

DMReview – The Problems of Megadata Searches

By Clive Longbottom, Service Director, Quocirca Ltd

Imagine that you are newly arrived on planet Earth and faced with going to a public library. You are completely unaware of how a library works, but you understand books themselves. You enter and start to look at every single book until you find one on your chosen subject - possibly something along the lines of *Social Etiquette for Earth Visitors*. Having found this book, you go away, read it and return the next day to find another book on the same subject. You start again from the very beginning, looking at every book (including all the ones you looked at yesterday and found no interest in), until you find another book on the same subject. You patiently continue to do this, day after day, until someone offers to help you.

Why am I harping on about such a case? Well, it seems to me that the above method is the main means the majority of companies utilise in dealing with their own data. When a report is required, it is run against the complete database, often day after day, replicating the same searches it has done many times before.

In the library, the person who will help will be the librarian - a person who has built up a wealth of knowledge and multiple means at their disposal to help identify and locate what you need. In the data center, it has tended to be the use of relational databases, with multiple indices and fast and expensive hardware platforms.

Increasingly, however, massive databases on expensive hardware don't seem to be enough. Running reports against such megadata stores is still slow, with some reports taking many hours or even days to complete. Set against this is the need for organisations to be far more fleet of

foot, responding to changes in the market at near real time.

Unfortunately, the continuing growth in data quantity, often combined with a lowering of data quality, does not fit well with the need for speedy reporting.

Back to the librarian. If a library took the same approach as a historical database, all new books would just be put on any shelf and a member of the public would have to search through each book until they found the one they wanted, much like our alien earlier. Luckily, librarians have spent time in coming up with what appears to be a simple means of enabling books to be identified far more rapidly.

Let's take this at its most basic level - a librarian using a paper-based system. The librarian has a set of cards, each of which covers a certain aspect of indexing books. When a new book comes in, the librarian takes specific information from the book, and adds this to each card as needed. For example, the author is added to one card, the subject of the book to another, where the book is physically located to another. Many of these cards will refer to other cards, so that the librarian can easily move from one item to another as required. When a member of the public comes looking for the book, the librarian can easily retrieve information on the book, no matter what the information is that the member of the public provides. If the person wants more books by the same author, it will be under the *author* card, if more books on the same subject, look it up on the *subject* card. The librarian hasn't needed to refer to the overall database of

all available books. Instead, they work against very small, dedicated subsets of information.

This is the basis behind the use of standard indices, but the one thing that seems to be missed by many of the database vendors is the one key fact. Each new piece of data is dealt with as it enters the system, not once it is already there. The incoming book details are dealt with immediately and each new book is just that - one single item, rather than an increment in the overall massive database. For many organisations, a 10 million record database will only have a few hundred main equivalents to the librarian's cards, which means that in-line, real-time reporting becomes possible. Each new piece of incoming data can be dealt with in a very rapid manner, with the pertinent information being added to the relevant cards before the main record is created in the master database.

So, let's look at how this would work in the online world with the example of a customer on a Web site creating an order for an item. What information do we already have on them? If it's held in a standard database, we run a standard search against the existing information generally using an indexed field and pull up their existing record. However, if the customer provides information that is held as a non-indexed field, the search becomes massively slower. But, if we use this in-line mode, all we have to do is to go to the equivalent of the *customer* library card and see if they are already there. If not, we add them as a new record. If they are, we see which other cards this card points to for other information, such as previous purchases, payment records, possible upsell items and so on. What provides the main speed here is that these individual virtual cards are permanent records - but are far smaller than the underlying

database itself, by several orders of magnitude. Therefore, searching through these is almost instantaneous and incoming queries can be dealt with immediately and effectively.

A further advantage of this approach is in the information that can be uncovered. For this, let's look at a different example - fraud prevention in the financial markets. A person wants to take out a loan and applies online. They know they have a bad record and, so, use an alternative name. Normally, a check would be run against the name of the person applying and, if that came through clear, the loan may be approved. Now, first the name is searched for, which comes through clear. The address is also run. As the list of information is run against a small subset of the main database, more than just a straightforward one-for-one comparison can be done. For example, is 123 Main St. the same as 123 Main Street or even 123 Mian St.? Intelligence can be applied to the comparisons and uncover far more fraudulent attempts in the process.

A prime player in this market is IBM, which bought Systems Research and Development (SRD) in 2005, gaining an evangelical resource in the form of Jeff Jonas, the founder of the company. Jonas has been applying this approach for fraud identification in Las Vegas (amongst other places), where real-time identification is necessary to prevent a fraudulent gambler from rapidly fleeing a casino.

When used in conjunction with other data optimisation technologies (e.g. information deduping, transactional data management approaches, in-memory data caches and so forth), in-line data analysis can reap benefits for organisations looking to, not only, be more responsive to incoming data events, but also ensure that fraud is minimised.

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>