

Open Comment – Master Data management and Deduplication

By Clive Longbottom, Service Director, Quocirca Ltd

Although the cost of storage devices has plummeted over the last few years, the cost of managing stored data continues to grow. Unfortunately, an organization's predilection for creating new data does not abate, and many see data volumes doubling every year to 18 months.

Not only is the cost of managing all this data becoming a major issue, but the speed of reporting against massive data sets is causing headaches too. It is not just about size, but that much of the data is being stored in silos under monolithic applications, which limits how effectively organizations can respond in dynamic business markets.

A further issue revolves around the changing landscape of legal and additional regulatory controls that often require more information to be kept for longer periods of time.

It is important to look at what data is being created, how it is being created and stored and what steps can be taken to provide a more streamlined and manageable environment. There are two major ways to approach the problem: master data management (MDM) and data deduplication.

MDM is a great step toward gaining control over data. The idea behind MDM is to take monolithic applications and ensure that the main information is codified in a standard manner, enabling searches, reporting and business

intelligence (BI) to be carried out across multiple data sets at the same time. This involves taking referential data and creating a separate database to hold the information.

A prime example where MDM comes into its own is with customer data. An organization's customer relationship management (CRM) system will hold a customer's name, address and other contact details. This data may also be needed by the enterprise resource planning system, if it manages delivery and supply chain issues. More often than not, the two sets of data will be separate from each other, in data fields with different names and may even have different details due to errors or differences in the ways that individuals put information into systems. For example, my details may be in the CRM system as Mr. Clive Longbottom, 1 High St., Reading, RG4 7HS. In the ERP system, it could be C.S. Longbottom, 1 High Street, Reading, Berks. To the human eye, it's pretty obvious that these two items are the same. To a computer, they look completely different.

MDM aims to create a single reference record for inconsistent data, so I will be known by the same main contact details on all systems. When any system wants to find information about me, it goes to an MDM master data set and picks up my name and contact details.

MDM is not about creating the world's largest data warehouses; applications still retain their

own data sets covering the data that they need. However, the shared information is held in a separate data set, and because this set should be a lot smaller than any of the application data sets themselves, the initial response speeds should be faster.

MDM does not impact data volumes in any major way however - and for this a different approach is needed. Much of the data held within an organization is heavily redundant. For example, most email systems hold physical copies of each message. Therefore, if a 1MB document is sent to 10 people, 10MB of data storage is needed. If 50 percent of recipients save a local copy, an additional 5MB of storage is needed. If minor changes are made to the document and sent back by two people to all recipients, another 20MB of storage is required. If the organization has basic backup procedures in place, then each of the documents will be duplicated to backup disks and tapes. Whether the documents are 98 percent alike or greater is neither here nor there as far as storage systems are concerned - each document is complete in itself and will be held as such on disk.

Let's look at basic deduplication approaches. If a means of identifying when identical documents are being stored can be put in place, a virtual pointer to a single copy of the document can be created, saving the overhead of storing multiple copies. How about saving the changes when they are made, rather than the whole changed document? Such approaches can drastically reduce storage requirements. Email vaulting solutions from vendors such as Symantec and CA can really help with managing the main

cause of this - the overuse of email as a document workflow and review system.

However, deduplication can be taken further. The majority of storage management vendors, such as Symantec, IBM and EMC, now provide capabilities to look at data at a binary code level and identify where blocks of data are identical. These blocks can then be stored as single master records, and only where changes are noticed are they stored - but they are stored as delta changes, rather than new full data records.

Such an approach can collapse data storage needs by 60 percent or more, and this can be magnified when you look at backup storage requirements. After all, when you back up your system as a complete image, it is unlikely that more than 10 percent of that will have changed when you next back up. If you have applied deduplication techniques to the original data as well, then everything becomes far more compact. Even with the overhead of rebuilding data sets from the initial master and applying the changes, response times are improved, due to the much smaller data sets involved.

Bringing together MDM and deduplication gives organizations just what is needed in today's markets - a far more responsive and manageable data environment for supporting the business.

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, Dell, T-Mobile, Vodafone, 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>