

Comment Article

Information Management -

By Clive Longbottom, Service Director, Quocirca Ltd

I recently heard a worrying statistic : It is predicted that by 2012, the amount of data being stored will double every 11 hours. I have to say that I view this with a healthy dose of scepticism, but whichever way you look at it, there's still going to be a lot of data around.

Standard approaches of trying to deal with rapidly expanding data volumes have been met with relatively low levels of success – trying to get people to stop copying files to multiple different storage devices, sending files to large distribution lists via email and getting people to only use centralized storage just does not match with people's own ways of working, so we do tend to end up with large duplicate amounts of data being stored.

Data deduplication has been around for a while, but it still has to reach a reasonable degree of use in the mainstream. Even where it is in use, it is often misunderstood and poorly implemented.

So, what can we do with data deduplication? At the most basic level, we can remove all the copies of the files that we have in the system. Here, files are compared against each other, and where there is an exact match, the files are collapsed down to a single physical file, with use of virtual pointers to point to this file from other places. This can save the average company somewhere between five to 15 percent of its storage volumes.

This is, in reality, only scratching the surface. Let's take a much deeper look at how an organization deals with its data and identify the sort of savings that are possible.

In creating a document, a user will get it to a level where it becomes a working draft. At this point, it needs to be sent out for review by others. Although workflow and document management systems are well-suited to this, very few documents at initial draft level will get

into such systems, due to perceptions of high cost and overly formalized work practices.

Instead, the majority of users will tend to use email to send the document to a group of possible reviewers. For the sake of argument, let's assume that there are four people in this group and that each participates in the review. Unless the review process is tightly controlled, each person will tend to review in isolation, saving a copy of the document to his/her drive and working on it. After the review is complete each will then send the document back to the original owner, who will then aggregate comments.

So, we start with one document. This document is copied as a single email attachment and depending on how the organization's email system works, it will then become four additional documents – one in each recipient's inbox. There will be four more copies when each person copies the attachment to his/her own file system as well as four modified files being sent back to the original owner, leaving four more copies in the inbox. The original owner will then make a copy of each of these and will make a new overall document reflecting the comments being made.

In one review, there could be 23 similar or identical versions of the one document. If we have three iterations of the review, we end up with 70 versions of the document being stored in different places around the organization. If we assume a file size of 200KB, we suddenly have 14MB of data being stored.

But, if block-level data deduplication is used, the content of files is compared at a more granular level, comparing blocks of data against each other. Therefore, if there are two documents where only some text has been changed, only these changes will be physically stored, along with a small amount of data that shows how the actual document itself needs to be rebuilt.

Block-level deduplication can also work against other types of data – image, encrypted, video and voice. An organization can expect to save around 60 to 80 percent of its storage volumes by taking such an approach.

Sound like a silver bullet? In many ways it is, but I also recommend a degree of caution. Unless business continuity is taken in to account, data deduplication can lead to massive problems. Let's assume that there is disk failure or data corruption somewhere in the system. Many parts of the storage will be made up of data fragments and pointers, and it will be almost impossible to rebuild these so that data can be successfully rescued. You may have been thanked for cutting down on capital and operating costs for storage while it was all going well – now you're suddenly the villain of the piece for not foreseeing what could happen.

But, if you are saving 70 percent of your storage volumes, mirroring the data still means a saving of 40 percent - and you will have created a solid data business continuity capability at the same time. Data deduplication is not all smoke and mirrors, it really does work, and newer approaches apply the approach across all data stores. Indeed, companies capture the data before it hits the storage itself, deduplicating it on the fly while enabling intelligent filing. Here, the data can be tagged as it is deduplicated, and the system can then make sure that it is stored in the right place, whether for governance and audit reasons or for cost reasons, using old storage as near-line storage for less important data. Further, as the data is tagged automatically and is indexed against text contained in documents, you get a fully searchable, organization-wide data store.

If organizations do continue to see storage growth and if the statement at the beginning of this article is anywhere near the reality, then organizations have to do something. You can try educating the users to be less profligate in their storage use, but it's probably a lot smarter to take control through automated means and finally grab the problem by the horns.

About Quocirca

Quocirca is a primary research and analysis company specialising in the business impact of information technology and communications (ITC). With world-wide, native language reach, Quocirca provides in-depth insights into the views of buyers and influencers in large, mid-sized and small organisations. Its analyst team is made up of real-world practitioners with first hand experience of ITC delivery who continuously research and track the industry and its real usage in the markets.

Through researching perceptions, Quocirca uncovers the real hurdles to technology adoption – the personal and political aspects of an organisation’s environment and the pressures of the need for demonstrable business value in any implementation. This capability to uncover and report back on the end-user perceptions in the market enables Quocirca to advise on the realities of technology adoption, not the promises.

Quocirca research is always pragmatic, business orientated and conducted in the context of the bigger picture. ITC has the ability to transform businesses and the processes that drive them, but often fails to do so. Quocirca’s mission is to help organisations improve their success rate in process enablement through better levels of understanding and the adoption of the correct technologies at the correct time.

Quocirca has a pro-active primary research programme, regularly surveying users, purchasers and resellers of ITC products and services on emerging, evolving and maturing technologies. Over time, Quocirca has built a picture of long term investment trends, providing invaluable information for the whole of the ITC community.

Quocirca works with global and local providers of ITC products and services to help them deliver on the promise that ITC holds for business. Quocirca’s clients include Oracle, Microsoft, IBM, O2, T-Mobile, HP, Xerox, EMC, Symantec and Cisco, along with other large and medium sized vendors, service providers and more specialist firms.

Details of Quocirca’s work and the services it offers can be found at
<http://www.quocirca.com>