Postprint: Storage Solution for Historical Solar Physics Observational Data in China

Authors: Lin Ganghua

Date: 2017-10-20T00:00:00+00:00

Abstract

Astronomical users achieving optimal outcomes from data services depends on factors including astronomical data storage methodologies, storage convenience, data security, and the maintainability of data storage services—requirements that are fundamental to data storage and sharing in all medium- and large-scale data-centric projects. This paper analyzes the storage requirements of domain-specific projects, examines differences among various storage service architectures, proposes the adoption of a cloud storage architecture, and designs a domain-specific cloud storage service architecture. This architecture not only fulfills requirements spanning from data processing to unified storage and unified external services, but also delivers an optimal experience for user data query services. Finally, the paper addresses the establishment of disaster recovery systems and related specifications.

Full Text

Preamble

Astronomical Research and Technology
Vol. 13 No. 2, Apr. 2016
Storage Solution for Historical Solar Physics Observation Data in China

Lin Ganghua
Key Laboratory of Solar Activity, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China
Email: lgh@nao.cas.cn

Abstract

Astronomical users achieve optimal results from data services when key factors are considered: storage methods, convenience, data security, and ease of maintenance. These represent essential requirements for data storage and sharing

in every large and medium-sized data-related project. This paper analyzes the storage requirements of domain-specific projects and explores differences among various storage service architectures. Based on this analysis, we propose adopting a cloud storage architecture and provide a design scheme for a domain cloud storage service architecture. This architecture not only meets the demands from data processing to unified storage and unified external services, but also offers the best user experience for data query services. Finally, we discuss the establishment of a disaster recovery system and related specifications.

Keywords: Cloud storage; Network; Data; Service

Historical Context and Data Characteristics

Due to limitations of historical astronomical observation technology, observation data were recorded on fragile media such as film and paper. These storage media deteriorate over time—for example, silver bromide shedding from film causes images to lose their original integrity and become unrecognizable, while paper media yellows and molds, rendering the data unusable.

Solar physics observations in China originated at the Shandong Qingdao Observatory, which conducted sunspot observations and manually recorded sunspot parameters, continuing for multiple solar activity cycles. Other domestic stations conducting joint sunspot observations include the Zijinshan Station of the Purple Mountain Observatory and the Fenghuangshan Station of the Yunnan Observatory. Historical solar observation data in China also include solar transverse magnetic field data. Departments providing historical observation data include the Huairou Solar Observing Station at the National Astronomical Observatories in Beijing, the solar activity forecasting department at the headquarters of the observatory, and the Space Science Institute at Nanjing University. Storage volumes range from several terabytes to petabytes.

These departments constitute the main organizations for solar activity monitoring and forecasting in China. China' s solar physics observation data possess regional advantages and feature multiple varieties. Internationally advanced observation equipment has produced first-class data that are also scarce internationally. Currently, preliminary data processing work is conducted at multiple locations, presenting characteristics of data dispersion. These are non-renewable precious resources that can provide systematic or case-specific data for scientific research, fill data gaps, and serve solar activity forecasting research.

Therefore, the digitization and standardization of China' s solar physics observation data have received special funding from the Ministry of Science and Technology as a basic work project, enabling these precious data to be preserved and ultimately serve solar physics research in China and worldwide. After data processing, digitization, and standardization are completed, the system will eventually make the data available for user queries on the China Solar Physics Portal according to established rules, with corresponding processing software available for use. According to plan, after various types of historical data are processed,

they will be gradually uploaded to the portal website' s servers. Once the storage system is completed, various departments can directly process their data within this system, which will automatically continue to upload them to the portal website' s servers. These historical data can be integrated with current daily observation data to form more complete Chinese solar physics observation data across solar activity cycles for user query and use.

Data backup and disaster recovery must be considered, and the system design incorporates comprehensive data backup and disaster recovery solutions. While data providers specialize in their own data processing and simple data archiving is relatively easy to achieve, for most data, proper long-term preservation requires more sophisticated solutions.

Cloud Storage Architecture: Rationale and Advantages

Introduction to Cloud Storage

Cloud storage [**?**] is a cloud computing system centered on data storage and data management. It represents a new concept extended from the cloud computing paradigm, referring to a system that integrates various types of storage devices in a network through cluster applications, grid technology, or distributed file systems to work collaboratively and provide data storage and business access functions. It avoids the cumbersome problems of traditional storage technology that require knowledge of specific storage information such as device models, interfaces, and transmission protocols.

The storage architecture consists of the storage layer, basic management layer, application interface layer, and access layer. The storage layer is the most fundamental and important part of the cloud storage system. Storage devices can be fiber channel or other hardware; in this system, they consist of multiple storage arrays, which is also the core component of cloud storage.

The basic management layer features a storage device management system that provides centralized management of these storage devices, including logical virtualization management, storage status monitoring, and storage maintenance and upgrade services. With distributed file systems, network computing, and cluster technologies, the basic management layer—though the most difficult part of cloud storage implementation—can fully enable collaborative work among heterogeneous storage devices. The cloud storage system can coordinate operations to provide users with high-quality services. The basic management layer also includes data content distribution and other services such as data backup. Since these services are directly experienced by users, the success of the basic management layer determines whether the cloud storage system can successfully serve users.

The application interface layer serves as the communication bridge between cloud storage and applications, and is the most flexible component. This flexibility is fully realized through the development of different program interfaces

according to various user needs. This layer is responsible for network access, permission management, and other functions. The access layer directly serves users, who can access the cloud storage system according to different needs and obtain various services. It provides multiple service types and access forms to serve diverse user needs with unified services.

Key Considerations for Adopting Cloud Storage

Compared with previous storage solutions, we value cloud storage for the following characteristics:

(1) **Convenience of use**: The same storage system provides storage services for various terminals including servers and personal computers. The storage system adopts a mount approach, allowing each data provider's servers or computers to use cloud storage space as if it were local storage. This reduces development and maintenance costs for data providers, allowing each to focus more on their expertise in data processing without needing to consider storage infrastructure design.

(2) **Convenience of maintenance**: For data providers offering storage and computing resources, services including automatic collection of relevant content, unified content structure, automatic computing resource management, file search within relevant scopes, network drive file sharing, and rapid publishing to clients are all conducted under strict security control methods including VLANs, firewall rules, and load balancing.

(3) **Data security**: Retaining file modification history versions, automatic synchronization and sharing, the ability to recover mistakenly modified or deleted files at any time, and automatic synchronization and backup of system data without manual operation.

(4) **Storage expansion advantages**: Due to the separation of metadata and data, cloud storage systems have nearly unlimited expansion capabilities. Cluster storage differs from traditional storage methods in that it is not pre-partitioned into independent data spaces but is simply a directory. By aggregating storage space from various storage nodes, it achieves scalability of user space, allowing each terminal' s mounted space to adaptively expand or contract.

(5) **Data file sharing**: Since each terminal mounts only a directory, users can specify sharing a particular file with one or several users, enabling these users to operate on the file and achieving data file sharing.

(6) **File retrieval speed advantages**: Due to the separation of metadata and data technology and aggregated I/O performance, the bandwidth improvement is particularly significant for large file reading. When facing massive numbers of files, user retrieval speed can be several times faster than traditional architectures.

(7) **Long-term development perspective**: Building a cloud storage system represents the current optimal choice in terms of convenience for data source providers, system maintenance, data security maintenance, storage device utilization, data application development, and data retrieval speed. This comprehensive development trend will lead to future cloud storage products with simpler structures, more powerful functions, lower prices, and higher data security. Its high quality and data management capabilities can meet the needs of large-scale data storage and computing for subsequent multi-band analysis [**?**].

Cloud Storage Architecture Design

Based on the specific objectives mentioned above, we designed a cloud storage architecture. The architecture diagram is shown in [FIGURE:N]. To implement the cloud storage architecture, the system is divided into several network types outside the virtual router: a shared network for data provider departments, a network for communication between management servers and system virtual machine management addresses, and a virtual local area network directly allocated for virtual machine use, which is divided into separate and shared types. Communication between the domain portal website and storage, or between storage virtual machines and storage, is all indicated by different colors.

Data provider departments are divided according to function into data provider segments and solar physics portal segments, each having internal and external networks. For data providers located on the public network end, an advanced resource domain network deployment mode will be adopted, using more network services and VLANs.

The cloud storage server side will be deployed at the headquarters of the National Astronomical Observatories, with maintenance and management handled by the information technology team of the Huairou Solar Observing Station. Clients refer to observation data production departments, such as the Huairou Solar Optical Observatory, Solar Activity Forecasting department, Purple Mountain Observatory, Yunnan Observatory, and Nanjing University Space Science Institute, among other related departments. Their main work is to process raw production data according to their expertise. Departments with more professional personnel maintaining data websites and large data output volumes can set up dedicated storage. For departments without sufficient professional personnel, the number of hard drives in ordinary desktop computers can be increased according to data volume.

To meet the digitization and standardization processing requirements, when the cloud storage system is fully implemented, all work will be conducted within this system. The database requirements are high, and an automatic processing mechanism already exists to timely add data to the database, enabling users to retrieve data through network services.

Regarding the data processing for the Huairou base group, one of the original

data providers in the data provider network, a single server can currently meet the data processing needs and will continue to be used for distributed computing clusters. The network server for the Huairou base group and the original storage array will be used for extended computing clusters and extended storage clusters respectively. All future data expansion will be based on this storage.

The virtual router provides address translation, virtual local area network (VLAN) allocation, virtual private network (VPN) setup, load balancing, and other functions for each customer account and each type of network. Virtual supervision servers typically complete virtualization functions. In the cloud architecture, primary and secondary storage are used together to achieve maximum efficiency and flexibility.

Disaster Recovery Methods and Specifications

Data Archiving Specifications

For different data types, create subdirectories under the main directory. Directory names should indicate data types. Data should be stored by type, then by year-month-day format. Corresponding processing software should establish directories under the main directory named after the software. Processing software naming should reflect the corresponding data type.

Each data production department should back up all data at least twice: one copy saved in the cloud, and another properly preserved locally on storage media. Simultaneously, designate specific personnel responsible for regularly replacing storage media with new-generation devices.

A disaster recovery server will be established at the Beijing Huairou Observing Base. For disaster recovery, regularly start the disaster recovery server remotely, transmit continuously updated data to this server, and designate specific personnel responsible for regularly replacing storage media with new-generation devices.

Consistency between each data producer and cloud data is achieved through login synchronization disks. According to data update frequency, determine the regular transmission schedule. After transmission is complete, shut down the server. What remains in the cloud is the final modified result, meaning data in the synchronization disk is the final data requiring backup, with automatic background synchronization. File and folder operations are completely identical to local resource manager operations, and changes by the data source provider will affect the other side.

Each data provider can adopt fixed backup rules based on data processing characteristics: daily, weekly, or monthly. To reduce workload and assign responsibility to individuals, establish a backup record table. Only perform incremental backups on data. Set all access permissions for backup data to read-only. The backup record table must include a field for the person responsible for executing the backup.

Conclusion

This system adopts a cloud storage architecture, providing the best approach for data security and sharing, and laying a good foundation for future solar physics observation data integration. This cloud storage system can be further applied to the storage and integration of modern solar physics observation data in China.

References

[1] Cloud Storage Analysis. Beijing: People' s Posts and Telecommunications Press.

[2] China Cloud Storage Development Report. Liu Bingwei, Chen Yu. Beijing: Electronic Industry Press, 1-10.

[3] Shen Dan et al. An adaptive process-based cloud infrastructure for space situational awareness applications. Proceedings of SPIE. 2014, 5450-5453.

[4] Megino F H B, Benjamin D et al. Exploiting virtualization and cloud computing in . Journal of Physics Conference Series, 32011-32022.

[5] Serfon C. Data management tools and operational procedures in ATLAS example of the German cloud. Journal of Physics Conference Series, 42053-42057.

Note: Figure translations are in progress. See original paper for figures.