How to make sense of the big data universe

The prospect of analysing big data can appear daunting, so here **Clive Longbottom** looks at what data you need, where to find it and how to break it down into manageable, meaningful chunks

IMAGE: HUBBLESITE.ORG

CW BUYER'S GUIDE
BIG DATA
part 1 of 3

As is the way with IT, as soon as one bandwagon begins to be understood by the general public, another one has to be rolled out. In this case, as cloud computing starts to become more of a reality, big data is rearing its head as – depending on the commentator – the next greatest opportunity or threat to the organisation.

As there was with cloud, there's a lot of confusion out about big data. Many of the database vendors tried to play big data as purely having a lot of data in one or more databases. But that is not big data, it's large data – a problem that can be handled with database federation, standard business intelligence and analytics.

Next, it was said to be a mix of data held in the organisation that needed to be brought together so decision makers could see everything the organisation held around a specific topic to make better informed decisions – but only through whatever information the organisation was already aware of. So if the organisation wasn't already aware of something, that was to be excluded from the results – see the problem here?

Many technology companies – aided by the PR organisations employed to monitor their brands – pushed the idea that big data was moving towards the field of social networking. They said big data was all about using the wisdom of the crowd and identifying the sentiment of the masses.

But social networking has not usurped much that went before, so any solution still has to include all the information feeds such as e-mail, call recordings, customer relationship management (CRM) records, scanned documents and so on.

All the approaches cover some aspect of big data, but they all miss the point as well. The best, simple definition of big data comes down to volume, velocity and variety.

The volume aspect of big data is actually the one that is the least important. Big data is not about petabytes of data – it can be down to relatively small volumes that need to be dealt with in a manner that requires a big-data approach.

However, for most organisations, big data will involve bringing together many different data and information sources which, by their nature, will tend to result in the overall amount of data under consideration being big. Therefore, volume is not something that is under the direct control of the organisation – what has to be considered is how the volume of data that ends up being analysed is minimised, (more on this later).

Again, the velocity aspect of big data may well be a moot point – everyone »

« wants results against their analysis of available data in as short a period as possible. However, everything is relative – for example, every millisecond added to providing results to a financial market trader can cost millions of pounds, whereas someone tracking variations in the global movement of tectonic plates may not be that worried if results take a few seconds to come through.

The one aspect that really matters is the variety of the information. Big data is all about the mix of data and where it is held at any time. Here, formal databases under the organisation's direct control are only a very small component of the overall mix. There are all the office documents held as files across the organisation and you may need to include voice and video files as well.

Then there's the information held in the value chain of suppliers and customers – information that is critical to the process or service being provided, yet isn't under the organisation's direct control. Then, there may well be a requirement to include information from the various social networks out there – and whatever approach is taken has to be inclusive.

Inclusivity of data sources

For example, it is pointless constructing something that is Facebook-specific, if most comments are appearing as hashtags in Twitter.

Further, it's a waste of time writing multiple connectors to cover all of today's social networks – remember MySpace, Bebo and Second Life? They were all the darlings of their day, but have faded to a withered existence or almost non-existence as newer players have taken over.

Sites such as Pinterest are showing signs of major interest – yet this was also the case with Google+, which more resembles a Western desert than a viable, active social network, after just a short time.

Any social network solution has to be able to embrace new platforms at minimal cost, so new networks that are just "spikes" on the continuum do not use up lots of money in creating connectors specifically for them.

Even the largest organisations will have little control over anything beyond a small percentage of the total available data. The two-edged sword of the internet raises its ugly head in that it does provide massive extra information resources – but then again, it also includes a massive amount of dross that doesn't add anything to the sum knowledge of an organisation.

So how are we to deal with this real big data challenge, without running into Dilbert's pointy-haired boss's dictat, "Just run me off a copy of the internet"?

Storage and structure

Storage needs must be fully considered. EMC, NetApp and Dell are now talking about object, block and file storage, rather than focusing purely on high-performance database object storage to cover the various types of big data that needs to be controlled.

Other storage vendors, such as Nutanix, Coraid, Amplidata and FusionIO provide systems that focus on one aspect of big data, partnering where necessary to cover others.

The need for structure around semi- or unstructured data is leading to an explosion in interest in noSQL-based databases, such as Apache Cassandra, 10gen MongoDB, CouchDB and so on. Systems such as Apache's Hadoop, (which enables a massively scaled-out platform for providing distributed processing for large amounts of data), can use MapReduce, (the use of "chunking" data analysis into packets of work that can be dealt with in a parallel manner across a large resource pool), approaches to minimise the amount of information that needs to be dealt with.

What is being aimed for here is to take the seemingly infinite amount of available data and filter it down into manageable chunks. Standard internet searches can feed into a Hadoop-based system, which can then act as a feed into either standard SQL-based database or into a noSQL-based one, depending on the type of information being dealt with.

Extra information can be added automatically via rules engines or manually, as required, as metadata that adds to the value of the information stored. Once the information is held in a recognised form, it is then down to being able to apply the right form of data analysis against it to provide suitable feeds to the decision maker.


The key for buyers is to treat big data as a journey. Set short and medium-term targets of what is required

IMAGE: HUBBLESITE.ORG

This is where the main problems still reside, but much work is being carried out. Unsurprisingly, a lot of this is coming from the incumbent business intelligence suppliers, such as SAS Institute, QlikTech, JasperSoft as well as those who have gained entry to the market through acquisition such as IBM (Cognos, SPSS), SAP (Business Objects) and Oracle (Hyperion, Endeca).

The storage suppliers are also making plays in the space – EMC acquired GreenPlum and Dell continues to acquire companies that will help it create a more cohesive and complete big data approach.

Buyer dos and don'ts

The key for buyers is to treat big data as a journey. Set short- and medium-term targets of what is required and then put in place solutions that help to move towards these targets.

Don't put in place anything that could result in a need for major forklift upgrades at a later date – embrace open standards, look for suppliers who espouse heterogeneity in storage systems and in tooling, as well as an approach that covers a hybrid mix of private and public clouds.

Don't fall for any supplier who says that the world is moving to or from "standard" SQL-based databases – the move is to a mixed environment of a Hadoop-style system paired with SQL and noSQL-based systems. Look for business analytics packages that enable links to be made to data sources of any kind that reside anywhere on the internet, and that can link into semi-structured systems such as social networking sites in a meaningful manner.

Big data may appear to be just another bandwagon at this stage – but it is important, and needs to be addressed carefully and sensibly, rather than in a bull-in-a-china-shop manner that seems to be pushed by many vendors. The journey can be carried out at a measured pace, leveraging existing systems in conjunction with new systems. It just needs a strategic plan built from careful planning – and an eye to the long-term future. ■

Clive Longbottom is a director of analyst organisation Quocirca


IMAGE: HUBBLESITE.ORG

more online

› Decision makers plan to spend big on big data projects

› Security Think Tank: Using big data for intelligence-led security

› Making sense of big data in the petabyte age