“Information Assessment System” Supporting the Organization and Utilization of Data Stored on File Servers

MUROI Yasuyuki, MUKAI Yoshikazu

Abstract

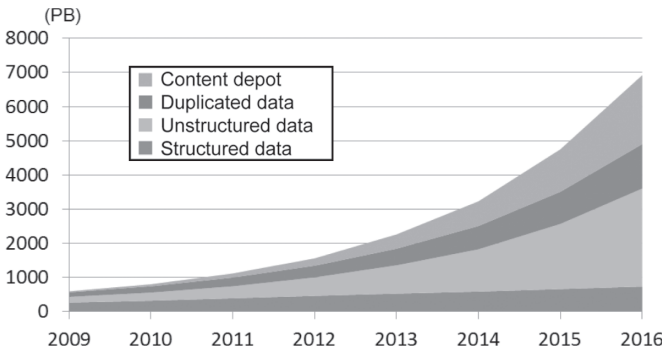
The information explosion is becoming a real issue, and the amount of information stored on file servers is continuously bloating, making the identification, organization and utilization of information on file servers difficult jobs. The latest version V2.1 of the Information Assessment Tool, a tool for “visualization,” “slimming,” “activation” and “optimization” of file servers, adopts the InfoFrame DataBooster high-speed data processing engine to deal with large-scale file servers and enable interactive analysis based on high-speed search/aggregation.

Keywords

file server, bloating, file organization, information utilization, memory DB, NAS, unstructured data, NIAS, information explosion

1. Introduction

The data used in enterprises can be roughly classified into structured data stored in databases, etc. and unstructured data such as documents, images and logs stored in file servers, etc. Compared to structured data, for which a data management method is defined, the standard management method for unstructured data is not established, so the growth of such data tends to make information management difficult. According to a survey by IDC Japan on the expected average annual growth rates from 2011 to 2016 of data stored on storage devices, the growth rate of structured data is 13.4% while that of unstructured data is 52.2%, which means an explosive increase in the amount of unstructured data (**Fig. 1**).



Source: IDC Japan, “Domestic market in disk storage systems: analysis of 2011 and forecast for 2012 to 2016”
TABLE112: Domestic disk storage system consumption model: shipping forecast classified by segments, 2009 to 2016

Fig. 1 Change in total storage capacity by data type.

With unstructured data increasing explosively, it is almost impossible to continue manual management of file servers. This makes it even harder to organize and utilize this continuously bloating information.

We noticed this unstructured big data existing in every enterprise and growing significantly. This paper introduces the concept and functions of the “Information Assessment System,” which was developed as a tool for facilitating the identification/analysis of file server situations and the arrangement/utilization of the stored data.

2. Development Background, File Server Issues

At NEC, we provided an “information assessment service” to more than 100 enterprises, which involved checking the customer’s file server status, analyzing the usage situation and proposing optimum information management and storage extension methods according to the usage situation.

Many customers of our information assessment service requested that we enable effective utilization of data and deletion/organization of unnecessary files based on the results of the information assessment, because simply extending storage cannot promote utilization of the data stored on the file

servers. To meet their needs, we advanced the development of tools that can be used by customers themselves to promote information utilization.

Problems Confronting File Servers

As described above, file servers used to store explosively increasing unstructured data are confronted by the following problems, from the viewpoints of system administrators and users.

One of the biggest problems from the viewpoint of administrators is the difficulty in identifying the file server usage situation (visualization). The expansion of the physical capacities of storage devices and the explosive increase in data quantity are interdependently making it hard to accurately identify the usage situations of file servers. If the usage situation is not identified accurately, it is impossible to deal with certain cases. For instance, when the available capacity of the file server is about to run out, the usual measure taken is to reserve capacity by deleting or moving files (slimming) because it is difficult to immediately extend file servers. However, the deletion or movement of files is not easy when the usage situation is not fully identified, because the system administrator does not have enough information to judge whether or not each item of stored information is necessary. Even when the administrator tries to get confirmation from users, it takes a long time to select the files to be reduced from out of a huge number of files.

A problem important for information security is the optimum management (healthy maintenance) of file server access rights, but it is not easy to determine all the access rights settings and maintain healthy conditions. To ensure information security in the operation of an enterprise's file servers, it is necessary to set appropriate access rights for each organization or project. However, as the authority management function of Microsoft Explorer is incapable of batch confirmation of authority conditions, much labor is required to check the settings of a very large number of individual file and folder authorities.

One of the problems from the viewpoint of file server users is the difficulty in the utilization of information (activation). It is not unusual that a file that should be there cannot be found or that the latest file cannot be identified from a large number of similar files. With the bloating of information, the search efficiency of stored information is dropping every year and a perspective on the information is often lost. How to find and use important or beneficial information from file servers out of a "chaotic mixture" situation is currently a big problem for all users.

3. Functions Provided by the Information Assessment System

To meet customer needs related to file server problems, we launched the "Information Assessment System V1.1" in August 2011. This product is a file server organization/utilization tool based on three functions: the "visualization" for identifying file server information that has already been provided by the existing information assessment service, "slimming" for the organization of bloating information and "activation" for promoting the utilization of information stored on file servers.

The Information Assessment System V2.1, shipped in July 2012, includes a file server "optimization" function that enables visualization and resetting of file server authority settings in addition to the three functions above (**Fig. 2**).

3.1 Visualization

The first important matter for the organization/utilization of file servers is to identify their usage situation. The Information Assessment System collects information on the files stored on file servers and "visualizes" the usage situation in the form of a graph or report. This visualization facilitates identification of the overall situation of the file servers and confirmation of the causes of their bloating in various aspects. For example, it makes it possible to read trends, such as the fact that a small number of files occupy a large portion of the disk or that there are a large number of files that have not been referenced or updated for a certain period or that are duplicated, and the results can be used as a reference for determining countermeasures.

3.2 Slimming

File organization not only contributes to disk capacity saving, it can also improve information search efficiency. To enable "slimming" through file server organization, the Information Assessment System provides a function to "narrow down" unnecessary files under specific conditions and organize file servers by means of deletion, movement and compression. It is not realistic to confirm all of the unnecessary files from out of a huge number of files. To save disk capacity

"Information Assessment System" Supporting the Organization and Utilization of Data Stored on File Servers

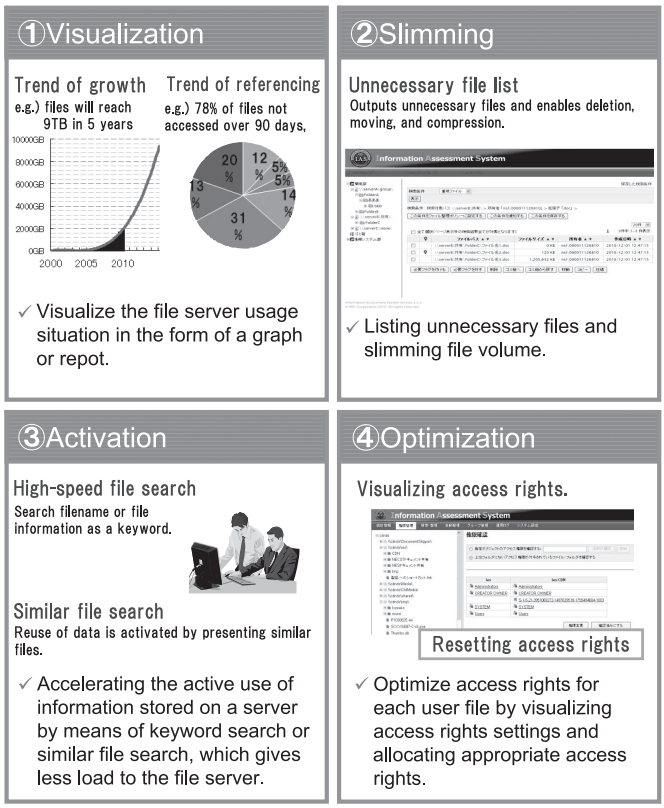


Fig. 2 Features of the Information Assessment System.

efficiently with less labor, the Information Assessment System organizes files efficiently by adjusting organization conditions, selecting a realistic number of organization target documents and applying the established organization guidelines.

Considering the difficulty for the administrator to confirm the necessity and importance of individual files for organization, the Information Assessment System defines three levels of management authorities for organization: the "system administrator" who manages the file servers, the "group manager" who performs management at the department level and the "general user." It requests organization and confirms files from these three levels and organizes files based on interconnections between these levels.

The system is capable of auto organization of files as well as manual organization. When the organization conditions are set in advance, files matching the conditions will be organized periodically. The auto organization function enables automatic operations according to a policy, for example moving files that have not been accessed for a certain period and periodically deleting temporary-domain folders.

With file organization based on movement, required files are moved to a secondary storage such as iStorage HS. This contributes to physical capacity saving by effectively using the deduplication/physical compression functions of the storage and also supports the effective storage of information.

3.3 Activation

The effective utilization of information stored on file servers requires the retrieval of desired information from out of a large amount of stored information. To meet this need for information search, the system provides a function for searching for target files by specifying various conditions such as filename and date of last update. This function can promote the active use of file servers in routine server jobs.

3.4 Optimization of Authority Settings

The system provides a function to optimize file server authority status by visualizing file server access rights settings, finding inappropriate access rights and resetting them.

This function can confirm access rights set inappropriately in the file/folder hierarchy as well as files that can be accessed by specific users simultaneously. Together with the file server proprietary rights modification function, it can deal, for example, with the access authority resettings following a personnel reshuffle.

4. Information Explosion Countermeasures

4.1 Data Management Issues with the Earlier Version

With the Information Assessment System V1.1 (hereinafter abbreviated as "V1.1"), we used a relational database (RDB) for the management of information on file servers. As the processing of big data collected from servers takes a lot of time, it was necessary to aggregate data by means of batch processing under the conditions decided at the time of information collection. Although the advancement of the information explosion has made file servers with capacities of some tens of TB no longer rare, V1.1 set the standard maximum capacity of each server handled by a single management server to around 10 TB, considering the RDB performance limit. A file server larger than 10 TB was handled by more than one

management server, but this led to the problem of an increase in the installation/operation costs of management servers.

4.2 Real–time Aggregation

With the Information Assessment System V2.1 (hereinafter abbreviated as "V2.1"), we reviewed the data management platform and adopted the InfoFrame DataBooster, a high-speed data processing engine capable of fast batch processing of big data - which was a weak point of the RDB - by using a memory database.

This has brought about a drastic increase in data aggregation speed, about 40 times that of V1.1, and has made possible interactive file server analysis/aggregation, which was impossible with V1.1 due to the problem of processing speed. Now file organization policy can be set effectively by narrowing down organization targets with detailed modifications of file organization conditions as required.

4.3 Challenge to Big Data

The adoption of the InfoFrame DataBooster allowed V2.1 to increase the standard maximum capacity that can be handled by each information management server from the 10 TB of V1.1 to 50 TB. This has made V2.1 compatible with the high-speed analysis/aggregation of large-scale file servers. If data quantity per file is assumed to be 500 KB, this 50 TB of data corresponds to 100 million files. We designed V2.1 to process this big data with a target of 32 GB of memory, which is the upper limit of Windows Server 2008 R2 Standard Edition.

If all the management data required for file server inspection and information aggregation were stored in the InfoFrame DataBooster, the memory required to operate the software would increase to an unrealistic value. To avoid this, we adopted a hybrid information management method in which the information required for file server information aggregation/search is stored in the InfoFrame DataBooster and other information is managed using the RDB and files.

Furthermore, thanks to the duplicated data compression function, managing data that is assumed to be duplicated frequently, such as file authority information, in a memory database has allowed V2.1 to increase aggregation/search speeds and save memory usage.

For performance considerations, the addition, updating and deletion of information are performed by batch processing, combining multiple data processing operations together to improve processing performance.

5. Conclusion

In the current information explosion trend, the information stored on file servers is continually bloating. With product technology that offers big data compatibility through the active use of memory databases and the Information Assessment System based on expertise cultivated through the information assessment service, we will continue to propose new ways to manage and utilize the information inside enterprises.

*Windows and Windows Server are registered trademarks or trademarks of Microsoft Corporation in the U.S. and other countries.

Authors' Profiles

MUROI Yasuyuki
Manager
3rd IT Software Division
IT Software Operations Unit

MUKAI Yoshikazu
Assistant Manager
3rd IT Software Division
IT Software Operations Unit

Information about the NEC Technical Journal

Thank you for reading the paper.
If you are interested in the NEC Technical Journal, you can also read other papers on our website.

Link to NEC Technical Journal website

Japanese

English

Vol.7 No.2 Big Data

Vol.7 No.2
September, 2012

Special Issue TOP