerability generated by uploading files. The storage requirements of the traditional file system for unstructured data must be the directory of the filesystem, which is 6-7 writable. Cloud storage is not required, and similar functions can be implemented using other methods, but technically advanced cloud storage has some advantages. Cloud storage is stored and read in the form of object storage, which is responsible for the actual content of the document. The high scalability, massive, high reliability and repeated file merging of cloud storage will help to improve the quality of storage service. The unstructured data cloud storage architecture is based on this design, as shown in Figure 2.

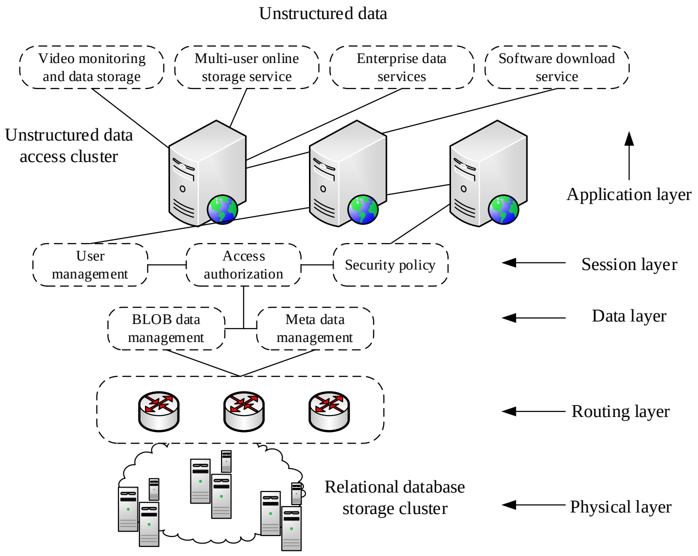

Figure 2. Cloud storage tier architecture of unstructured data

The application layer provides unstructured data application interface, the interface is composed of various types of data storage service providers to develop the storage applications, such as online storage, network drives, video, data hosting and software download service. At this point, users face a virtual, unlimited capacity cloud storage space without taking into account the physical location of the storage space and data when the user submits the data.

The session layer is responsible for user management, authority allocation, spatial allocation, and storage security policies. The layer relies on the security level and develops different security programs to ensure data security.

The role of data layer is the unified management of unstructured data and metadata. Unstructured data ranges from MB to GB, and size and metadata information, such as data identifiers, file length, type and other attribute information, the total length of not more than 1 KB, the difference in the amount of data between the two. Therefore, different data and metadata need to be stored on the network bandwidth, and computing resources should be used for different types of data storage strategies. Thus, figure 1 will be broken down into a data layer, a service data store, and a metadata store. The routing layer is responsible for cloud nodes, interoperability and storage path access interfaces and back-end storage device computing.

The physical layer of unstructured data storage provides storage space and computing resources, and is responsible for maintaining physical path storage nodes. The purpose of this system is to make full use of the existing communication subnet and equipment without adding hardware input.

Unstructured data cloud storage system structure design

In order to realize the effective management of unstructured data, many companies and individuals at home and abroad have done a lot of research. The most important management is divided into two categories: one is to convert the unstructured data based on the technology of semi structured data; the other is the conversion of unstructured data to structured data, the final data will be stored in a relational database. Unstructured to structured data conversion mostly adopts unstructured data, structured data and semi-structured data. Therefore, through the storage and management of relational database, the data structure is obtained. According to the requirements of the project, "structured data, unstructured data, semistructured data conversion method and gradually enlarge the application on the basis of the data structure of file metadata extraction concept structure standard to realize the conversion function of the template file name save the converted files, create a document template, unstructured data file table and structured data association, as shown in Figure 3 shows. The system is composed of database, file system, template library, file format definition module, metadata extraction module, template creation and management module, middle module, data representation and data conversion module. The whole system is divided into three layers: interface application layer, application logic layer and data storage layer..

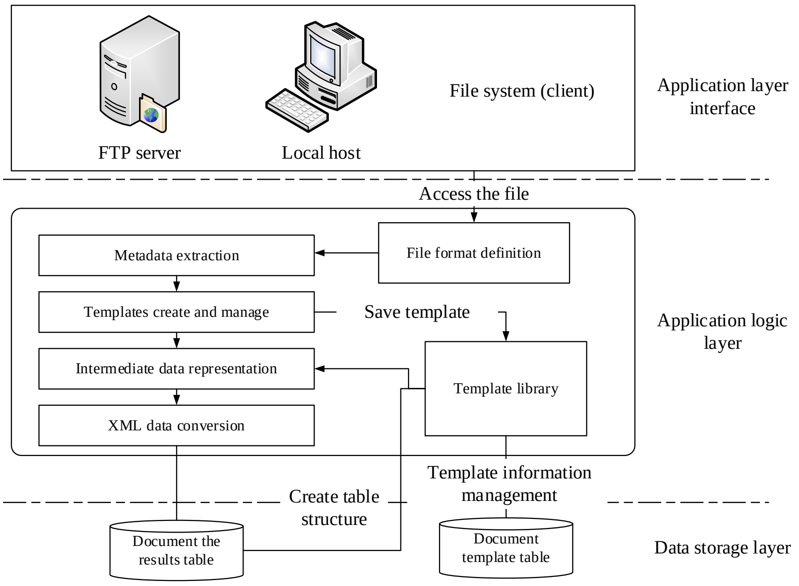

Figure 3. Unstructured data cloud storage system structure

The application layer interface provides the user interface graphical data interface of the application program, the user can use the structured unstructured data conversion operation, regardless of the specific data conversion.

The program logic layer is composed of five functional modules of the system structure, with emphasis on the implementation of the business logic structure to the unstructured data conversion system. The application layer interface client file system in the acquisition of analog output file, make a request for data conversion, then the application receives a request sent by the client, will need to convert the file transfer to the data conversion module. After the module receives the file, it determines which program to convert according to the type of file. Then, the work of the five functional modules, file metadata extraction, set up the appropriate document template, and then realize unstructured to semi-structured data conversion, the processed data is written in the database table simulation results. Then, the application results are converted back to the user, and the user is prompted for the next data conversion, and finally the entire process of data conversion is completed.

The data storage layer collects the database tables used by the system, such as document templates, document association tables, simulation results tables, and so on. The document template needs to create a document association table before the system runs. The data simulation table is the unstructured file data after the structured data is converted. When the data conversion is complete, the system will associate the relevant information to the file table.

Conclusion

Non structured data analysis based on the rapid growth trend on the Internet, introduces the solutions proposed by researchers at home and abroad, and non structured data storage, storage of these solutions can solve the massive unstructured data, and to ensure that the expansion of the system. However, different data types of unstructured data and different data have different storage characteristics, and how to store these different types of unstructured data in a unified way becomes an urgent problem.

This paper presents a unified unstructured data storage platform, unstructured data storage interface provides a unified model, combined with the underlying implementation of different types of unstructured data in heterogeneous storage, and in this heterogeneous storage infrastructure, to ensure the consistency of the data and the use of high. On this basis, combining a number of unstructured data structures on the storage platform, a large number of unstructured data resources and storage resources can be fully integrated in the processing process to achieve efficient data processing.

Acknowledgements

Fund Project: National Natural Science Foundation of China (51578057); State Key Laboratory of Rail Traffic Control and Safety (Beijing Jiaotong University) Independent research project (RCS2016ZT007) .

References

[1] Nicolae B. High throughput data-compression for cloud storage[M]//Data Management in Grid and Peer-to-Peer Systems. Springer Berlin Heidelberg, 2010: 1-12.

[2] Calder B, Wang J, Ogus A, et al. Windows Azure Storage: a highly available cloud storage service with strong consistency[C]//Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles. ACM, 2011: 143-157.

[3] Prahlad A, Muller M S, Kottomtharayil R, et al. Performing data storage operations with a cloud storage environment, including automatically selecting among multiple cloud storage sites: U.S. Patent Application 12/751,651[P]. 2010-3-31.

[4] Zhang D W, Sun F Q, Cheng X, et al. Research on hadoop-based enterprise file cloud storage system[C]//Awareness Science and Technology (iCAST), 2011 3rd International Conference on. IEEE, 2011: 434-437.

[5] Wang Q, Wang C, Ren K, et al. Enabling public auditability and data dynamics for storage security in cloud computing[J]. Parallel and Distributed Systems, IEEE Transactions on, 2011, 22(5): 847-859.

[6] Wang C, Ren K, Lou W, et al. Toward publicly auditable secure cloud data storage services[J]. Network, IEEE, 2010, 24(4): 19-24.

[7] Lin H Y, Tzeng W G. A secure erasure code-based cloud storage system with secure data forwarding[J]. Parallel and Distributed Systems, IEEE Transactions on, 2012, 23(6): 995-1003.