The dawn of Big Data

 Businesses need to adapt rapidly if they are to prosper in a world where massive, high-velocity data is everywhere. Smart businesses are already racing to harness its power in order to gain competitive advantage. Do advanced analytics hold the key to profiting from the Big Data explosion? Research reveals a widening divide, with those organisations already using analytics more than twice as likely to outperform their industry peers. What does your business need to become one of them?
Our ability to generate data has always outstripped our ability to organise it. Paper emerged around 100AD but it took another 1,800 years to come up with the filing cabinet. Today’s businesses can’t wait that long. By 2011, it was estimated that 90 per cent of the world’s data in circulation had been generated in the preceding two years. How can advanced analytics unlock the insights buried in the almost unimaginable amounts of data now being generated? The ability to analyse Big Data has already transformed the world of science, unravelling the human genome and identifying the elusive Higgs boson. Now it is transforming the way the world does business.
WHAT CAN BIG DATA DO?

By applying advanced analytics, businesses can tap into their own Big Data streams to gain deeper insights, enhance the efficiency of operations and deliver business transformation. It also allows organisations to differentiate themselves with new services for customers and new ways of working for the business itself.

This has implications for businesses of all types, especially those already working in a data-rich environment. For example, mobile telecoms providers are able to capitalise on the deluge of data from smart phones to provide customers with new services. Telecoms can access and analyse Big Data – such as browsing logs, download histories and GPS location data – to create better products and align services with customer needs.

Big Data also puts the customer centre stage. What makes Big Data analytics so valuable is its ability to pinpoint the needle in the haystack – and to do so in real time. This creates unique opportunities to deliver highly granular personalised services.

For example, network operators know that dropped calls and poor reception are drivers for customer churn; they also know who their high value customers are. What if they could identify a dropped call, identify the customer – and then offer that customer a personalised apology and a compensatory discount the moment something goes wrong? That would have been virtually impossible five years ago.

Today, Big Data provides the raw materials needed for businesses to develop a much deeper understanding of their customers through brand insight and reputational analytics. Businesses can now gauge customer sentiment and behaviour easily, and tailor products and services to achieve deeper market penetration.

In the past, the only way this could be achieved was through expensive market research and by using basic information provided by customers themselves. In the case of banks and financial services providers, that includes information such as when an account was opened and the frequency and nature of transactions.

Traditional structured data still has immense value – Big Data analytics takes this to the next level. It adds to the equation a whole raft of semi-structured and unstructured data that had previously been too big and too fast to handle. Now, virtually every interaction can be considered.

For example, when a customer contacts a call centre, the content of the conversation can be analysed; if the customer has exchanged information using the bank’s website, that interaction can also be taken into account. It is also possible to scan what customers are saying on blogs and social media websites. This could provide organisations with a 360° view of the customer for the first time. The insurance sector is also set to benefit from the Big Data revolution. Traditionally, insurers have invested in conventional data warehouse architecture to help them price products and provide management information. These systems provide business value, but they don’t provide the big picture view needed to deliver real transformation.

For insurers, one of the benefits of Big Data usage could be the ability to combat fraud more effectively, something that can be achieved by correlating claims with social media activity and GPS location data. This can provide a far more accurate view of what really happened in individual incidents than has previously been possible.

Big Data analytics also makes it possible to deliver pay-as-you-drive (PAYD) insurance for motorists. PAYD hinges on the ability to match premiums to actual driving behaviour, rather than statistical abstractions based on a driver’s history and risk profile. To deliver PAYD effectively, insurers need to make sense of massive data flows from individual vehicles.

Big Data could also meet the needs of the entertainment industry, making it possible to measure the likely success of upcoming film releases by analysing social media postings made by potential cinemagoers. Data about specific films could be filtered from tens of thousands of Twitter feeds, providing a clear indication of which films are getting talked about and which aren’t, to provide an accurate measure of audience sentiment.

Insights gained can be used to reallocate promotional resources based on needs, with last-minute fine tuning to pre-release campaigns that provide films with an extra push before launching at the box office. By leveraging Big Data from social media streams, marketers would be tapping into what are, in effect, the biggest focus groups on earth.

The ability to dig deeper, faster, has important implications for businesses that routinely generate huge amounts of data. Big Data analytics allowed one major multinational car hire business to transform customer operations by redeploying staff to meet demand surges at peak times. This insight was gained by centralising the processing of customer satisfaction surveys – something that had previously been conducted locally – and applying advanced analytics to parse data from multiple sources to identify patterns and problems that were previously invisible.

The ability to make sense of Big Data has significant implications for oil & gas businesses as well. The industry has no shortage of data – an oilfield, for...
example, can produce around one terabyte of data every day. What is in short supply are actionable insights.

Deploying advanced analytics can help engineers and other stakeholders get their hands on the insights they need, identifying key patterns and converting data into actionable insights to support better decision making. With oil & gas companies increasingly exploiting reserves in highly complex geological environments, the ability to make sense of growing amounts of data is likely to become a differentiator.

Some of the most exciting potential advances, though, could hinge on collaborative ventures between different data custodians such as telecoms providers and retailers. Businesses like these already have an intimate electronic knowledge of their customers’ habits.

By combining this conventional structured data with additional unstructured data streams – what a customer is saying on Facebook, Twitter, LinkedIn and blogs, for example – it is possible to reveal an even clearer picture of that customer’s tastes. By further combining that with real-time location data, the prospect of targeting the right offer, at the right time, to the right person, becomes a real one.

Contextual targeting of this sort is already a possibility and it should become easier to deliver – and more powerful – as Big Data analytics becomes ubiquitous. But businesses that go down this route need to make careful judgements about the reputational impacts of deeper engagement: the line between an unbeatable offer and unacceptable intrusion is, potentially, a thin one.

UNDERSTANDING BIG DATA
The term “Big Data” first emerged in the public arena during 2008. The word “big” reflects the fact that the volume of stored and streaming data had become too big for conventional analytics and business intelligence systems to handle. Big Data is generally considered to have three characteristics: volume, variety and velocity. Behind each of these three is a technology that is making analysis possible.

**Volume** refers to the scale of data being generated, currently estimated to be in the region of 2.5 quintillion bytes of data each day. However, it should be noted that individual businesses continue to manage high-volume data successfully, with complex terabyte-scale models capable of handling billions of observations. These already exist in conventional data warehousing and business intelligence, along with software that can handle semi-structured data and text analytics. The difference today is that data warehouses are evolving, offering ever-increasing power to tackle volumes that are only going to grow.

**Variety** is a reflection of the upsurge in heterogeneous data sources, most of them generating semi-structured or unstructured data – digital gibberish – that has traditionally proved resistant to easy analysis. These include feeds from sensors embedded in plant and vehicles, social media streams, audio, video and text.

To make sense of this volume and variety of data, the analytics involved must be distributed across a large number of machines rather than relying on a single, centralised database. From IBM’s perspective, Apache™ Hadoop® leads in this sphere, an open source software framework that allows for large data sets to be processed across clusters of servers.

**Velocity** adds a further layer of complexity. Traditionally, data has been batch processed and scrubbed clean before being used by businesses. Big Data, by contrast, is a moving target, with mass flows from a huge array of sources. Because there is too much to store, insight must be acquired in real time. One of the things that makes Big Data so useful to businesses is the value that can be extracted from it while it is still fresh.

Stream computing allows businesses to analyse and act on high-velocity, high-volume data – such as social media and GPS feeds – while it is in motion. This represents a fundamental shift away from traditional analytics, which depends on first accumulating and structuring data. Stream computing is intended to analyse and identify facts on the fly. This is an essential function of the Big Data environment because in many cases the sheer volume of raw data would be too great to store in the first place.

The combination of these three elements – Volume, Variety and Velocity – creates a unique opportunity for businesses to tap into the true potential of Big Data and allow organisations to create value in two ways:

1. Continue to develop actionable insight through business analytics but on a wider and deeper scale. Many consumer-facing organisations already employ

2.5 quintillion

(that’s $2.5 \times 10^{18}$) bytes of data have been created every day, as of 2012.
analytics in order to gain insights into their customers. Big Data analytics approaches could increase the depth and breadth of this capability.

2. Create new opportunities previously unimaginable or not practical due to technology constraints. Not only should organisations be able to analyse data previously out of reach, but they will be able to act upon the discovered insights in near real-time.

**BIG DATA TECHNOLOGY**

IBM's view is that Big Data should be treated as a platform in its own right. It is not a single solution, however, but a suite of technologies, concepts and practices.

Big Data does not mean moving away from traditional business intelligence applications or abandoning recent investment in predictive analytics applications and it certainly does not suggest shutting down the enterprise data warehouse. On the contrary: all of these can be plugged in to the Big Data platform. A number of technologies are shaping the shift to Big Data.

Attempting to manage Big Data without solutions like Hadoop® is a potentially risky and expensive strategy, particularly if that data is heading straight for a conventional data warehouse. This type of framework helps by allowing organisations to offload that data into a separate cluster where advanced analytics are applied. This can reduce the cost and complexity associated with conventional data extraction, transformation and loading (ETL), and can create a new way of enabling exploration and insight discovery. This could dramatically reduce the time to insight when considering the potential value of previously unexploited data sources.

Hadoop® is a framework rather than a solution. That means tools are needed to drive processes and derive insights. Data scientists – experts tasked with solving complex data problems – are increasingly being relied on to write the processes for analysing the data, which means it can be time-consuming to get Big Data analytics up and running.

Individual data scientists work in everything from data engineering to mathematics, statistics and computing, but their expertise may be limited to one or two of these areas – it’s unlikely for any one data scientist to cover all the bases. As a consequence, data science tends to involve teams and there have already been concerns raised over skills shortages in this area². However, the emergence of high-level tools – such as IBM InfoSphere® BigInsights™ – is playing a decisive part in democratising distributed analytics. BigInsights™ is built on top of Hadoop®, and allows much more complex analysis to be written by developers in a familiar user interface. The result is that business users and business analysts can have direct interaction with their data, without needing anybody to write new code for them.

Frameworks such as Hadoop® and analytics solutions such as BigInsights™ can leverage the power of data at rest to deliver insights. But what about data on the move? One of the most interesting and promising developments in Big Data is the ability to carry out real-time analytics on data as it enters the organisation.

IBM InfoSphere® Streams meets the need for real-time stream analytics, providing a high-performance computing platform that allows user-developed applications to rapidly ingest and correlate real-time information from thousands of different sources. Solutions of this sort have the capability to handle data rates that are measured in petabytes per day and they are critical if businesses are to leverage the full potential of Big Data.

The capacity to make sense of massive amounts of chaotic and unstructured real-time data is a game changer. Around 80 per cent of the world's data falls into this category. Any digital data stream is a potential candidate. These include everything from freeform text and images to audio, voice traffic generated by call centres, VoIP, video, web traffic, email, GPS data, financial transaction data, satellite data and sensor outputs. All of these digital streams can be turned into insight – often instantly – and delivered into the business ecosystem.

**HOW DO YOU PUT BIG DATA TO WORK?**

Businesses that have already successfully embraced business analytics have a head start. To succeed with Big Data, as with analytics, organisations need to develop three core competencies.

1. **The will to drive transformation**
   Management commitment is needed to translate insights and outputs into action and to ensure that those insights are readily accessible. Both the practitioners who carry out the analysis and the leaders of the organisation around them have a vital role to play in achieving this.

   Big Data analytics need to be embedded in business processes. Businesses may be swift to adopt analytics, but the effort is wasted if hard-won insights are simply shunted into reports that are filed and forgotten. This is an opportunity for organisations to change the way things are done and to expand into new opportunities, triggering new actions across the business on the path to greater value. The insights gained should link to business strategy in a way that is easy for end users to understand.

   Big Data demands a new approach, with exploratory and iterative analysis to cope with semi-structured and unstructured data. This requires a new level of dialogue between IT and business stakeholders. IT departments need to understand the technologies and what is possible; business stakeholders need to come forward with questions and use cases. Simply building a Big Data platform and sitting back to see what happens is not the right approach.

2. **Understanding the tools and methods**
   The ability to understand and apply analytics skills, tools and methods to data is of critical importance. Businesses need to recognise that the information
they hold – and are acquiring – is a unique strategic asset. Building a Big Data platform is just the first step. High-level frameworks such as Hadoop® can only be exploited with access to the right tools: organisations that want to get ahead of their competitors are either investing in high-level tools or coding for themselves, in some cases from scratch.

Organisational capability is a factor and Big Data provides an opportunity to make data scientists out of existing teams by taking advantage of the accelerators that IBM has to offer that reduce barriers to entry and time to insight.

3. An approach to veracity and relevance
Approaches to information management must evolve to take account of Big Data opportunities. Organisations will need to recognise the fundamental differences between traditional “refined” data and “raw” Big Data, which is much harder to handle.

Today, most of the data used by businesses is highly structured. Data is subjected to ETL processes in which it is first extracted, then transformed to meet operational needs and finally loaded into an end target such as a data warehouse or database.

Approaches of this sort will continue to be required for conventional high-value data. However, the time lag between acquisition and data being converted into something that is useful for the business – a data mart or a dashboard, for example – can sometimes be lengthy.

It’s important to remember that there is another “V” in the equation – Veracity, which introduces the idea that not all Big Data is high quality.

What’s more, organisations need to embrace the idea that not all of their Big Data is going to be relevant all of the time. In fact, the bulk of any given Big Data set may be of low or no relevance at all so it ought to be discarded or pre-processed into higher order information structures (eg, routes rather than raw location).

These Big Data sets we encounter allow us to preserve the concept of a high integrity, high data and information quality architecture as well as a core single version of the truth for the business.

WHERE NEXT?

● Decide what sort of capabilities you want to have as a business.
Then use a third-party partner to reduce the time to insight and help you get your proof of concept going.

● Recognise that your Big Data project is likely to be the first of its kind.
Big Data is virgin territory, so think creatively about data sources both internal and external. New data sources create new opportunities.

40% growth in global data generated per year is expected versus only 5% growth in global IT spending*
Information governance is key. Organisations are entering new territory, particularly with regard to how they use their customers’ information and how they deliver products and services. This raises important questions about data security, privacy and – equally important – customers’ perceptions of how their personal information is being managed.

Businesses must understand the legal constraints governing the way data is used and they must prepare for an open dialogue with customers to explore the mutual benefits of Big Data usage.

The skills needed to bring Big Data to life are already in your organisation. Big Data is an opportunity for businesses to unshackle their in-house analysts. They are vital to the success of Big Data in your enterprise. Today’s business analysts are tomorrow’s data scientists: allow them to be the drivers and proponents of the data revolution.

HOW IBM CAN HELP
IBM has the broadest Big Data capability in the industry with a strategy that is dedicated to helping businesses go further. To do this, IBM provides market-leading software and the hardware needed to run it. It also provides services and consultancy to help businesses benefit from Big Data – a capability that is unique in the marketplace.

IBM’S CORE BIG DATA CAPABILITIES INCLUDE:
- Hadoop®-based analytics to process and analyse any data type across commodity server clusters.
- Stream computing to enable continuous analysis of massive volumes of streaming data with sub-millisecond response times.
- Data warehousing to deliver deep operational insight with advanced in-database analytics.
- Information integration and governance that allows organisations to understand, cleanse, transform, govern and deliver trusted information to enterprise-critical initiatives.

In addition to this, IBM provides a range of supporting data platform services. These include:
- Visualisation and discovery to help end users explore large, complex data sets.
- Application development to streamline the process of developing big data applications.
- Systems management to monitor and manage Big Data systems for secure and optimised performance.
- Accelerators to speed-up time to value with analytical and industry-specific modules.
IBM Contacts

Jonathan Batson
Associate Partner
UKI Advanced Analytics Leader
IBM Business Analytics & Optimisation (BAO)
+44 (0)7710 189256
jonathan.r.batson@uk.ibm.com

Matin Jouzdani
Senior Managing Consultant
UKI Big Data Leader
IBM Business Analytics & Optimisation (BAO)
+44 (0)7841 495021
matin.jouzdani@uk.ibm.com

© Copyright IBM Corporation 2012

IBM United Kingdom Limited
76 Upper Ground
South Bank
London
SE1 9PZ

The IBM home page can be found at ibm.com

IBM, the IBM logo and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries or both. If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or ™), these symbols indicate US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries.

A current list of IBM trademarks is available on the web at “Copyright and trademark information” at www.ibm.com/legal/copytrade.shtml

References in this publication to IBM products or services do not imply that IBM intends to make them available in all countries in which IBM operates. Copying or downloading the images contained in this document is expressly prohibited without the written consent of IBM. This publication is for general guidance only.

All rights reserved.

NIW03034

References

5. Ibid.