Enterprise-wide Storage Resource Management Solutions

Leverage innovation for file management efficiencies

Enterprise-wide Storage Resource Management Solutions

Leverage innovation for file management efficiencies

Table of Contents

Executive Summary

A clean house runs cost-effectively.

Managing storage is a universal problem for IT administrators. A company employee opens new documents and creates files on a daily basis in the course of business. In the course of non-business, that same employee may also download MP3 files, store pictures from the company picnic and watch video clips, all stored on his or her hard drive. Multiply that activity by the number of employees in an office and the number of regional offices in an organization and that's a lot of data to store. Additional storage will solve the problem, at least until that new space is filled, and all that expansion comes with substantial costs.

The key to gaining a firm grip on the information contained in the millions of directories and billions of files contained in an enterprise is learning more about what types of data are clogging the network. Data profiling and recovery allows administrators to differentiate between business and non-business information. Different data takes up larger or smaller amounts of space; critical business data needs to be stored, while MP3 files and home movies take up unnecessary space and need to be jettisoned from the network. This "house cleaning" enables administrators to begin the task of recovering space, which, in turn, leads to more efficient network operations.

Compare this to a homeowner who hates to clean. He continues to add furniture and clutter to his house. As rooms fill to capacity, the homeowner looks for a solution to his storage needs. Rather than looking internally and clearing out the clutter, he adds another room to his house. Once that room is full, he adds another room, and so on until he runs out of money for new rooms and time to manage all the additional clutter. Adding new space only serves as a temporary solution. He never learns how to manage the space he has.

Organizations today don't have the budgets to "add rooms" to their storage. They know they need to make an effort to go into each of these rooms and do some "spring cleaning," but see that effort as time-consuming. That's where storage resource management (SRM) comes into play. Many enterprise customers are slow to implement SRM solutions, failing to see the long-term benefit from taking the time now to put such data organizing systems in place. However, without SRM, administrators run the risk of losing control of their data, losing the capability for capacity planning and ultimately saving time and money.

FileCensus provides enterprise-wide storage resource management (SRM) solutions for organizations that need a cost-effective way to effectively manage their storage space. FileCensus looks at all files, determines what type of data is being stored, how much of that data is relevant to the organization's operation, and how space can be recovered within existing space. It shows where the largest storage problems are occurring, determines what information should be archived and where that long-term storage should be housed, particularly taking into consideration corporate governance and compliance issues that many organizations face. In addition, the software manages security issues that take into account restricted file access between groups within an organization. FileCensus gives IT administrators the power to clean up storage, not manage more, and plan for future storage needs.

Understand the Data

Recognize data storage patterns to recover capacity.

Data profiling, data recovery and capacity recovery work hand-in-hand in helping IT administrators gain a better understanding of their organization's data. Once they know the type of data taking up space in the network, they can see where cleanup is necessary. They learn what type of data is causing their storage issues, can work to remove the storage bottlenecks, organize the data and create a plan that allows for future storage growth.

For example, an organization might have a problem with employees copying music, pictures and movies onto their storage. A data profiling exercise reveals that 30% of the storage consists of non-business related data. The next step is to clean up that data. With this pattern of behavior detected, IT administrators can implement a process that runs a nightly report showing non-business data stored on the network and alerts the end-users who generated this storage that they need to clean that data off the network or it will automatically be deleted by the end of the week. This becomes a standard policy that runs automatically, leaving the administrator free to handle more strategic matters.

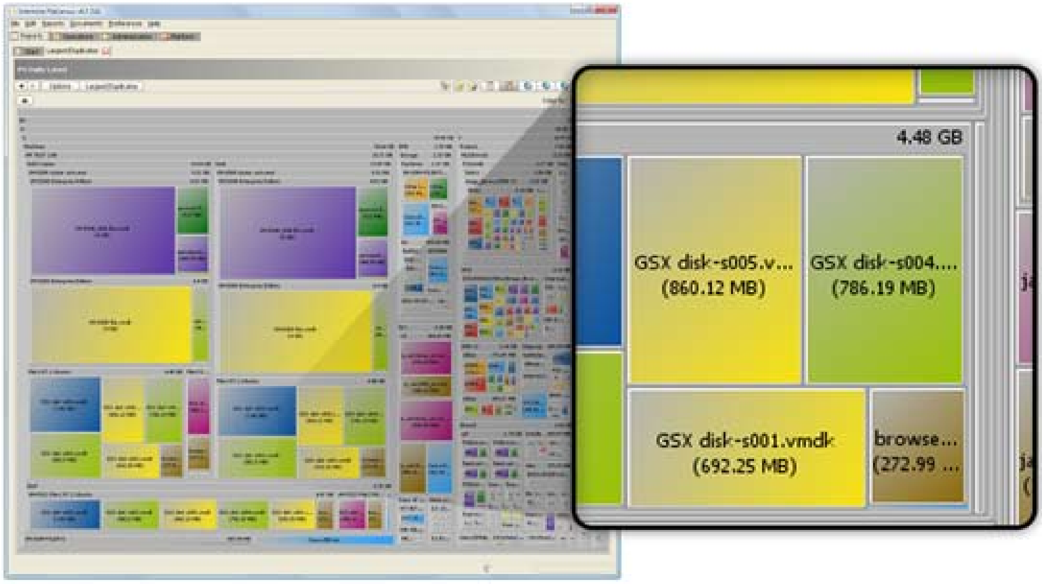

4.48 GB
GSX disk-s005.v... (860.12 MB)
GSX disk-s004.... (786.19 MB)
GSX disk-s001.vmdk (692.25 MB)
browse... (272.99 ...

Figure 1
Data Profiling & Recovery

Appropriate data storage saves money.

The biggest problem when trying to manage data is uncontrolled growth of information in the enterprise. This uncontrolled growth causes many complicated and interrelated problems in the network. Because this growth is moving so rapidly, administrators are having difficulties managing the data. Organizations also have only limited human resources to deal with these growth issues.

Differentiating between business data and non-business data is another issue. An example of non-business data would be employees storing movies, pictures, and music from home on their work computers, or downloading from the Internet. Clients have experienced more than 30% of their primary storage taken up by non-business data.

The impact of excess storage on an organization's primary tier of storage is expensive to consider. Primary tiers, which include Storage Area Network (SAN) and Network Attached Storage (NAS), are high performing and high quality storage solutions, but also expensive. Most companies will invest millions of dollars into the hardware infrastructure they have and are spending more annually to keep up with their data growth rate.

Clients report an average of 30% primary storage taken up by non-business data

Maintaining large amounts of storage on the primary tier has a huge effect on the entire organization. Every day, companies typically conduct incremental backups, only copying files that changed that day. Full backups are completed at the end of each week. With that, every file on the organization's network is stored on back-up tapes. If non-business related data fills 30% of the company's storage space, that means that 30% of the back-up tapes are filled with non-business related data. Backup tapes are kept for months, so eventually a company could have 20 or more copies of this unusable data. The overhead required to maintain this quantity of data involves not only the cost of storage but also the time it takes to back up this amount of information. Doubling storage every year on the primary tier might alleviate storage issues, but that also means the backup system needs to keep pace with that growth.

Most non-critical data is duplicated or potentially illegal

Restoring capacity can be time consuming when administrators have insufficient information regarding the data on their networks. Mission-critical data needs to be first and foremost, meaning administrators must work through and prioritize volumes of files in order to create efficiencies within the network. Without knowledge of what information is most important and where it's located in the network, the task of prioritization takes time and with that time, increases costs.

Find solutions that handle all data.

EMC Technology and Symantec each offer products that try to look at all the data stored on an enterprise network. The problem with these companies' products, however, is scalability. When looking into a large enterprise that has terabytes and, in some cases, petabytes of storage, that means billions of files. Many products currently in the market only scan for targeted data that they know potentially are a problem, such as MP3 and AVI files.

60 days of scanning in FileCensus would create a 270 Gigabyte database versus a traditional database at 9+ Terabytes

FileCensus is different because it acts as a census, while competing products act only as surveys. With a survey, results are based on a review of a small number of people. Trying to gain more widespread statistics by extrapolating those small-group results can skew statistics. However, with a census, the same question is asked to everybody, making it easier to gain more accurate statistics. A census results in the exact answer, providing an exact picture of what's happening.

FileCensus collects every file and folder with all associated Meta information about that server. That Meta information includes approximately 30 different properties – file name, file extension, file ownership information and all file attributes. The software scans the entire network, collects the information and sends it back to the FileCensus server. That means the image is kept historically and can be accessed by an administrator who may need to retrieve data from a week ago. The administrator can even run reports from that historic data. It's like rolling back the clock.

Time Warner Cable recovers space in one day.

Five years ago, Time Warner Cable contacted Intermine to archive their data. At that time, Time Warner had only 1.5 terabytes for storage, but that was a large amount of storage for that period of time. Their data was growing at a phenomenal rate but they couldn't figure out what was happening. Their administrators were spending time every day looking through a huge amount of data and they could not understand what was causing this phenomenal growth rate because they weren't expecting this type of behavior on their network.

With FileCensus, Intermine took only one day to discover what was causing the problem. Intermine recovered over 30% of their storage in one day. They found out that one employee had been saving data into his home directory every day and making duplicated copies of that directory. This was becoming an exponential problem. This one user was utilizing more than 1/3 of the storage on the company's network, and it was basically all duplicates. The administrators deleted all the duplicates and recovered that space.

This comes back to data profiling. Time Warner did not have a tool that would enable them to identify their problem and they were spending a huge amount of resources trying to find out what was causing the problem. It was like trying to find a needle in a haystack.

Analyst views SRM as mandatory for shared storage environments.

According to a 2007 Gartner report on storage resource management and SAN management software[1] *, an understanding of the need for SRM is growing. With that growth, two SRM tools have emerged: There are traditional, comprehensive offerings with added functionality and an emphasis on product integration and there is a relatively new, more-focused solution that deals with individual pieces of the storage management space.*

As noted in the report, a key aspect of SRM is the ability to identify new storage objects and collect that store them in the appropriate places in the network. Administrators must be able to collect and store data for review not only in its current state, but historical data views, with the goal of managing current data and planning for future storage needs.

[1] Dave Russell and Robert E. Passmore, *Magic Quadrant for Storage Resource Management and SAN Management Software, 2007* (Gartner RAS Core Research Note G00146578), March 19, 2007.

Plan for Storage Needs

FileCensus stores billions of records in a space more than 30x smaller than traditional technology

Analyze trends now to predict the future.

Capacity planning is the process of determining what type of storage you will need in your environment in the future, once the network has been cleaned up. FileCensus analyzes trends in stored information, using a collection of statistical and modeling techniques.

There is a finite capacity to a company's existing storage and employees keep filling up that space. Once that capacity fills up it can actually stop people from doing their job. For example, when an employee works on a Word document, that work increases the amount of storage space necessary for the revised document. When the employee tries to save that document, if there's not enough space on the network to store it, a message will pop up that states that the document cannot be saved due to insufficient storage space. Expand that further, and consider that there are thousands of people in that same situation. Storage fills up very quickly.

The IT administrator now has to manually expand storage and may also have to establish policies for saving files to the network. This complex process can be avoided with a capacity planning report that shows a more accurate determination of when an organization is going to run out of space on its servers. IT administrators can plan for the future more easily if they can see how storage is growing.

With this advance planning, companies can reduce or eliminate server over-provisioning, identify and repurpose underutilized servers, reduce IT operational and capital expenditures, reduce server downtime, improve server performance and availability, improve IT budget accuracy and leverage future technologies.

Not all files grow at the same rate.

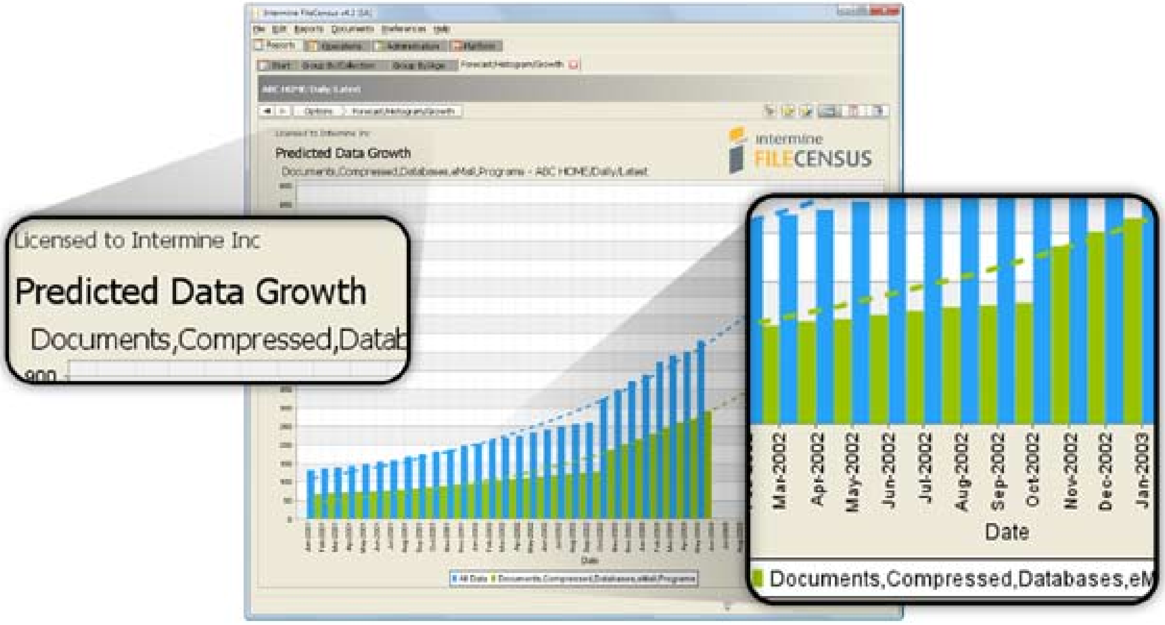


Figure 2
Capacity Planning

Large files, such as MP3s and video clips, take up significantly more space than Word documents and Excel files. However, if the network is only being scanned for certain files, they may not be the files taking up the majority of the space. Deleting smaller files won't alleviate storage issues. Part of the challenge is finding out exactly where growth is occurring most rapidly.

MP3s and video clips take up much more space than Word and Excel files

To get a better picture of what's filling up storage space, all of the storage must be scanned. A company considering an expenditure on hardware would be more comfortable knowing that the data collected and scanned included the company's entire storage, not just a small survey of one of its volumes. It's not effective to extrapolate data from one volume over the entire network. Not all volumes grow at the same rate. FileCensus scans everything and reveals the exact growth rate for each volume. Companies can also learn which files are causing that growth in the different volumes and at different rates.

In addition, there's the issue of predicting how much a company's storage needs to grow and at what rate that growth will occur. FileCensus reports extrapolate and determine a company's normal growth rate. Statistics will show the rate of growth if no changes are made in the storage environment. Further, FileCensus reporting can filter existing storage – only look at Word and Excel files and movies, pictures, music and other non-business related data from the growth rate – and predict the growth rate without the non-business data. In the same graph, companies using FileCensus can see what will happen in a year if they do nothing, but if the storage were cleaned up the growth rate would be less. This is important when considering future hardware purchases to handle storage. Two terabytes of storage, for example, could cost $100,000 to $200,000 in capital expenditures as of February 2008. But, with some cleanup work in the storage environment, the company may only need to buy one-half terabyte of storage, a savings of $150,000.

2 terabytes of storage could cost up to $200,000 in capital expenditures as of February 2008

Timing is another factor in the equation of when a company should purchase additional storage hardware. FileCensus forecasting reports can tell companies exactly what they need regarding storage space and when they need it. An accurate growth rate enables a company to accurately determine when they need to buy storage. That's important, as storage costs are always dropping. The importance of being able to wait a little longer means the price of storage will actually drop. This could allow a company to set up a phased storage increase implementation, possibly buying one terabyte of storage space in nine months and another terabyte six months later.

With FileCensus's forecasting capabilities, companies can predict the growth of data on a per volume basis. This report uses multiple historical snapshot images to track used space and total capacity for each volume over time. These observations are then used to predict future data growth by calculating a linear, straight line or trend line. Looking at high water mark data, this report attempts to predict when a volume's data will exceed capacity based on the line of best fitness. Companies can drill into each volume in the report to see the actual observation points and prediction curve for that volume.

A file-level view of data is best.

Most companies use an infrastructure tool for capacity planning, but that tool only provides volume information. That information is inaccurate, in that it doesn't see everything. That solution also doesn't go into the file level and cannot tell the difference between business and personal data. That is a really critical point. Their software just says you need to buy more storage.

Hardware vendors like these types of volume-predicting tools. With volume predictions, vendors can project that the storage is going to double in a year's time; therefore, the company needs to buy more storage. But this perspective overlooks what would happen if a company cleaned up its storage and removed some of the unnecessary data. Only in the last few years have companies looked for help in cleaning up their networks.

In addition to FileCensus, TeamQuest offers a capacity planning solution that surveys data to see where changes could alleviate storage problems. This offering tracks trends in data usage; however, this solution relies on extrapolation techniques to determine effects over time, rather than analyzing all data currently housed in a company's storage network enterprise.

More time for tactical management.

Capacity planning is such a useful part of the equation that nearly every customer uses this feature. Once data profiling, data recovery and capacity recovery have been completed, administrators have more time to be strategic and look at planning for the future. They can put more effort into the planning stages, which they can't do if they're always putting out fires.

Industry view on capacity planning.

As noted in a 2007 Gartner report on storage resource management and SAN management software[2]*, once data is identified and properly stored, and storage space has been recovered, the capacity planning component of SRM software requires analysis of the remaining data to assist in the prediction of future storage needs. The software analyzes trends within the data stored within the network and uses that information to predict future capacity requirements by server, department, application and/or enterprise. With this capacity, organizations can plan in advance for potential hardware or other storage needs purchases.*

[2] Dave Russell and Robert E. Passmore, *Magic Quadrant for Storage Resource Management and SAN Management Software, 2007* (Gartner RAS Core Research Note G00146578), March 19, 2007.

Financial Accountability for Storage Usage

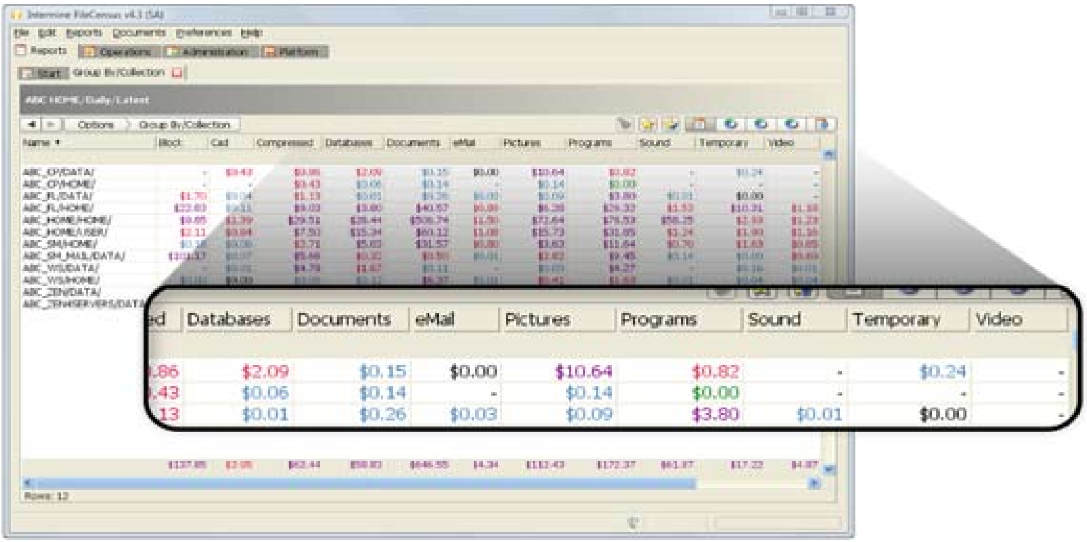
Customers typically face 60 – 80% storage capacity growth

Create accountability for data growth.

Chargeback, which works best with a clean storage environment, is a fundamental decision companies make to associate explosive data growth with internal storage budget constraints. The concept that storage infrastructure is a commodity has placed many organizations in a precarious position, where the true cost of storage and associated operational practices have dominated IT spending. Without active management practices, internal customers are typically facing 60-80% storage capacity growth, while hardware price reductions are not keeping pace.

To recover some of these internal costs, many organizations will charge internally for their storage. A large enterprise with thousands of users maintains storage in each of those centers. There is an internal process of buying the storage and administrators will "charge back" to each of those areas for using certain amounts of storage space.

For example, a company with offices nationwide would bill each office for their storage usage per month. If an office wants more storage space, they have to pay for it. That money goes back to the central administrative group, which can use those funds to buy more storage and roll it out to the rest of the enterprise. It's a more mature storage management process. Sometimes, companies will implement this to force regional offices to clean up their storage. Finance departments ask why the increased storage costs are occurring and can then help enforce the cleanup policy as a way of cutting costs.



ed	Databases	Documents	eMail	Pictures	Programs	Sound	Temporary	Video
.86	$2.09	$0.15	$0.00	$10.64	$0.82	-	$0.24	-
.43	$0.06	$0.14	-	$0.14	$0.00	-	-	-
13	$0.01	$0.26	$0.03	$0.09	$3.80	$0.01	$0.00	-

Figure 3
Chargeback

This also comes back to the fact that FileCensus gathers information about all the data, determining specifically what's business-related and what's non-business related. Invoices sent to regional offices note exactly where the problems are and what information should go on their bills. Central office administrators need to be able to itemize bills for the chargeback from regional offices. A bill can be broken down into individual users and individual directories and files, by type of file, and so on. Then the regional office can clean up based on the specific items on the bill. It makes the regional offices aware of where their storage problems are and ultimately why they were charged that amount. The bill is part of the chargeback process but it's also a way of identifying where the problems are and where their storage needs to be cleaned.

This allows larger companies to delegate branch office storage management to administrators in those regional offices and not managed by one central IT team. The regional IT people are better able to handle their storage issues anyway. This frees up IT management for more long-term planning and strategic efforts.

The payoff is ultimately for a company to get the regional offices to pay for what they use. It keeps tabs on uncontrolled growth rates and keeps the central office from having to pay for increased storage needs in the regional offices.

The end result for the users is that they have more space for storage. However, it may come at the cost of having to take time to clean up their files on a weekly basis. The administrators gain a lower chargeback cost.

Drill down to the file level.

Just as with the other features, the few chargeback products in the market operate at the volume level only. They measure volume growth rate and the cost of that growth, but provide no detail about why the growth occurred and where it is concentrated.

Scan millions of files every day.

Two of the largest U.S. banks chargeback, integrated with FileCensus; one has used chargeback for six years and the other for more than three years. Both "Global Financial Service Providers" use chargeback very dynamically, scanning over one billion files every day to generate chargeback reports.

Data Management for Compliance and Security

Meet Government Regulations Effectively

Public companies, as well as organizations within the legal and medical fields, benefit from the file organization and storage capabilities of FileCensus. Corporate and medical governance regulations enacted within the past several years required these types of enterprises to maintain large amounts of documents. The most efficient way to do that is with electronic file storage. However, finding those files when called upon to produce them can be problematic.

Maintain long-term files with short-term pain.

For corporate compliance purposes, these organizations need to organize and manage their data and store it on their networks forever. Medical professions must comply with the federal Health Insurance Portability and Accountability Act (HIPAA). Enacted by the U.S. Congress in 1996 and effective in 2003, this is a set of national standards that deal with the accessibility of patients' medical records. The act provides for the protection of certain health information, while at the same time allows the flow of information necessary to provide proper health care. The act requires medical entities to keep records of their patients for at least 5 years after their death.

In the legal and corporate world, public companies must comply with the Sarbanes-Oxley Act of 2002. Also known as the Public Company Accounting Reform and Investor Protection Act, this is a federal law enacted in response to a number of major corporate and accounting scandals, which involved financial disclosure issues with public companies. The act requires public corporations to maintain corporate and financial documents, and to make them available for audit purposes. Managing the data on a network is important to them.

Move data to a safe, cost-effective location.

With the need to maintain large volumes of data for extended periods of time, the security surrounding how that data is stored is vital. Security is an important part of the meta data that goes along with a company's files. System administrators can access files to learn who owns them and who can update them. Security becomes more important as organizations want to monitor their data. At that point, they need a tool to help them find out what they have. Accessing that information is the big payoff.

Manage logins and user groups.

The security component is specifically for administrators. Implementing security into a large environment is quite complex. Users have individual logins, but individual users also belong to groups. Groups have common areas they work in, but members of other groups do not have access to that group's information. Administrators managing networks with 50,000 users and 20,000 different groups need help controlling all the information generated by those users. The networks they oversee contain millions of directories with billions of files.

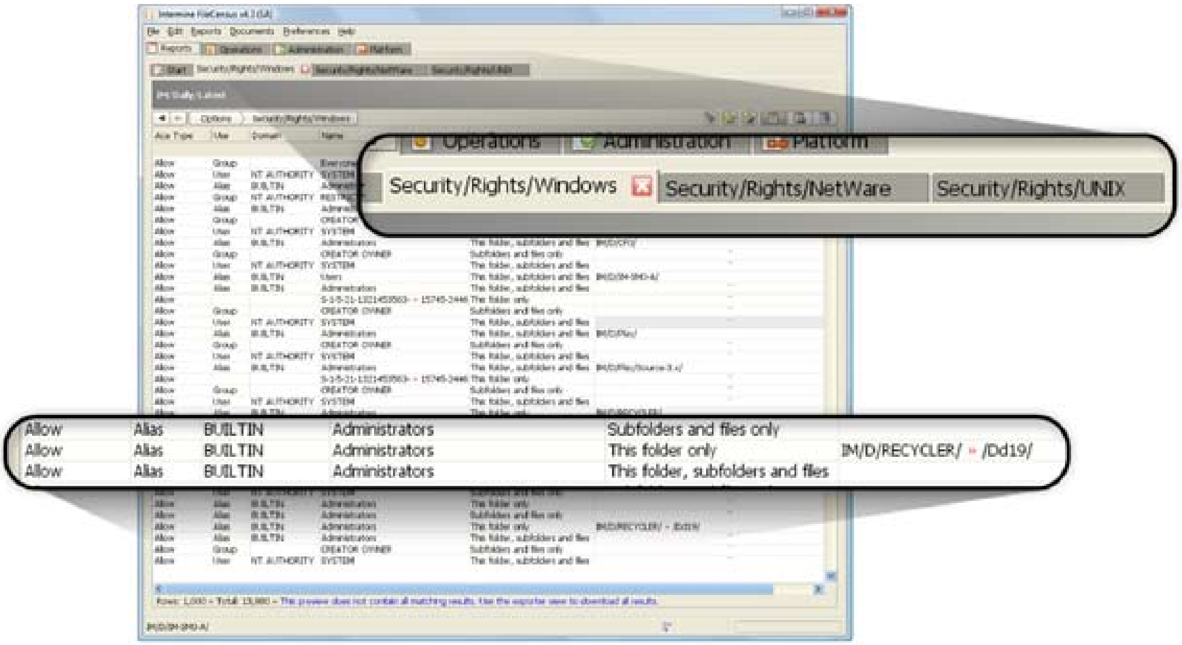
Security information is quite volatile. Users create new data and change existing data on a daily basis. Just having information in external repositories leads to questions that need to be answered, such as what occurred in the system a week or a month ago.

In addition, security-related data is quite difficult to collect, particularly in a large scale environment. The collection process is slow. If you have a tool that is trying to collect all the information centrally it can negatively affect the architecture of your data collection. At that point, it becomes more of a burden for an outside tool to collect.

Secure your network.

The key to the security process comes back to having an understanding of how the network works. FileCensus creates reports that allow administrators to clean and tighten up security and to take away "rights" in a certain area. There are no other tools out there that can provide a centralized view of security meta data.

FileCensus collects this information as part of its day-to-day operations and brings it back to one central location. The information has already been collected with the meta data, and with that, can then answer questions about the data kept within an organization's network. It's building along the theme of the FileCensus solution in that it provides security information from a day, week or month in time, and does some in a timely manner, as the information is gleaned from information that has already been gathered.



Allow	Alias	BUILTIN	Administrators	Subfolders and files only	
Allow	Alias	BUILTIN	Administrators	This folder only	IM/D/RECYCLER/ » /Dd19/
Allow	Alias	BUILTIN	Administrators	This folder, subfolders and files	

Figure 4
Information Security

With security, it's also important to have a physical baseline to know that an unadulterated version of the company's storage exists and that it can be referred to as needed. Some backup software products have these capabilities, but they cannot handle a large database. It's often hard to use the information that's been collected during a backup to determine what security is out there. It could take days or weeks to gather the information, based on what questions are asked.

The few existing security tools are expensive and specialized. The BindView Control Compliance Suite, now the Symantec™ Control Compliance Suite automates the management of deviations from security configuration standards.

Organizations could try to set up something like this on a proprietary basis, using specific scripting languages. It's easy to collect general information but security information is quite tricky to collect, particularly in a large environment. Just trying to do anything from a central location in a big environment is problematic. A system administrator might be able to create scripts at the server level. However, reaching beyond that to the file level, with billions of files out there, causes a completely different set of problems.

For large organizations, the difficulty in gathering this information is compounded by the number of security policies implemented. With FileCensus, administrators can ask to see all the permissions on the directory, enabling them to deal with all policies as one.

Finally, there are some products designed to be used on single machines, but not on large enterprises. The system administrator would have to go from machine to machine within an organization, taking valuable time and effort.

Archival

Prioritize data and store in tiers.

Creating tiers of storage is a cost-effective and efficient way to manage data accessed on a regular basis, as well as data that need long-term maintenance, in keeping with certain compliance regulations. FileCensus assigns different categories of data different types of storage to reduce the cost. That means that data is all from the same category and the different tier is a second, less expensive, tier of storage. The recommendation is to only have data on the more expensive primary level of storage that is accessed more frequently. The secondary tier has storage for information that is accessed less frequently. Information that needs to be kept forever, such as that necessary for HIPAA or Sarbanes-Oxley, can be stored on this secondary level of storage.

The top 20% of users take up 80% of all storage

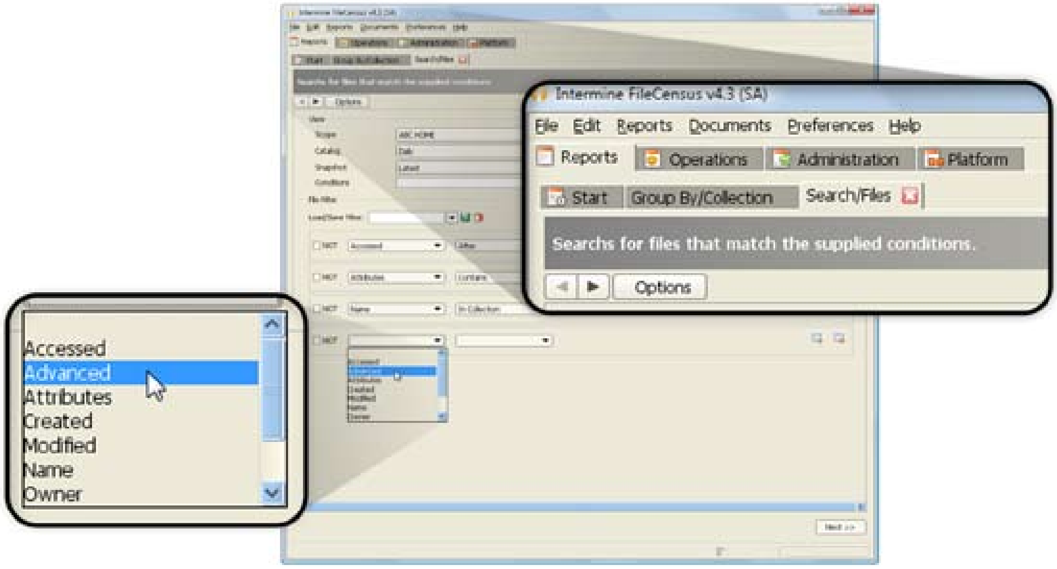
Which data is most important?

How can a network administrator create and manage more storage on a smaller budget? That question resonates through companies large and small. Tier 1 storage is expensive and storage requirements are continually growing. The top 20% of users take up 80% of all storage, so administrators don't want them all using high-cost, high-performance Tier 1 levels of storage, particularly given that 90% of all files generated are not re-accessed after 30 days.

90% of all files generated are not re-accessed after 30 days

Typically end users are not aware of storage problems. They think it is their right to use as much storage space as possible. The end user doesn't comprehend what their usage does to others in the company when it comes to taking up storage space on the network.

Data ownership is another issue involving end users. They don't want to see someone else moving their files, but they also don't want to do what is necessary to manage their files in a way that decreases their storage space on the server.



Searchs for files that match the supplied conditions.

Figure 5

File Archival

The challenge is how this is implemented. It's hard for administrators to understand individual data, so it is difficult for them to understand its importance. They don't' know which data needs to be accessed frequently and which data just needs to be stored? IT's definition of what's archival may not be the same as the end-user's definition of what needs to be archived.

Establish behavior-based policies.

Tiered storage assigns different categories of data to different types of storage to reduce cost. FileArchive creates behavior-based policies based on what is actually happening in the network. Primary, tier 1 storage is reserved for mission-critical documents used on a daily basis. Secondary storage is reserved for business data which is important but not accessed very frequently. A long-term level of storage is held for fixed content, tape libraries and deep archives, such as those required for organizations to meet the requirements of the Sarbanes-Oxley Act or HIPAA.

FileArchive gives the end user control to say what needs to be archived and what needs be accessed on a regular basis. What is deemed unimportant to the IT administrator might be a critical document to the end user. By reviewing behavior, the end user has a greater say in the fate of his or her documents in the storage level hierarchy. The end effect is a much cleaner archival process and IT doesn't have to manage it.