Key Technology Research for Unstructured Data Cloud Storage: New Exploring

Julan YI[1, a]

[1]XINYU UNIVERSITY, Xin Yu 338004, China

[a]julanyi@126.com

Keywords: Unstructured Data; Cloud Storage; Database

Abstract. From traditional to today's data network text documents, pictures, audio and video mainstream, the Internet is gradually changing the structure of the data, the data from the non-structured data, which are growing and a wide variety of unstructured data, the Internet data Storage management has brought new challenges. In this paper, for all kinds of massive unstructured data storage problems of the proposed solutions, summed up the key issues to achieve unified storage of unstructured data, design and implement a batch framework of a non-structured data using a unified data storage features, to solve the problem of unified treatment of various types of unstructured data.

Introduction

With the rapid development of the Internet today, the relationship between businesses and the Internet increasingly running close, many of the data flow of information through the Internet, so that the data on the Internet now reached a magnitude difficult to predict. Maintenance management information takes a lot of manpower and technology and other valuable resources. These data filled in on the Internet, the vast majority have their own different format documents, pictures and videos unstructured data [1-2]. The management of unstructured data is considered to be today's Internet technology is a major problem, because in the past can be effectively structured data management tools and techniques for unstructured data, it does not apply. Many commercial applications have proven traditional relational database can well manage structured data, but in recent years many rely on unstructured data network applications online media growth spawned found in the management of non-relational database structure When exposed more and more obvious limitations of the data, particularly in unstructured data after the rapid expansion of the magnitude demonstrated performance and reliability issues [3].

This paper studies the problem for all types of mass storage of unstructured data solutions proposed, analyzed all the problems of the storage system, which summed up the key issues to achieve unified storage of unstructured data. Then, for the storage problem has massive, heterogeneous, and other characteristics associated with unstructured data, put forward a unified storage management platform for unstructured data by addressing metadata management, unified data interface, heterogeneous storage and high availability of data and The key question consistency, integration, and other types of storage facilities, and heterogeneous storage facility by the selection mechanism to address efficient mixing of various types of data storage problems. At the same time, based on a unified storage platform, designed and implemented a batch framework of a non-structured data using a unified data storage features, to solve the problem of unified treatment of various types of unstructured data,

Cloud storage technology

The main cloud storage for storing massive amounts of data to actively solve the problem, it will not only be able to provide specialized storage solutions can also be released separately storage business. Cloud storage is a Web-based application model unique pattern, which is characterized by low cost, scalability, etc., is a concept of service, not real memory, nor is it specific device. Use connected to the Internet, users enjoy the ability to share cloud storage access storage pool. Users do not need to understand the content of the system, do not need to know how to store, for users of all devices it is transparent, at any time and space a legitimate authorized users are able to use the network to connect cloud storage, using cloud services [4-5]. Cloud storage data architecture model shown in Figure 1.

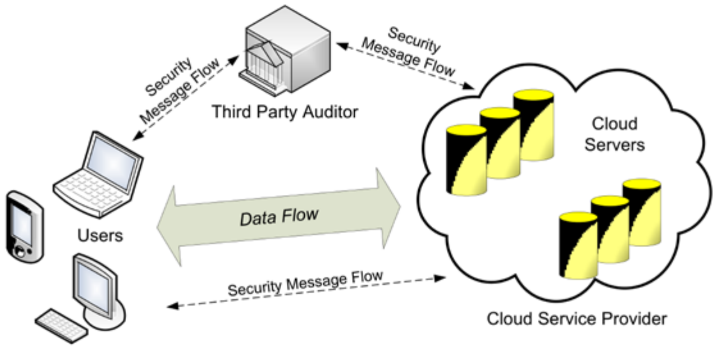

Figure 1. The architecture of cloud data storage service.

With the rapid development of modern information technology network, the data amount of information is growing exponentially, in the era of the formation of large-scale data, users generate a higher demand for storing data in a cloud environment for users to store data needs to address: (1) efficient storage and access massive data requirements, user appears every month up to hundreds of millions of dynamic, in a relational database using SQL query billions of data records table, inefficient, big data at times, the urgent need to solve the problem for efficient storage and access large amounts of data; positive development; (2) high concurrent read and write database requirements, Internet, Web emphasis on the user as the center, according to the user personalized information needed to generate dynamic pages and information, such as The current micro-Bo, this application form a high concurrent access to data load requirements, usually formed tens of thousands of times per second read and write demand; (3) high availability and scalability of database requirements, Web-based architecture , it is difficult to extend the database level, when a rapid increase in the amount of users and access to the database server cannot simply take advantage of hardware and service node scalability and load balancing, some of the requirements for the site to provide uninterrupted service in terms of maintenance and upgrade form Stop and migration data, will reduce the user experience: C4) support for handling unstructured data requirements, relational database significantly constrains the processing of data and data types, cannot be achieved in the future user requirements for various data types.

Unstructured data cloud storage hierarchy

Unstructured data storage use is very common, many systems will have to upload attachments, images, press releases and document management functions. However, most of the current implementation is by creating a writable directory on the server to store. Unstructured data is often relatively large, it will take up more bandwidth and a certain server computing power, which has some influence on some of the high performance requirements of servers. Server cluster synchronization. When the application needs to scale clustering support, the traditional way will encounter more challenges. In order to synchronize data between the nodes within each server, we need some similar sort of network storage technology to solve. Many servers are invaded by Trojan uploaded to the server implementation, and most of these implementations is because the vulnerability upload files generated. Traditional file system storage requirements for unstructured data must be a directory of the file system is writable [6-7]. Cloud storage is not required, to use, and other methods can achieve similar functionality, but technically advanced cloud storage has certain advantages. Cloud storage in the form of object storage for storing and reading in charge of the actual content of the document. Cloud storage of high stretch, massive high reliability, duplicate files merge, will help improve the quality of storage service. Unstructured data cloud storage architecture built on this design, the hierarchy shown in Figure 2.

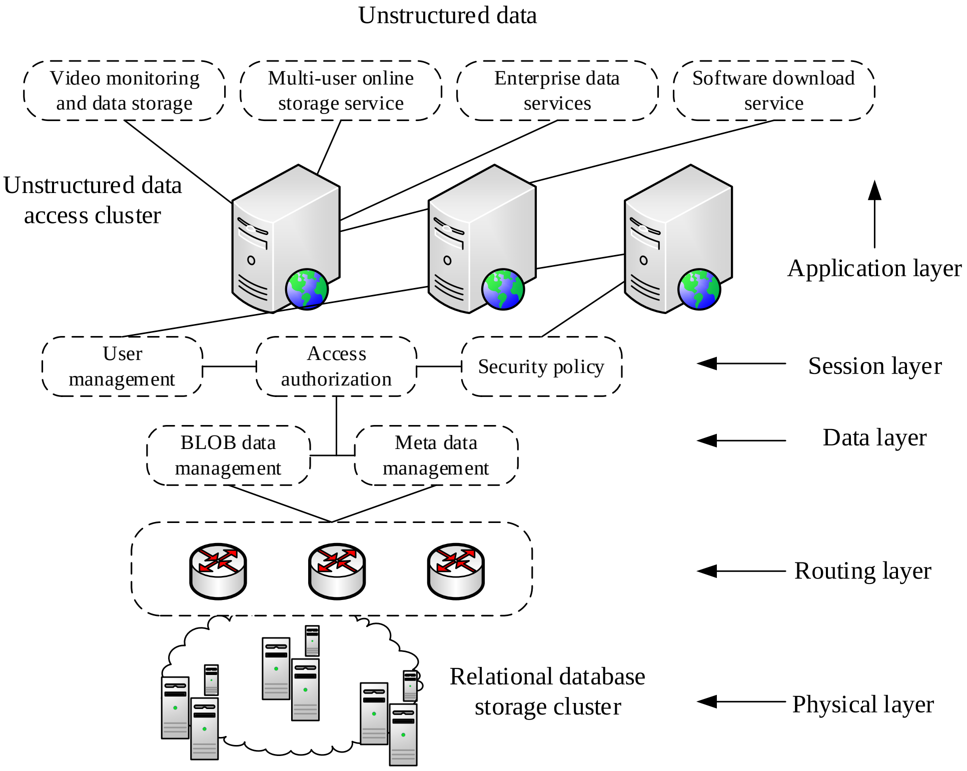

Figure 2. Cloud storage tier architecture of unstructured data

Application layer provides unstructured data application interfaces, these interfaces various types of storage applications developed by the data storage service provider in displayed, such as a variety of online storage, network drives, video, data hosting and software download services. At this point, the user is faced with virtual, capacity can be infinitely scalable cloud storage space, without regard to the physical location of the storage space and data when the user submits the data.

Session layer is responsible for user management, rights assignment, space allocation and storage security policy, the layer depending on the security level, the development of different security programs to ensure data security.

The role of the data layer is a unified management of unstructured data and metadata. Unstructured data volume level from MB to GB level, sizes, and its metadata information, such as data identification, file length, type and other attribute information, the total length of no more than 1 KB, the amount of data on the differences between the two. Thus, the different needs of BLOB data and metadata is stored on the network bandwidth and computing resources should be used in different types of data storage policies. Thus, Figure 1 will be decomposed into BLOB data layer service data storage area and a Metadata store. Routing layer is responsible for the cloud node, interoperability and storage path access interface and back-end storage devices calculations.

Physical layer unstructured data storage provides storage space and computing resources, and is responsible for maintaining physical path storage node. For the purposes of this system can make full use of existing communication subnet and devices without the addition of further investments in hardware.

Unstructured data cloud storage system structure design

In order to achieve effective management of unstructured data, many domestic and foreign companies or individuals be a lot of research. The most important management is divided into two: one is based on technology, semi-structured data to unstructured data conversion; the other is unstructured data to structured data conversion, data will eventually be stored in a relational database in. Unstructured to Structured Data Conversion mostly used the "unstructured data, structured data half a structured data" gradual conversion. Thus, the structure of the data obtained through the conversion of its relational database storage and management. Based on the project requirements, the use of "unstructured data structured data half a structured data" gradual conversion method and further expand on its basis, the concept of the standard structure members to implement the data structure of the file name conversion versatility introduction of templates to save the converted file to extract the file metadata, create document templates, documents related table to achieve the association unstructured data with structured data, as shown in 3 show. System consists of database, file system, template libraries, file format definition module, metadata extraction module, template creation and management module, intermediate module data representation and data conversion modules and other components. On the whole system architecture is divided into three levels: the interface application layer, application logic layer, data storage layer.

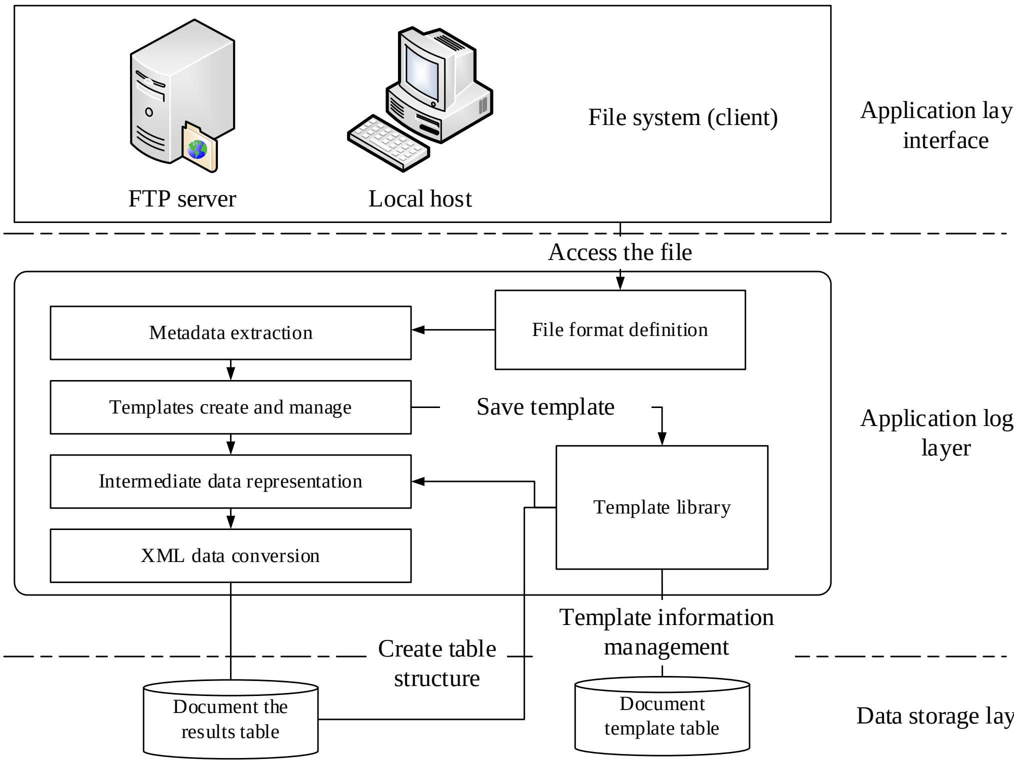

Figure 3. Unstructured data cloud storage system structure

Interface application layer provides a graphical interface to the user data conversion, through the application interface, users can use unstructured to structured data conversion related operations, without having to be concerned about the specific data conversion.

Program logic layer consists of five functional modules of the system structure, work focused on achieving business logic structured to unstructured data conversion system. Interface application layer client file system after obtaining simulation output file, issue a request for data conversion, then, the application receives the request sent by the client, will need to convert the file is passed to the data conversion module. After the module receives the file, depending on the file type classification to determine which program to use to convert. Then, five functional modules to work, extract metadata of the file, establish the appropriate document templates, and then implement the unstructured to semi-structured data conversion, the processed data is written to the simulation results table in the database. Application and then convert the result back to the user, and prompts the user whether the next data conversion, to finalize the whole process of data conversion.

Data storage layer collection system used by the database table, such as document templates, documents associated table, the simulation results tables. Document templates, documents associated table needs to be created before the system is running. Data simulation result table is unstructured file data after converted structured data. After the data conversion is completed, the system will associate the relevant information into the file table.

Conclusion

Based on the analysis of unstructured data on the internet rapidly growing trend, introduces the research staff at home and abroad for unstructured data storage problems caused by the proposed solutions, these solutions can solve the massive non-structural data storage problems, and to ensure the expansion of the system. However, a variety of data types unstructured data, different data have different storage characteristics, it remains an urgent problem that how can these different kinds of unstructured data stored unity.

This paper presents a unified storage platform outside of unstructured data to provide a unified model of unstructured data storage interface, the underlying combined and implemented for different types of unstructured data across heterogeneous storage, while in this heterogeneous storage infrastructure ensure high availability and consistency of data. Then combined on the basis of this storage platform on the frame of a batch of unstructured data architecture, making the process vast amounts of unstructured data in computing resources and storage resources can be fully integrated to achieve efficient data processing.

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.