



Why understanding your data must be the starting point for intelligent storage management

This paper discusses why understanding and automating the management of data and storage should reduce problems, downtime and “fire-fighting” activities for IT administrators.

It also considers why accurate decisions and actions around storage management issues such as Backup and recovery, High availability, Storage consolidation, Server, consolidation, NAS, SAN, Archiving etc cannot be taken without understanding the content of the data that is being held.

A white paper from otium software ltd.

Author: Andrew martin

Date: 1/1/2004

Contents

Background	page 2
What you should know about your data	page 3
How you should benefit from knowing your data	page 8
Pre-requisites when choosing a storage management software solution	page 11
Conclusion	page 12

Background

In IT today one factor is a given, as long as a company or organisation remains in business, the amount of data that they store and manage will grow. This is true of traditional forms of data held on paper, but even more so of electronic data.

The amount of electronic information that even small companies store and manage would not have been dreamt of even a decade ago. E-mail, presentations, multi media files, scanned documents and application data all take up ever increasing amounts of disk space in IT environments.

We never know what data we will need to look back on so we tend to store rather than delete our data. Every piece of information that is stored requires, time effort and expense to manage.

Every bit of data needs to be stored on a hard drive. Hard drives, raid controllers, disk management software, server room space all cost money.

Every bit of data we store will be backed up. Tape devices, tape media, backup software, backup time all cost effort and money

Decisions are consistently being made regarding storage infrastructure. DAS, SAN or NAS. What level of High availability should be put in place, what plans for data protection should be implemented? What type of disk technology should be employed, the list goes on.

This paper considers the viewpoint that until the IT manager is more informed about the actual content and context of the data itself, it can be misguided to make decisions around the infrastructure to accommodate this data.

What you should know about your data

Legality

An unfortunate consequence of the electronic age is that data is at best offensive and at worst illegal has become very easy to get hold of. Companies can be held liable for the kind of data that employees store on company equipment. It is important for companies to actively seek out unsuitable data such as pornographic material and remove it from company equipment.

The problem is serious. It is not enough to hope that employees will use correct judgement in deciding whom they share such data with. If the offensive data is stored on company IT equipment in an area where other users have access. It is possible that employees may come across and be offended by such data. In such scenarios companies can and have been found liable to prosecution.

Duplication

In today's IT environments data is freely, quickly and easily distributed. Users do not view electronic data as tangible; they do not associate cost and effort with storing electronic information. As a result users tend not to look for efficiencies in what electronic data they are holding in the same way they may have done with personal paper based filing systems.

This has resulted in the identical data being stored by numerous users. It can also result in the same user storing multiple copies of an identical file in numerous locations. An example of this is when a particular user downloads a useful file from the internet. A few months later that same user requires use of the same file. They are pretty sure that they have downloaded it before, but can't remember if or where they saved it. After two or three minutes of unsuccessfully searching for the file they simply download it again and save it in another place on the network.

Likewise individuals send out e-mails with useful files attached. Users often detach and save their own versions of these files. This can lead to the same file being stored by many users.

Quite simply duplication wastes time, effort and money for IT departments. Some IT administrators try to keep on top of this type of issue by periodically running manual searches. However the time they spend on such manual time intensive exercises does not usually justify the means. In order to keep on top of the build up of duplicated data automated tools are the answer.

Stale Data

Most human beings share a trait. That trait is that we like to hoard things. It is in our nature. As time goes by most of us find that our houses become full of “stuff” that we no longer require. When things become too much we have a clear out. It transpires that the electronic age has not changed this characteristic, users hoard data. We store data but very rarely delete it again, even when we no longer require it. For IT administrators running IT for 10’s 100’s or even 1000s of users the chance to have a periodic “clear out” is essentially non-existent.

IT administrators are usually too busy to trawl through listings of files attempting to guess what files may not be required by users. The users themselves tend to be too busy to go through the same exercise.

It is an unfortunate fact that once quantities of stale data start to build, there is very little that IT departments can do to fight it. Automated tools alone will not be enough. The problem with identifying truly stale data is that automation alone cannot provide the answer. Ultimately the owner of the data needs to make a decision regarding a particular file as to whether it is required or redundant.

Therefore in order to effectively identify and act on potentially stale data, IT managers need simple automated tools that do the “grunt” work in identifying candidate stale data. They need to combine this with an inclusive policy that provides a method for allowing users to very quickly feedback to IT whether files are indeed required or not. As an example an automated tool could be used to find potentially stale files (e.g. those that have not been accessed for 6 months). The same tool could be used to generate an automated e-mail to each owner of these identified files, letting them know that the identified files will be deleted in 30 days unless the user e-mails IT to request particular files not be deleted.

Useless data

As we create data, along the way the applications and operating systems we employ have been designed to create temporary versions of our files, this can be useful when users forget to save their work or systems shutdown before users have had a chance to save their data. Other applications need to create temporary files in order to complete calculations and executions of commands.

Over time large amounts of disk space can be taken up by what is essentially useless data. It serves no ongoing purpose, is not required by any application or needed by any users. Yet IT departments faithfully look after this useless data in the same way that they do even the most business critical files they hold.

If IT managers are unable to easily locate and identify this type of data, then the chances are it is residing along side other data taking up space on expensive disk arrays, perhaps using SAN infrastructure, being backed up every night, having high availability technologies (such as clustering or replication) applied to it to ensure that it always remains on-line and available.

It is vital that this kind of data is identified and purged at appropriate times. This data is not just useless, it is damaging it has a negative impact on IT systems, efficiencies and performance.

Multimedia Files

Audio files, video files and picture files tend to be disproportionately large from normal data files. Many businesses have a legitimate need to have these kind of file formats on the corporate network.

However because this type of data is so space intensive it makes good “housekeeping” sense to keep a vigil on multimedia file growth in your IT environment. Even in environments where certain types of these file are necessary (e.g. design companies), the danger is that a large percentage of these files will not serve any business purpose.

We all receive humorous multimedia attachments on our e-mails. Most companies are happy to foster reasonable and vibrant e-mail cultures. However when individuals receive a mail of this kind that they like it is not unusual for them to save a copy for themselves. This quickly accumulates. It is not unusual for companies with data growth problems to find that massive reduction in data consumption can be achieved by identifying multi media files amongst there users.

The tools exist to identify and eliminate this type of data, however in most situations such draconian action is not the answer. Once again a combination of automated tools combined with sensible policies usually provides the answer.

Companies may decide that they do not want to deprive users the right of storing this kind of data, it is often good for moral. One answer is to use extremely cheap jbod type disk as an archive area for this type of data.

The correct tools can be used to identify this data, give users a chance to verify that the data identified is not business critical, and then migrate such data to the provisioned archive area.

White Space

Imagine this. You have ten servers in your environment. Each has a 1TB hard drive. Nine of these servers are only using 100GB of the 1TB available. One of the ten servers has reached 999GB capacity. The IT administrator has to go out and buy extra capacity for that one server, despite the fact he has over 8TB of unused disk space sitting amongst his other servers.

This extreme and unrealistic example demonstrates in clear terms the waste that is caused by the amount of white space that exists wherever servers are installed.

Unfortunately the nature of servers, disks and file-systems is such that identifying this white space is not easy to do, and can be time consuming. Automated storage management solutions can very quickly provide insight into where this white space may be. Once IT administrators have this information to hand they can work out how to best maximise the utilisation of their existing disk infrastructure.

Trending

A massive issue that data growth causes for IT departments is its unpredictability. If we don't have a fairly clear idea of where data is likely to increase in volume and what rate it becomes very difficult to plan for data growth.

Bad forward planning around data growth can lead to a host of issues:

- Inaccurate budget forecasting
- Inefficient use of existing resources.
- Reactive (and hence costly) purchases of new storage and new servers.
- Unplanned and unscheduled downtime for maintenance
- Increased downtime for maintenance and problem resolution.

By using tools that monitor disk usage and data growth, the opportunity to start doing basic trend analysis lends a greater degree of accuracy to planning for managing this data into the future.

Profiling

Different users and different departments have different usage patterns of company IT. The amount of data they create and maintain can also vary enormously.

In providing the right resource for the right departments and individuals knowledge is king. The ultimate aim of any IT department is to keep the rest of their company productive. Providing the right resources at the right time is key to achieving this.

It is not uncommon even in very forward thinking IT departments to find that disk space tends to shared equally across departments and individuals. However it is also very common to find that this tends to lead to problems. With some users and departments running out of disk space on a regular basis whilst others never get near 100% utilisation.

Using automated tools can allow IT departments to have a much clearer idea of which departments and individuals are creating more data than others.

Capacity Monitoring

Almost every IT department wants to avoid unplanned downtime and generally like to minimise planned downtime also.

One cause of unplanned downtime is having to perform reactive maintenance. Dealing with volumes and disks unexpectedly running out of space and capacity is one such cause for having to take on unplanned maintenance.

As companies increase the number of servers they use and also increase the number of disk volumes they set up on those servers, the task of monitoring usage levels on each of those volumes becomes increasingly time intensive. Ultimately time and resource constraints will result in servers going down prior to administrators discovering that disk space available has been reached.

Simple centralised alerting solutions that keep a consistent watch of disk and volume usage levels are available that can provide advance warning to administrators of impending capacity based difficulties. Such tools allow administrators to gain advance knowledge of problems and plan to conduct appropriate maintenance at least disruptive times.

Performance Monitoring

Businesses often acquire highly specified server technology in order to ensure that their business applications perform quickly. This is key to many business environments, as it is important that application performance enhances rather than slows down employee productivity. It is also fairly standard for companies to use server based monitoring tools to check that each server in their environment is providing required performance levels.

Application performance is not purely a function of servers. There are a number of other factors that contribute. The most notable are networking and storage performance. From a storage perspective there can be numerous reasons why a disk performance may start to diminish.

The most important factor in these instances is to identify early that a disk is starting to under perform. From that point the administrator can investigate, understand the issue, work out the corrective action and plan the appropriate maintenance before the performance problem begins to impact user productivity.

How you should benefit from knowing your data

IT environments change on a near daily basis. Small reactive changes are made and implemented often. More detailed changes requiring specific planning, project management and dedicated IT purchases are part of every IT department's evolution. At the heart of all IT is data. Without the data the technology would be pointless. It would seem obvious that this fact alone would govern that know decisions could be taken without first having a firm grasp of all the factors relating to data already discussed in this paper.

However, this is usually not the case. Companies make large IT decisions, spend significant budget buying equipment and implementing solutions often without having spent due consideration assessing the data that these IT solutions are designed to create, deliver and protect.

Accurate knowledge of your data will facilitate more appropriate and often more cost effective solutions being implemented than in circumstances where no real grasp of the nature of data is known. Some of the most significant areas are highlighted below.

Hardware Purchases

If a company remains in business it will continually create more data. As more data is created, servers will run out of disk space. The solution for the vast majority of IT managers in this situation is to purchase more disk or another server.

The IT manager utilising automated storage management products will be able to assess the nature of the data on these servers. This will enable IT administrators to reclaim disk space and reduce the amount of new server and disk hardware that needs to be purchased.

Ongoing, good storage management policies will halt the rate of data growth and reduce the amount that needs to be spent on server and disk hardware into the future.

Storage decisions

DAS, NAS, or SAN. IT administrators often have to consider what type of disk storage they want to implement as their data volumes grow. Working out the right type of technology to move to is not always easy. It can be made more difficult by vendors with a vested interest. If you are speaking to a vendor that specialises in NAS based technology, the chances are they will attempt to convince you that NAS is the right technology for your situation.

The truth is that different types of disk storage are suited to varying types of data. Most companies have many different types of data (e.g. databases tend to work at block level, where as file and print tends to be via a file-system).

When an IT administrator has a strong grasp of the actual content of their data, they can start to work out that a number of different storage technologies may be appropriate and should perhaps co-exist. By knowing your data you are able to make informed decisions regarding what type of storage should be used for different types of data. This is not restricted to fundamental differences such as NAS and SAN, but also to the quality of disk that is purchased. If an IT administrator is able to identify non-critical data from business critical data. They can then purchase expensive high performing disk for the business data and cheap job type disk for the non-critical data.

Server or Storage consolidation

The two main reasons for looking at storage or server consolidation tend to be the pursuit of infrastructure efficiencies and management efficiencies.

The nature of any consolidation project is hardware led. It is often the case that IT managers will turn to hardware vendors to lead these projects. In many ways this is a sensible decision with one caveat. Hardware vendor's start from the perspective of what hardware are we consolidating. In order to maximise efficiencies the starting point needs to be "What data are we consolidating?"

In practical terms the difference comes down to the following base lines.

The hardware centric approach:

How much disk capacity do you have?

Add 20% for growth and that's how much consolidated disk you should purchase.

The data centric approach:

How much disk do you have?

How much of that disk is not used?

How much of the data on that disk can be removed?

How much of that data is of a nature that it could be stored on old re-purposed centralised disk?

You can actually purchase 20% less capacity than you currently have across your enterprise.

Good storage management practice actually allows you to reap the promised benefits of projects such as consolidation.

High availability solutions

As data is viewed as more critical to business success and continuity, companies start to be more serious in their consideration of implementing high availability solutions such as replication or clustering.

Implementing these solutions is not cheap and the more that requires high availability protection the more expensive the solution and it's ongoing management costs will be.

The curve is exponential. As the level of high availability protection increases (e.g. 99.999 availability) the cost of implementing and running that level of protection also increases accordingly.

IT departments that have advanced storage management policies and tools are able to get very granular in working out what data actually requires high availability protection and which data may be more appropriately protected by more cost effective means such as tape backup.

Having the ability to analyse data in this way makes it possible for companies with even small IT budgets to consider implementing intelligently focused cost appropriate high availability technology for their most business critical data.

Backup and Recovery

Backup procedures are a part of every day life in every IT department. Generally backup is regarded as a necessary evil, it has to be done but it causes big administration overheads, and as the amount of data being backed up increases the intrusiveness of the backup process (e.g. longer backup windows) grows.

Also as IT environments become more complex the chances of some files simply not being backed up also increases.

Numerous answers exist for upgrading backup solutions in order to increase backup rates and decrease backup windows. However, the price to move to enterprise level backup solutions is often a barrier even for the largest of companies.

Using automated storage management tools, can allow administrators to use in depth knowledge of their data combined with the existing backup infrastructure in order to develop new backup policies that reduce the amount of data being backed up at any one time.

Using the knowledge gained to reclaim disk space will also reduce the total amount of data that is being backed up on a daily basis.

Pre-requisites when choosing a storage management software solution

Sensible storage management should not be a nice to have, it should be a pre-requisite for any IT department that wants to maximise their own staff time and productivity, ensure that efficiencies and budgets are preserved, and provide excellent service, application performance for their users, and generally maximise company productivity and results.

A storage management product should be simple to install and be very sparing in the amount of system resource it requires to run.

Storage management products should be focussed on storage alone, if they start to venture into other areas they enter the realms of framework systems management solutions. These have their place but are a different proposition in terms of functionality and resource required to derive benefit from them.

The price of a storage management solution should be appropriate to the benefit it delivers. A number of storage management software providers have attempted to price their solutions very high and make sales and justification based on ROI arguments. The only reason for buying into any storage management solution should be for the functional benefits it delivers. Once installed, any Return on Investment that is delivered should be viewed as an extra benefit.

When deciding on a storage management software solution IT managers should look for flexibility in licensing. It may be that not all servers need to be monitored all of the time, however it is likely that at some point every server will need some level of storage management investigation. Some products and companies do not make redeployment of licences easy to do. Others have designed their products to cater for movement of licenses between servers on a need to use basis. This type of flexibility can be very useful.

Storage management is a massive field. Two products that are described as storage management tools can provide totally different functionality (e.g. data archiving and data interrogation tools can both be classified as storage managers. Be sure that you pick a product that deals with your issues. Some products will deal with your issues and a whole load more areas that are not problems in your environment. It is important to be sure that you do not end up paying for a host of functionality that you do not need. At the same time, check that if you suspect that other issues may arise in the future, look for a technology that allows you to build and add in features and functions as and when you need them.

One of the main reasons for investing in storage and data management technology is to reduce the workload on IT staff, freeing up their time to concentrate on more proactive business beneficial activities. When evaluating any storage management automation software, be sure to check that the administration effort to run and manage the software itself is minimal. Some products may deliver great information but at the expense of being cumbersome and time intensive to use and manage. To a large extent storage management software with a high management overhead is self-defeating.

Conclusion

Many people feel that paying a service provider to conduct a one off storage audit is the same as investing in storage management tools.

A storage audit is a useful exercise, but limited in as much as it can only provide you with a snap shot view of the point in time that the audit occurred.

Storage and data is constantly growing and evolving, if you allow it, data will grow out of control. Skilled IT administrators can keep on top of storage issues most of the time, but at the expense of time effort and budget.

Using automated storage management tools, ensures that IT administrators can remain in control of data **all** of the time, with significantly reduced effort and time, whilst reaping the benefits of greater efficiencies and budget savings.