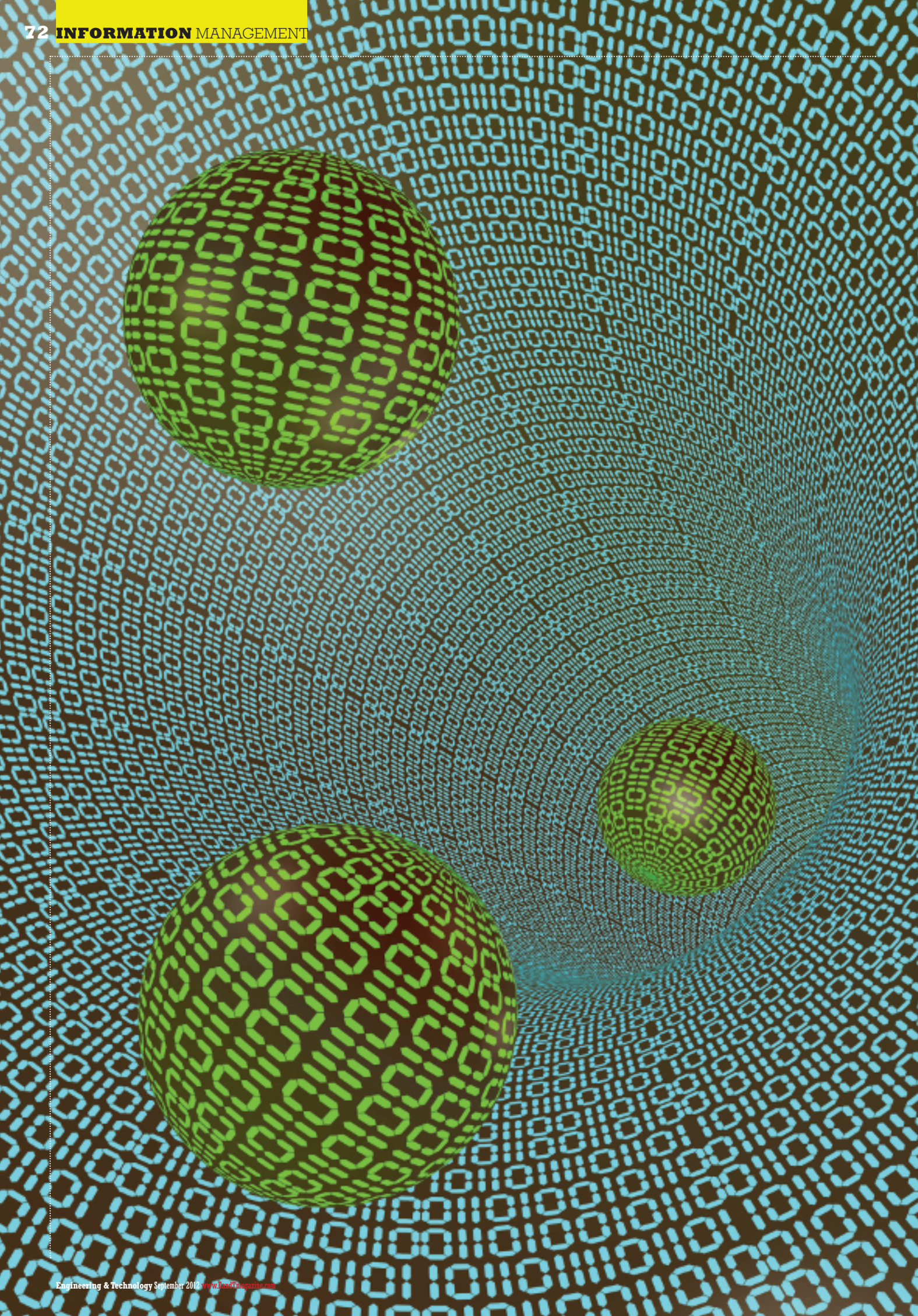

THE LARGING-UP OF BIG DATA

'Big data' is a buzz-term that is resonating big-time with IT solutions providers and end-user organisations. But are 'big data' applications really so different from the business intelligence and analytics tools that have been around for decades? **Martin Courtney** investigates.

THE TERM 'BIG DATA' has been getting big much exposure in IT circles over the last year or two, on a scale that is bound to cause seasoned industry-watchers to sniff the air for the familiar aroma of industry hyperbole. There is the customary amount of hype, of course, but there is more to it than the covert repackaging and repurposing of existing products.

In one sense 'big data' is a classic misnomer. The implication is that the volume of electronic information being generated and stored is now so large that existing database systems are no longer able to handle it.

It is certainly true that the world is generating data on an unprecedented scale, and it is going to escalate as trends such as machine-to-machine applications roll-out. However, it is not so much its size as the diversity of formats that data now comes in – particularly unstructured sources like text, email, instant messages, Web pages, audio files, phone records, videos – and what people want to do with it that presents the bigger problem.

"Most vendors are now realising that big data has actually very little to do with databases and more to do with information management," according to Clive Longbottom, director of analyst firm Quocirca. "Eighty per cent of an organisation's data is now electronic, yet 80 per cent of that is not held in a database, so cannot be dealt with just by throwing a big database at it." It is a question of "how you pull data from a Microsoft Office or whatever into an environment where it can be dealt with", Longbottom believes.

"Companies have always done big data – escalating amounts of information – but that is not really the definition of the term. It is more about the variety of the data and the velocity at which it comes at you," explains Dr David Schrader, director of marketing and strategy and data warehousing software firm Teradata. "Traditional customers may have a lot of data in tabular format – customer credit ratings tables, for example – which they need to join together in a variety of ways. For some customers it's megabytes, gigabytes, terabytes – the biggest with petabytes, like eBay, say." However, with entities like the Web, and social media sites like LinkedIn, the kind of analytics on those data sets are semi-structured. Schrader says it is "hard to force them into a relational database. It is far easier if you have database systems with the required speed to be up and running already to handle non-relational database data, systems able to run queries in parallel".

Compliance versus intelligence

Maintaining separate storage systems to handle all those different forms of data is generally inefficient, particularly if an individual or organisation wants to exploit all of the information it stores to use or for meaningful insight, and to do that fast enough to make the most of any business opportunity the exercise might subsequently provide. Most organisations keep data archived for compliance and regulatory purposes, at least on a temporary basis, before deleting. But others see the value in the information itself, and apply business intelligence and analytics tools to pull out statistics and patterns which they can turn to their advantage before discarding it.

Archiving data as insurance against potential e-discovery requests is relatively easy as the organisation does not need to know precisely what information is being kept, only that they can search it if necessary, while modest investment in the required capacity is easily offset against the cost of potential litigation. Arvind Krishna is IBM's general manager of information management, which, like Teradata, EMC, Oracle, and a host of other software application vendors, is making a big play for big data customers, albeit from a slightly different approach (see 'Big Data Goes To Work', p75). He recalls the case of a utility industry customer in the US running a power plant offering nuclear and fossil fuel.

"It had a bunch of systems from 20-30 years ago, and wanted to cut down storage and IT costs, but because of compliance and regulation it had to keep the old systems going to show the auditor what systems they were running to avoid accidents," Krishna says. "Now it can use metadata to search them, build a new archive [to house them] and keep it in a place where they can easily query it, and shut down all the stuff sitting in the main database. It can be much more cost effective than having two systems where there is some accountability, and can pay for itself in six months."

Quocirca's Longbottom agrees: "If this [stored data] is going to be something about people's mortgages, for example, we need to be able to prove how we put everything together to prove that opportunity, so when mis-selling cases hit the headlines it is maintaining that auditability as well."

Onboard the 'big data' bandwagon

When applying business intelligence and analytics tools to large repositories of structured and unstructured data on a regular basis, there is a danger that companies will spend time and money on new systems that are able to sift through information on an industrial scale, only to find that the data contains little or no value to the business anyway. As such, there are certain industries that are far more likely to gain advantage from big data projects than others, with healthcare, retail, utilities, and transport sectors top of the list.

We are already seeing the healthcare sector benefitting, because it has so much information that is not in databases, or is spread across multiple databases.

Longbottom argues that the retail sector "could do a lot with it because it has lots of stuff held in databases around loyalty cards, for example, and they often want to be >

'Most of an organisation's data cannot be dealt with just by throwing a great big database at it'
Clive Longbottom, Quocirca

JARGON BUSTER

Business intelligence (BI) Computer-based techniques used in identifying, extracting, and analysing business data, such as sales revenue by products and/or departments, or by associated costs and incomes. BI technologies provide historical, current and predictive views of business operations. Common functions of business intelligence technologies are reporting, online analytical processing, analytics, data mining, process mining, complex event processing, business performance management, benchmarking, text mining and predictive analytics. Until recently, BI applications have been seen mainly as the preserve of very large enterprises and organisations.

Big data Describes data sets that have grown so large that they become awkward to work with using on-hand database management tools. Typical difficulties include capture, storage, search, sharing, analytics, and visualising. This trend continues because of the benefits of working with larger and larger data sets, allowing analysts to discern and validate trends, such as tracking (and preventing) the spread of diseases.

Data mining Interdisciplinary field of computer science that describes the process of discovering new patterns from large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics and database systems.

Data analytics (DA) The science of examining raw data with the purpose of drawing conclusions about that information. Data analytics is used in many industries to allow companies and organisation to make better business decisions and in the sciences to verify or disprove existing models or theories. Data analytics is distinguished from data mining by the scope, purpose, and focus of the analysis. Data miners sort through huge data sets using sophisticated software to identify undiscovered patterns and establish hidden relationships. Data analytics focuses on inference, the process of deriving a conclusion based solely on what is already known by the researcher.

SOURCES: WIKIPEDIA, TECHTARGET, E&T RESEARCH.

What level of growth are you seeing in the following types of data within your organisation?

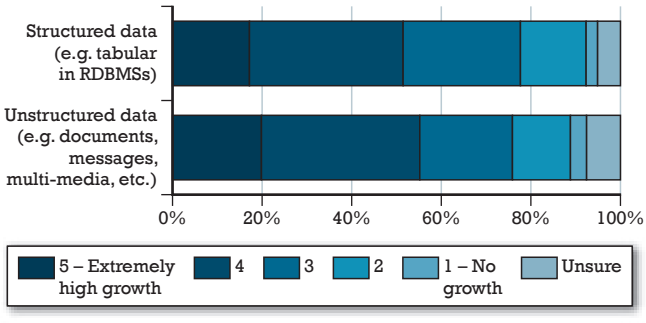

Table 1: Organisations are seeing data volumes increase, with unstructured data looking set to grow even faster than structured data in some cases, according to a survey by IT industry analyst Freeform Dynamics.

Considered overall, to what degree does your organisation exploit its information assets for analysis and decision making purposes?

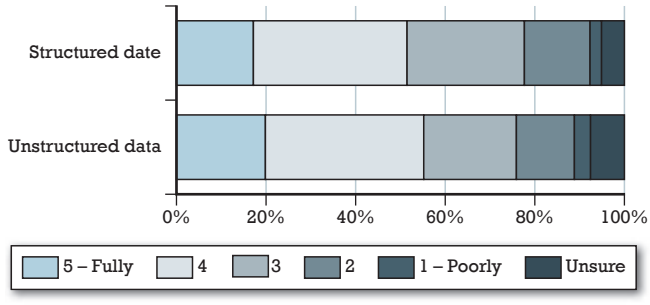

Table 2: Lack of clear return on investment is one key reason why so few organisations are extracting value from information held outside systems designed for handling structured data.

WWW.FREEFORMDYNAMICS.COM

< pulling data in from social networks to get a better idea of what customers and prospects are thinking".

He adds: "The utility companies have masses of data that is not being mined correctly, and they are not pulling in external information. Security agencies – MI5 or MI6, for example – have got to be thinking about picking-up patterns of information going across things like mobile phone records, email and what's happening on Twitter and Facebook so they can pull it all together and say 'right, this is the door that we go and knock down'."

Teradata's Dr David Schrader identifies the telephone companies or individual company call-centres as those which can benefit from big data analytics that interrogate call detail records (CDRs) to identify patterns around customer behaviour, as well as examples in the retail and transport industries.

"Think about eBay, the rate and volume of transactions, and the active intelligence you can gather from the data and put it in a database, for example," Schrader says.

"It is also about situational awareness in real-time – British Airways uses similar tools to replan operations in the event that a volcano blows and screws up [its schedules], with information on grounded planes, crew and passengers all at their fingertips in order to be able to construct an alternative [route and schedule]."

Profit wedge curve

Return on investment is always an ephemeral concept when it comes to business intelligence and Web analytics, but Schrader insists that big data solutions that are able to process so many different types of information in real-time provide better predictions on the effect of new sales or product strategies than earlier tools. So much so, according to Schrader, that a profit wedge curve – the classic V schematic, where growing revenue is offset by reduced costs to deliver increased profit – is very much a reality.

"A retailer, for example, would use Web analytics to see what would happen if they dropped something [a new product] into their website," he says, "but alongside traditional measures like sales or net promoter scores, you can now capture user tweets, which do not use tabular data, and get back an idea about who is happy with a new product, and who is not happy. Those can be critical."

Cloud services are the future

That return on investment depends to a large extent on the capital cost of the storage, processing and analytics resources to handle big data in the first place, which is generally not cheap.

Oracle, EMC, and Microsoft have rushed to introduce big data solutions based around Apache Hadoop, a platform that was created by Google to index the vast amounts of text and other document metadata it was collecting via the Internet to help improve its own search engine performance. Apache Hadoop is customised towards specific tasks and data types on an open source licence running on a specialised hardware appliance designed to be installed on the customer's premise.

BIG DATA GOES TO WORK

JSTART STARTS TO PULL IN BIG DATA BACK ENDS

IBM's Watson supercomputer: from TV gameshows to grappling with 'big data'

Many companies have successfully applied business intelligence or Web analytics tools to existing data warehouses or other databases, often integrating data from various unstructured sources into other applications, such as customer relationship management (CRM) or enterprise resource planning (ERP). Whether this constitutes a big-data solution or not is a moot point; but one company taking a different approach is IBM through its Jstart client engagement team, part of the IBM Software Solutions Group.

Car hire company Hertz engaged Flash tool Jstart to implement an analytics project based on a MindShare Technologies application to gather daily 'customer sentiment' information from unstructured sources as diverse as Web surveys, emails, and text messages in order to consistently analyse feedback and provide insights on problem areas which could be immediately addressed. As a result, Hertz says, it was able to identify areas for improvement in its Philadelphia office around delays for the return of vehicles at specific times of day and solve them by adjusting staffing levels at peak times.

Jstart was also the catalyst behind a big-data project for US healthcare company UNC Healthcare. The company used IBM's own text analytics software to mine patient data from various forms and databases to discover which patients were at greater risk of re-admittance to hospital, take proactive steps to minimise that risk, and therefore the chance of it being penalised under the new Medicaid regulations which fine healthcare providers who have excessive numbers of re-admittances.

More significantly for the future, IBM Jstart will help commercialise IBM's legendary Watson supercomputer platform – as used in a special 2011 edition of the US TV quiz show 'Jeopardy', whereby the system competed against humans to come up with correct answers as quickly as possible while receiving all information electronically as a text file. Ignore the artificial intelligence and text-to-speech technology which enabled Watson to perform, and the mighty computer is actually a data analytics and insight engine that uses a combination of text analytics, natural language processing and semantic systems to analyse the vast stores of information contained within its gigantic complement of hard disks and RAM.

After calling for proof of concept implementations last year, IBM is now looking to apply Watson's talents to tackle business orientated big data solutions. That means working with Jstart to produce lighter-weight configurations of Watson for different, application-specific tasks, then using that processing power to gather, organise and sift large volumes of unstructured data, understand the context of items of interest within the broader scope of the text it is discovered within to identify trends and patterns more accurately.

IBM general manager Arvind Krishna points out that Watson's real strength here is the ability to perform very sophisticated statistical analysis on the information gathered to understand 'relationships' and 'correlations' between data – many of which are either not immediately apparent or are very subtle. "If all you have is the book-keeping table," says Krishna, "it does not tell you whether the information is important or not. It needs to become application aware and business context aware, you cannot just do it by looking at a disk or a tape."

WEBLINK

There's more online...
Data management - will we ever press 'delete'?
http://bit.ly/eandt-dump-data
Unstructured data: nail it - then mine it
http://bit.ly/eandt-unstructured
IET Event: Big Data seminar, December 2012
http://conferences.theiet.org/big-data/index.cfm

That thinking is starting to change, with all the vendors looking to deliver more flexible, hosted big-data solutions available through cloud services which – in theory – could trim costs through an on-demand, pay-as-you-go model, as long as customer concerns around security and performance can be addressed. IBM led with the launch of its Hadoop-based InfoSphere BigInsights distributed data processing and analysis platform as a service (PaaS) in October 2011, with rivals seemingly set to follow.

"IBM's Watson makes more sense as a cloud solution rather than selling somebody a shedload of powerful [on-premise] systems," says Quocirca's Clive Longbottom. "Business Intelligence vendors are also moving towards the cloud – look at what they are doing when digging through 12TB of data in Facebook and other environments, it is much better that they have that control, their own security and data centres."

The problem with big data and the cloud: pushing large volumes of information over any network invariably risks performance and availability issues. This opens up the market to vendors keen to sell additional bandwidth optimisation solutions, and one reason why Teradata prefers to stick with the on-premise approach.

"That is a key engineering challenge," reckons Teradata's David Schrader. "Typically you want to push the computation as close to the data as possible – you don't wants bits and pieces all over the place, especially with call detail records (CDRs) for example. You would never want to copy 100 billion CDRs into the cloud to do the calculation, and that is why a lot of big companies prefer to have data at their fingertips in one system. Other than cloud surge capabilities, they have mostly tended to keep stuff in-house."

Is 'big data' actually that big?

Despite the continued frenzy of hardware and software vendors keen to sell their wares on the back of big-data initiatives, any project does necessarily require investment in new hardware and software if it is done correctly, says Quocirca's Clive Longbottom. He believes it is more about tweaking existing systems in the first instance. Deduplication, its advocates claim, can make a significant contribution to stripping away the 'slag' that can make data mining initiatives daunting at first sight.

"When you start looking at big data you find much redundant data: the same file in 48 different places, so if you can delete 47 of them, and just maintain pointers to all the rest, you instantly need less storage," Longbottom points out. "Once you get single instance you get less network traffic, so it can all be done correctly; but you need to plan correctly. As with anything to do with information management, it is a case of 'garbage in, garbage out' – you need to do data cleansing and de-duplication across the whole environment first so you end-up with something far cleaner, and look at master data modelling before you look at a big data solution." *

IBM, OD ILLUSTRATION