bcs The Chartered Institute for IT

Standards that create the outstanding

Where are we with big data?

Go with the flow or drown

Introduction

There have been many descriptions of big data of late – mostly metaphors or similes for 'big' – deluge, flood, explosion... but not only is there a lot of talk about big data, there is also a lot of data. What can we do with structured and unstructured data? Can we extract insights from it? Is 'big data' just a marketing puff term?

There is absolutely no question that there is an awful lot more data around now than there was. BCS's data migration blogger recently commented on the fact that some telco's are 'de-tuning or switching off entirely some of their monitoring platforms due to the sheer volume of data and their inability to store it, never mind process it meaningfully.'

IBM say that 'every day, we create 2.5 quintillion bytes of data – so much that 90 per cent of the data in the world today has been created in the last two years alone.'

Clearly there is enough data around that if we can get any meaningful analyses from it then 'marketing puff' concern can be allayed. But where is all this data coming from?

Sources

Social media platforms produce huge quantities of data, both from individual network profiles and the content that influencers and the less influential alike produce. Short form blogging, link-sharing, expert blog comments, user forums, 'likes' and more all contain potentially useful information.

There is also data produced through sheer activity: machine-generated content in the form of device log files, which could be characterised as the internet of things. This would include output from such things as geo-tagging.

Yet more data can be mined from software as a service and cloud applications, data that's already in the cloud but mostly divorced from internal enterprise data.

Another large, but at this stage largely untapped, area is the data languishing in legacy systems, which includes things like medical records and customer correspondence.

Caveats

A recent post from BCS's future blogger called into question some of the behind the scenes story: 'For the big data commercial advocates, there must be algorithms that can trawl the data and create outcomes better, that is to say more cost effective, than traditional advertising. Where is the evidence that such algorithms exist? How will these algorithms be created and evaluated and improved upon if they do exist. One problem is that in a huge dataset, there may be many spurious correlations and the difference between causation and correlation hard to prove.'

As we would perhaps expect, the likes of IBM say this goes beyond hype: 'While there is a lot of buzz about big data in the market, it isn't hype. Plenty of customers are seeing tangible ROI using IBM solutions to address their big data challenges.'

Big Blue go on to quote a 20 per cent decrease in patient mortality by analysing streaming patient data in the healthcare arena; a telco that enjoyed a 92 per cent decrease in processing time by analysing networking and call data and a whopping 99 per cent improved accuracy in placing power generation resources by analysing 2.8 petabytes of untapped data for a utilities organisation.

Tools

To handle large datasets in times gone by enterprises used relational databases and warehouses from proprietary suppliers. But they just can't handle the volumes of data being produced. This has seen a trend towards some open source alternatives such as Hadoop, which Wikipedia defines as 'an open-source software framework that supports data-intensive distributed applications, licensed under the Apache v2 license. It supports the running of applications on large clusters of commodity hardware.'

Wired recently reported on Cloudera — one of several companies that help build and use Hadoop applications, which is offering a Google-style search engine for Hadoop called, uninspiringly, Cloudera Search.

Interestingly, Wired point to a recent Microsoft paper on whether customers really need to put all their data in Hadoop. It argued, says Wired, that 'most companies don't (have) data problems that justify the use of big clusters of servers. Even Yahoo and Facebook, two of the companies most associated with big data, are using clusters to solve problems that could actually be done on a single server.'

Despite that interest is on the up and big organisations are taking advantage. A recent piece from The Sun Daily mentions that 'analyst firm International Data Corp projects the global big data technology and services market will grow at a compound annual growth rate of 31.7 percent – about seven times the rate of the overall information and communications technology market.'

The same article reports further investment in the perceived future of big data with the recent announcement by Dell, Intel Corporation and Revolution Analytics of the Big Data Innovation Centre in Singapore. The new centre brings together expertise from all three organisations to provide training programmes, proof-of-concept capabilities and solution development support on big data and predictive analytic innovations catering to the Asian market.

How and when

The 'when' of embracing any new technology is massively variable depending on your organisation's aims, business sector and so on.

Some of the things that could affect your timing are neatly summed up by Redmond magazine in a recent article, simply by listing some of the possible motivators. They mention that you could utilise 'CRM systems and data feeds to tweets mentioning their organisations that can alert them to a sudden problem with a product.' If this kind of real-time feedback is of benefit then dipping a toe into the deluge of the big data waters is best done sooner rather than later.

Another area mentioned is 'potential market opportunities spawned by an event'. Not as business-critical as product feedback, but important in a time of global austerity.

Redmond magazine also mentions things such as online and big-box retailers using big data to automate their supply chains on the fly and law enforcement's analysing huge amounts of data to thwart potential crime and terror attacks. The scope and motivations vary widely, but potential benefits are both long- and short-term.

As to how to go about it, some of the tools are mentioned above, often oriented around Hadoop. Microsoft recently launched Windows Azure HDInsight and Redmond magazine also cited VMware's key application infrastructure and big data and analytics portfolio called Pivotal.

There's plenty to read about, as the following list shows...

Further reading

Microsoft's special report on using clusters for analytics:
http://research.microsoft.com/apps/pubs/default.aspx?id=179615

Victor Mayer-Schonenberger and Kenneth Cukier, 'Big Data' review:
http://www.bostonglobe.com/arts/books/2013/03/05/book-review-big-data-viktor-mayer-schonberger-and-kenneth-cukier/T6YC7rNqXHgWowaE1oD8vO/story.html

IBM on big data:
www-01.ibm.com/software/data/bigdata

Wired on Cloudera: **www.wired.com/wiredenterprise/2013/06/cloudera-search**

Asian market for big data:
www.thesundaily.my/news/734375

The hardware perspective:
www.techrepublic.com/blog/big-data-analytics/are-we-headed-for-a-platform-change-for-big-data/445?tag=content;blog-list-river

Big data sources:
www.zdnet.com/top-10-categories-for-big-data-sources-and-mining-technologies-7000000926/

Hadoop:
http://en.wikipedia.org/wiki/Hadoop

Things you should know about implementing big data:
http://redmondmag.com/articles/2013/05/01/buried-in-big-data.aspx

About BCS

We help global enterprise align its IT resource with strategic business goals. We work with organisations to develop people, forge culture and create IT capabilities fit to not only lead business change but to meet companywide objectives and deliver competitive advantage.

IT has been gaining momentum within global business for decades and we've been there from the beginning, nurturing talent and shaping the profession into the powerhouse that's now driving our digital world. Today organisations partner with us to exploit our unique insight and independent experience as we continue to set the standards of performance and professionalism in the industry.

Call us on +44 (0) 1793 417 755 or visit us at enterprise.bcs.org

BCS The Chartered Institute for IT
First Floor Block D North Star House North Star Avenue Swindon SN2 1FA
T +44 (0) 1793 417 655 **enterprise.bcs.org**

© BCS, The Chartered Institute for IT is the business name of The British Computer Society (Registered charity no. 292786) 2013

If you require this document in accessible format please call +44 (0) 1793 417 600

BC267/LD/ART/0613