THE WHITE BOOK OF...

Big Data

The definitive guide to the revolution in business analytics

shaping tomorrow with you

FUJITSU

FUJITSU

THE WHITE BOOK OF...

Big Data

The definitive guide to the revolution in business analytics

THE WHITE BOOK OF...
Big Data

Contents

Acknowledgements 4

Preface 5

1: What is Big Data? 6

2: What does Big Data Mean for the Business? 16

3: Clearing Big Data Hurdles 24

4: Adoption Approaches 32

5: Changing Role of the Executive Team 42

6: Rise of the Data Scientist 46

7: The Future of Big Data 48

8: The Final Word on Big Data 52

Big Data Speak: Key terms explained 57

Appendix: The White Book Series 60

Acknowledgements

With thanks to our authors:

- **Ian Mitchell, Chief Architect, UK & Ireland, Fujitsu**
- **Mark Locke, Head of Planning & Architecture, International Business, Fujitsu**
- **Mark Wilson, Strategy Manager, UK & Ireland, Fujitsu**
- **Andy Fuller, Big Data Offering Manager, UK & Ireland, Fujitsu**

With further thanks to colleagues at **Fujitsu in Australia, Europe and Japan** who kindly reviewed the book's contents and provided invaluable feedback.

For more information on Fujitsu Big Data capabilities and to learn how we can assist your organization further, please contact us at AskFujitsu@us.fujitsu.com or contact your local Fujitsu team (see page 62).

Preface

In economically uncertain times, many businesses and public sector organizations have come to appreciate that the key to better decisions, more effective customer/citizen engagement, sharper competitive edge, hyper-efficient operations and compelling product and service development is **data – and lots of it**. Today, the situation they face is not any shortage of that raw material (the wealth of unstructured online data alone has swollen the already torrential flow from transaction systems and demographic sources) but how to turn that amorphous, vast, fast-flowing mass of "Big Data" into **highly valuable insights, actions and outcomes**.

This Fujitsu *White Book of Big Data* aims to cut through a lot of the market hype surrounding the subject to **clearly define the challenges and opportunities** that organizations face as they seek to exploit Big Data. Written for both an IT and wider executive audience, it explores the different approaches to Big Data adoption, the issues that can hamper Big Data initiatives, and the new skillsets that will be required by both IT specialists and management to deliver success. At a fundamental level, it also shows how to map business priorities onto an action plan for turning Big Data into **increased revenues and lower costs**.

At Fujitsu, we have an even **broader and more comprehensive vision for Big Data** as it intersects with the other megatrends in IT – cloud and mobility. Our **Cloud Fusion innovation** provides the foundation for business-optimizing Big Data analytics, the seamless interconnecting of multiple clouds, and extended services for distributed applications that support mobile devices and sensors.

We hope this book offers some perspective on the opportunities made real by such innovation, both as a Big Data primer and for ongoing guidance as your organization embarks on that extended, and hopefully fruitful, journey. Please let us know what you think – and how your Big Data adventure progresses.

Neil Jarvis
CIO
Fujitsu America, Inc.

1 What is Big Data?

In 2010 the term 'Big Data' was virtually unknown, but by mid-2011 it was being widely touted as the latest trend, with all the usual hype. Like "cloud computing" before it, the term has today been adopted by everyone, from product vendors to large-scale outsourcing and cloud service providers keen to promote their offerings. But what really is Big Data?

In short, Big Data is about *quickly deriving business value from a range of new and emerging data sources,* including social media data, location data generated by smartphones and other roaming devices, public information available online and data from sensors embedded in cars, buildings and other objects – and much more besides.

Defining Big Data: the 3V model

Many analysts use the 3V model to define Big Data. The three Vs stand for volume, velocity and variety.

Volume refers to the fact that Big Data involves analyzing comparatively huge amounts of information, typically starting at tens of terabytes.

Velocity reflects the sheer speed at which this data is generated and changes. For example, the data associated with a particular hashtag on Twitter often has a high velocity. Tweets fly by in a blur. In some instances they move so fast that the information they contain can't easily be stored, yet it still needs to be analyzed.

Variety describes the fact that Big Data can come from many different sources, in various formats and structures. For example, social media sites and networks of sensors generate a stream of ever-changing data. As well as text, this might include, for example, geographical information, images, videos and audio.

Data speed
In a Big Data world, one of the key factors is speed. Traditional analytics focus on analyzing historical data. Big data extends this concept to include real-time analytics of in-flight transitory data.

Photograph: iStockphoto

Linked Data: a new model for the database

The growth of semi-structured data (see 'Data types', right) is driving the adoption of new database models based on the idea of 'Linked Data'. These reflect the way information is connected and represented on the Internet, with links cross-referencing various pieces of associated information in a loose web, rather than requiring data to adhere to a rigid, inflexible format where everything sits in a particular, predefined box. Such an approach can provide the flexibility of an unstructured data store along with the rigor of defined data structures. This can enhance the accuracy and quality of any query and associated analyses.

Data sources
Big Data not only extends the data types, but the sources that the data is coming from to include real-time, sensor and public data sources, as well as in-house and subscription sources.

Value: the fourth vital V

While the 3V model is a useful way of defining Big Data, in this book we will also be concentrating on a fourth, vital V—value. There is no point in organizations implementing a Big Data solution unless they can see how it will give them increased business value. That might not only mean using the data within their own organization—value could also come from selling it or providing access to third parties. This drive to maximize the value of Big Data is a key business imperative.

There are other ways in which Big Data offers businesses new ways to generate value. For example, whereas traditional business analytical systems had to operate on historical data that might be weeks or months out of date, a Big Data solution can also analyze information being generated in 'real time' (or at least close to real time). This can deliver massive benefits for businesses, as they are able to respond more quickly to market trends, challenges and changes.

Furthermore, Big Data solutions can add new value by analyzing the sentiment contained in the data rather than just looking at the raw information (for example, they can understand how customers are feeling about a particular product). This is known as 'semantic analysis'. There are also growing developments in artificial intelligence techniques that can be used to perform complex "fuzzy" searches and unearth new, previously impenetrable business insights from the data.

In summary, Big Data gives organizations the opportunity to exploit a combination of existing data, transient data and externally available data sources in order to extract additional value through:

- **Improved business insights** that lead to more informed decision-making
- **Treating data as an asset** that can be traded and sold.

It is therefore important that organizations keep sight of both the long-term goal of Big Data—to integrate many data sources in order to unlock even more

Data types

IT people classify data according to three basic types: structured, unstructured and semi-structured.

Structured data refers to the type of data used by traditional database systems, where records are split into well defined "fields"(such as "name", "address", etc) which can be relatively easily searched, categorized, sorted according to certain criteria, etc.

Unstructured data, meanwhile, has no obvious pre-defined format, for example image data or Twitter® updates.

Semi-structured data refers to a combination of the two types above. Some aspects of the data may be defined (typically within the information itself, e.g. location data appended to social media updates but overall it does not have the rigidity associated with structured data.

The drive to maximize the value of Big Data is a key business imperative.

potential value – while ensuring their current technology is not a barrier to accuracy, immediacy and flexibility.

In many respects Big Data isn't new. It is a logical extension of many existing data analysis systems and concepts, including data warehouses, knowledge management (KM, business intelligence (BI, business insight and other areas of information management.

Big Data: the new "cloud"

The trouble with all new trends and buzz-phrases is that they quickly become the latest bandwagon for suppliers. As noted at the start of this chapter, all manner of products and services are now being paraded under the "Big Data" banner, which can make the topic seem incredibly confusing (hence this book). This is compounded when vendors whose products might only pertain to a small part of the Big Data story grandly market them as "Big Data solutions", when in fact they're just one element of a solution. As a marketing term, then, be aware that "Big Data" means about as much as the term "cloud" – i.e. not a great deal.

When is "big" really big?

History tells us that yesterday's big is today's normal. Some over-40s reading this book will probably remember wondering how they were ever going to fill the 1 kilobyte of memory on their Sinclair ZX81. Today we walk around with tens of

gigabytes of memory on our smartphones. Big Data simply refers to volumes of data bigger than today's norm. In 2012, a petabyte (1 million gigabytes) seems big to most people, but tomorrow that volume will become normal, and – over time – just a medium-to-small amount of data.

What's driving the need for Big Data solutions over traditional data warehouses and BI systems, therefore, isn't some pre-defined "bigness" of the data, but a combination of all three Vs. From a business perspective, this means IT departments need to provide platforms that enable their business colleagues to easily identify the data that will help them address their challenges, interrogate that data and visualize the answers effectively and quickly (often in near real time). So forget size – it's all about "speed to decision." Big Data in a business sense should really be called "quick answers."

Near enough or mathematically perfect?

IT departments need to provide platforms that enable their business colleagues to easily identify the data that will help them address their challenges.

When the concept of Big Data first emerged, there was a lot of talk about "relative accuracy." It was said that over a large, fluid set of data, a Big Data solution could give a good approximate answer, but that organizations requiring greater accuracy would need a traditional data warehouse or BI solution. While that's still true to a degree, many of today's Big Data solutions use the same algorithms (computational analysis methods) as traditional BI systems, meaning they're just as accurate. Rather than fixating on the mathematical accuracy of the answers given by their systems, organizations should instead focus on the business relevance of those answers.

Big Data is so yesterday

Since Big Data has only been in common use since mid-2009, it might seem natural to assume that early adopters face the usual slew of teething problems. However, this is not the case. That's not because the IT industry has become any better at avoiding such problems. Rather, it's because although the term 'Big Data' may be relatively new, the concept is certainly not.

Consider an organization like Reuters (whose business model is based on extracting relevant news from a mass of data and getting it to the right people as quickly a possible) – it has been dealing with Big Data for over 100 years. In more recent years, so have Twitter, Facebook®, Google®, Amazon®, eBay® and a raft of other well-known online names. Today, the bigger problem is that so much data is thrown away, ignored or locked up in silos where it adds minimal value. Being able to integrate available data from different sources in order to extract more value is vital to making any Big Data solution successful. Many organizations already have a data warehouse or BI system. However, these typically only operate on the structured data within an organization. They

seldom operate on fast-flowing volumes of data, let alone integrate operational data with data from social media, etc.

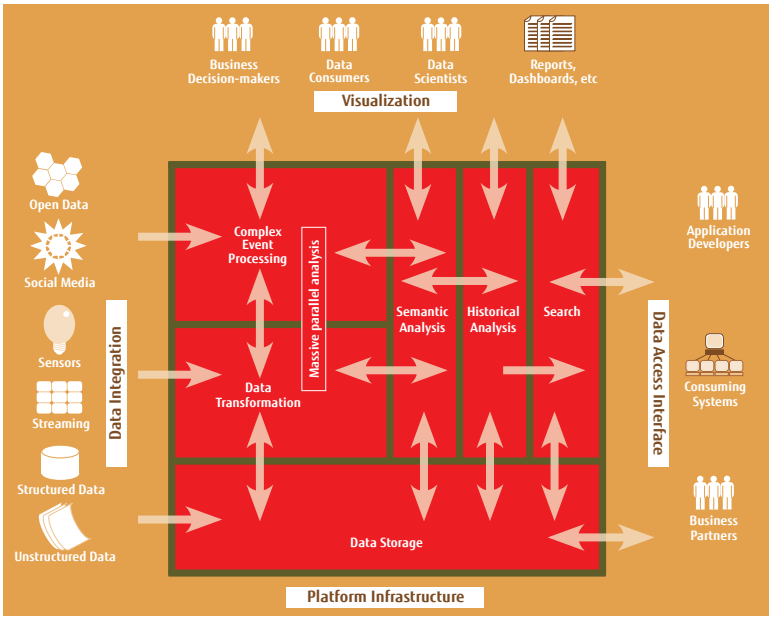
Isn't Big Data just search?

A common misconception is that a Big Data solution is simply a search tool. This view probably comes from the fact that Google is a pioneer and key player in the Big Data space. But a Big Data solution contains many more features than simply search. Going back to our Vs, search can deal with volume and variability, but it can't handle velocity, which reduces the value it can offer on its own to a business.

The IT bit: structure of a Big Data solution

CIOs are often concerned with what a Big Data solution should look like, how they can deliver one and the ways in which the business might use it. The diagram below gives a simple breakdown of how such a solution can be structured. The red box represents the solution itself. Outside on the left-hand side, are the various data sources that feed into the system – for example, open data (e.g. public or government-provided data, commercial data sites), social media (e.g. Twitter) or internal data sources (e.g. online transaction or analytical systems).

Structure of a Big Data Solution



The first function of the solution is"data integration"—connecting the system to these various data sources (using standard application interfaces and protocols. This data can then be transformed (i.e. changed into a different format for ease of storage and handling) via the "data transformation" function, or monitored for key triggers in the "complex event processing" function. This function looks at every piece of data, compares it to a set of rules and raises an alert when a match is found. Some complex event processing engines also allow time-based rules (e.g. "alert me if my product is mentioned on Twitter more than 10 times a second").

The data can then be processed and analyzed in near real time (using 'massively parallel analysis') and/or stored within the data storage function for later analysis. All stored data is available for both semantic analysis and traditional historical analysis (which simply means the data is not being analyzed in real time, not that the analysis techniques are old-fashioned).

Search is also a key part of the Big Data solution and allows users to access data in a variety of ways – from simple, Google-like, single-box searches to complex entry screens that allow users to specify detailed search criteria.

The data (be it streaming data, captured data or new data generated during analysis) can also be made available to internal or external parties who wish to use it. This could be on a free or fee basis, depending on who owns the data. Application developers, business partners or other systems consuming this information do so via the solution's data access interface, represented on the right-hand side of the diagram.

Finally, one of the key functions of the solution is data visualization – presenting information to business users in a form that is meaningful, relevant and easily understood. This could be textual (e.g. lists, extracts, etc) or graphical (ranging from simple charts and graphs to complex animated visualizations).

Furthermore, visualization should work effectively on any device, from a PC to a smartphone. This flexibility is especially important since there will be a variety of different users of the data (e.g. business decision-makers, data consumers and data scientists – represented across the top of the diagram), whose needs and access preferences will vary.

Privacy and Big Data

With the rise of Big Data and the growing ease of access to vast numbers of data records and repositories, personal data privacy is becoming ever harder to guarantee – even if an organization attempts to anonymize its data. Big Data solutions can integrate internal data sets with external data such as social media and local authority data. In doing so, they can make correlations that de-anonymize data, resulting in an increased – and to many, worrying – ability to build up detailed personal profiles of individuals.

Today organizations can use this information to filter new employees, monitor social media activity for breaches of corporate policy or intellectual property and so on. As the technical capability to leverage social media data increases, we may see an increase in the corporate use of this data to track the activities of individual employees. While this is less of a concern in countries such as the UK and Australia, where citizens' rights to privacy and fair employment are a major focus, such issues are not uniformly recognized by governments around the world. These concerns have led to a drive among privacy campaigners and EU data protection policy-makers towards a 'right to forget' model, where anyone can ask for all of their data to be removed from an organization's systems and be completely forgotten.

With the rise of Big Data personal data privacy is becoming ever-harder to guarantee – even if an organization attempts to anonymize the data.

Many of the concerns are borne out of stories such as people being turned down for a job because an employer found a comprising picture of them on Facebook, or companies sacking people for something they've posted in a private capacity on social media. But as today's younger generation becomes the management of tomorrow, it is likely to be more relaxed about both data privacy issues, and about what employees reveal about what they get up to in their own time. As a result, we're likely to see a move towards more of a "right to forgive" model – where individuals feel able to place more trust in organizations not to misuse their data, and those organizations will be less likely to do so.

The generation that has grown up with social media understands, for example, that if a photograph of someone inebriated at a party is posted on Facebook, it doesn't mean that person is an unworthy employee. Once such a more relaxed attitude to personal privacy becomes pervasive, data will become more accessible as people trust it won't be misinterpreted or misused by businesses and employers.

So when is the right time to adopt a Big Data solution? Just as has happened with mobile phones, our dependency on data will increase over time. This will come about as consumers' trust in the data grows in line with it becoming both

more resilient and more accessible. Given that Big Data is not actually new (as discussed earlier), late adopters may – surprisingly quickly – come to suffer the negative business consequences of not embracing it sooner.

The new KM model

For the past decade or so, businesses have often categorized data according to a traditional knowledge management (KM) model known as the DIKW hierarchy (data, information, knowledge, wisdom). In this model, each level is built from elements contained in the previous level. But in the context of Big Data, this needs to be extended to more accurately reflect organizations' need to gain business value from their (and others') data. A better model might be:

- **Integrated data** – data that is connected and reconnected to make it more valuable
- **Actionable information** – information put into the hands of those that can use it
- **Insightful knowledge** – knowledge that provides real insight (i.e. not just a stored document)
- **Real-time wisdom** – getting the answer now, not next week.

Of course, some organizations have put significant investment into traditional knowledge management systems and processes. So in regard to KM and its relationship with Big Data, it is worth noting the following:

1. KM is an enabler for Big Data, but not the goal
2. KM activities achieve better outcomes for structured data than for unstructured or semi-structured data
3. The principles of KM are still important but they need to be interpreted in new ways for the new types of data being processed
4. KM focuses much effort on storing all data, but that is not always the focus with Big Data, particularly when analyzing "in-flight" (transient) data.

In that sense Big Data has a librarian's focus. The archivist wants to store data but is less interested in making it accessible. The librarian is less interested in storing data as long as he or she has access to it and can provide the information that their clients need.

Hadoop: the elephant in the room

In a conversation about Big Data, it won't be long before someone (usually the techie in the room) mentions Hadoop®. Hadoop is an open source software product (or, more accurately, "software library framework") that is collaboratively produced and freely distributed by the Apache Foundation® – effectively, it is a developer's toolkit designed to simplify the building of Big Data solutions.

In technical terms, Hadoop enables distributed processing of large data sets across clusters of computers using a simple programming model. It can be extended with other components to create a Big Data solution. It is popular (as is most Apache Foundation software) because it works and it is free.

If this all sounds too good to be true, it's worth remembering that downloading the software is only the start if you want to build your own Big Data solution. In some cases, Hadoop projects distract businesses away from using Big Data to solve their business problems faster and instead tempt them onto the rocky road of developing their 'ideal Big Data solution' – which often ends up delivering nothing.

In short, Hadoop provides an important technical capability but it is merely one enabler for a complete Big Data solution (it incidently doesn't address the kind of semi-structured data challenge that a Linked Data solution is designed to handle). It is the capabilities beyond Hadoop that provide the real differentiator for Big Data solutions. Businesses should instead look out for cloud-based Big Data solutions which are scalable and offer "try-before-you-commit" features, not to mention an extensive range of built-in features.

Towards successful implementation

The key to successfully implementing a Big Data solution is to identify the benefits and pitfalls in advance and ensure it meets company objectives while also laying a foundation for broader business exploitation of the data in the future.

The following chapters will examine in more detail how to go about this.

2 What does Big Data Mean for the Business?

Every organization wants to make the best informed decisions it can, as quickly as it can. Indeed, gleaning insights from data in as close to real time as possible has been a key driving force behind the evolution of modern computing. For example, the very first computers – developed in the UK by World War II code-breakers – were designed to crack encrypted enemy communications fast enough to inform critical military and political decisions. Back then, any failure to do so could have potentially fatal consequences.

After the war, organizations began to realize that computing was also the key to securing business advantage – giving them the opportunity to work more quickly and efficiently than their competitors – and the IT industry was born.

Today IT has spread beyond the confines of the military, government and business, playing a part in almost every aspect of people's lives. The consumerization of IT has meant that most people in developed societies now own powerful, connected computing devices such as laptops, tablet PCs and smartphones. Combined with the growth of the Internet, this means an immense and exponentially growing amount of data is being generated – and is potentially available for analysis. This encompasses everything from highly structured information, such as government census data, to unstructured information, such as the stream of comments and conversations posted on social networks.

The challenge for organizations now is to achieve insightful results like those of wartime code-breakers.

The challenge for organizations now is to achieve insightful results like those of the wartime code-breakers, but in a very much more complicated world with many additional sources of information. In a nutshell, the Big Data concept is about bringing a wide variety of data sources to bear on an organization's challenges and asking the right type of questions to give relevant insights – in as near to real time as possible. This concept implies:

- **Data sets are large and complex,** consisting of different information types and multiple sources

Photograph: Corbis

- **Data is relevant** up to the second
- **Data collection** is automated and takes place in real time from people, systems, instruments or sensors
- **Analytical techniques** enable organizations to anticipate and respond dynamically to changing events and trends
- **Benefits may apply** to individuals, organizations and across society.

For different businesses and roles, this will mean different things. How someone assesses and balances factors such as value, cost, risk, reward and time when making decisions will vary according to their particular organizational and operational priorities. For example, sales and marketing professionals might focus on entering new markets, winning new customers, increasing brand awareness, boosting customer loyalty and predicting demand for a new product. Operations personnel, meanwhile, are more likely to concentrate on ensuring their organizations' processes are as optimal and efficient as possible, with a focus on measuring customer satisfaction.

The challenge is to find gold in the ever-growing mountain of information and act on it in near real time.

Finding gold in the data mountains

All these drivers for business success depend on information. But today the quantity of information available is not the issue. As the world has increasingly moved online, people's activities have left a trail of data that has grown into a mountain. The challenge is to find gold in that ever-growing mountain of information by understanding and acting on it in near real time. Companies already adept at doing so include the likes of Google, Amazon, Facebook and LinkedIn®.

But an organization doesn't need to be an Internet giant to benefit from Big Data—and successful solutions aren't always vast, expensive exercises that take months to implement. Even a simple mash-up (where someone thinks laterally, bringing together two or three different sources of information and applies them to a problem) can give a unique and fresh perspective on data that delivers clarity to a problem and allows an organization to take instant action.

For example, how do supermarkets ensure there's plenty of barbecue meat on the shelves whenever the weather is fine? They do it by combining and analyzing data they own and control (such as that from sales, loyalty card and logistics systems) with long range weather forecast data, as well as an understanding of suppliers' ability to meet any surges in demand for certain products. That's a fairly simple example, but more and more organizations are looking into their information hoard to see if it can be turned into a library for use today or in the future.

An explosion of information sources

The variety of available information sources is growing rapidly. As well as social media data, for example, there's telemetry data generated by cars, GPS data generated by smartphones, information collected on individuals and organizations by banks and governments—and much more data is coming on stream all the time.

The question is how all these sources can be applied in a way that is not only beneficial to a business but also allows people to trust in the integrity of the organizations and institutions collecting, handling, integrating, analyzing and acting on that data. In addition, businesses must understand the implications of relying on particular data sources, and what they would do if these became unavailable for any reason.

Big data in action

Today there are many examples of Big Data applications in action—both in a social and business context. From agriculture and transport to sustainability, health and leisure, Big Data has implications for just about every aspect of business and people's lives. For instance:

- **Financial services organizations** can use it to detect fraud and improve their debt position
- **Leisure companies** can examine data across their franchises of theme parks, hotels, restaurants, etc, looking for patterns that can help them enhance the customer experience
- **Disaster relief organizations** can aggregate data from both government and non-government sources (e.g. campaigning organizations, social media, etc) to visualize the situation and work out how best to deploy their resources.

Formula 1: Pole position for Big Data

Motor racing is at the leading edge of technological innovation. The margins between winning and losing can be measured in split seconds. Formula 1 (F1) teams would not be able to compete without real-time insight. They gain this through telemetry data supplied from hundreds of sensors on the cars. In a single race weekend, these sensors can generate a billion points of data.

The teams have invested millions of dollars in high-speed networks and vast amounts of computing resources. The car can be racing anywhere, but the data arrives instantly at a team's headquarters – which may be on the other side of the world. Strategic responses to situations in the race are generated in milliseconds, faster and more accurately than human team members would be capable of.

In the words of Geoff McGrath, managing director of the Applied Technologies division at F1 team McLaren, this gives the team access to "prescriptive intelligence" – the ability to anticipate the future and suggest winning moves. While this is primarily about driving competitive advantage, much of the data is also made available to the public (e.g. via television) and feeds back into the ecosystem of suppliers – driving innovation in the sport and, indeed, the entire automotive industry.

Ask the right questions

Organizations need to understand what real-time insight they can apply to make the most impact on their business in a particular situation. The key here is to ask the right questions, since these will determine both the data sources a business may wish to access and its choice of potential partner organizations (since pooling data on a given target market may make a proposition even more compelling).

The first question anyone in business should ask is what they would most like to know in order to have a greater positive impact on their business. They must then understand how to gather and process this information (i.e. what data sources are appropriate, what they need from these sources and what level of trust and reliance each offers), as well as working out what criteria they will apply to make decisions.

Start small and fine-tune later

The next stage is to run a pilot project and act on the insights it presents. Like most information system programs, with Big Data it pays to start small. After all, every journey begins with the first steps. Absolute accuracy isn't the goal – ballpark figures are good enough to gain useful real-time insights (for example, whether a trend is up or down). Processes can be fine-tuned as the journey progresses, through continual feedback and testing to hone the validity of the answers.

New opportunities and smart environments

The Big Data journey can lead to new markets, new opportunities and new ways of applying old ideas, products and technologies. One example is the widely discussed idea of "smart environments." For instance, smart cities might feature embedded sensors collecting data from buildings, cars, people and the environment.

By aggregating and analyzing this data in real time, many opportunities will emerge for new applications to improve everything from public health to traffic management and disaster response. Similarly, smart energy grids could link together new and existing energy generation technologies to maximize the use and sustainability of resources, among other benefits.

The Big Data journey can lead to new markets, opportunities and ways of applying old ideas, products and technologies.

A monumental impact

Real-time insight will have a huge impact on everyone's lives – as big as any historical technological breakthrough, including the advent of the PC and emergence of the Internet. By 2017, it's likely that:

- **Complex systems** which can sense and respond will be ubiquitous
- **Social objectives** will focus on the proactive rather than the reactive (for example, "maintaining wellbeing over providing treatment")
- **Everything** will speed up.

Summary and further considerations

- **Big Data can provide real-time insight** to answer an organisation's big questions—as long as that organization has a clear understanding of its goals and asks the right questions.

- **Big Data is about fully leveraging** the value that can be obtained from existing and new data sources, both within and outside the organization.

- **Thinking about data in new ways** can create new value. External or alternative perspectives on an organisation's data can open new pathways to success.

- **Combining different sources** of data that have not yet been analyzed together provides unique insights.

- **Where an organization does business** will affect the data sources it uses and how it can use the information—since data legislation varies around the world.

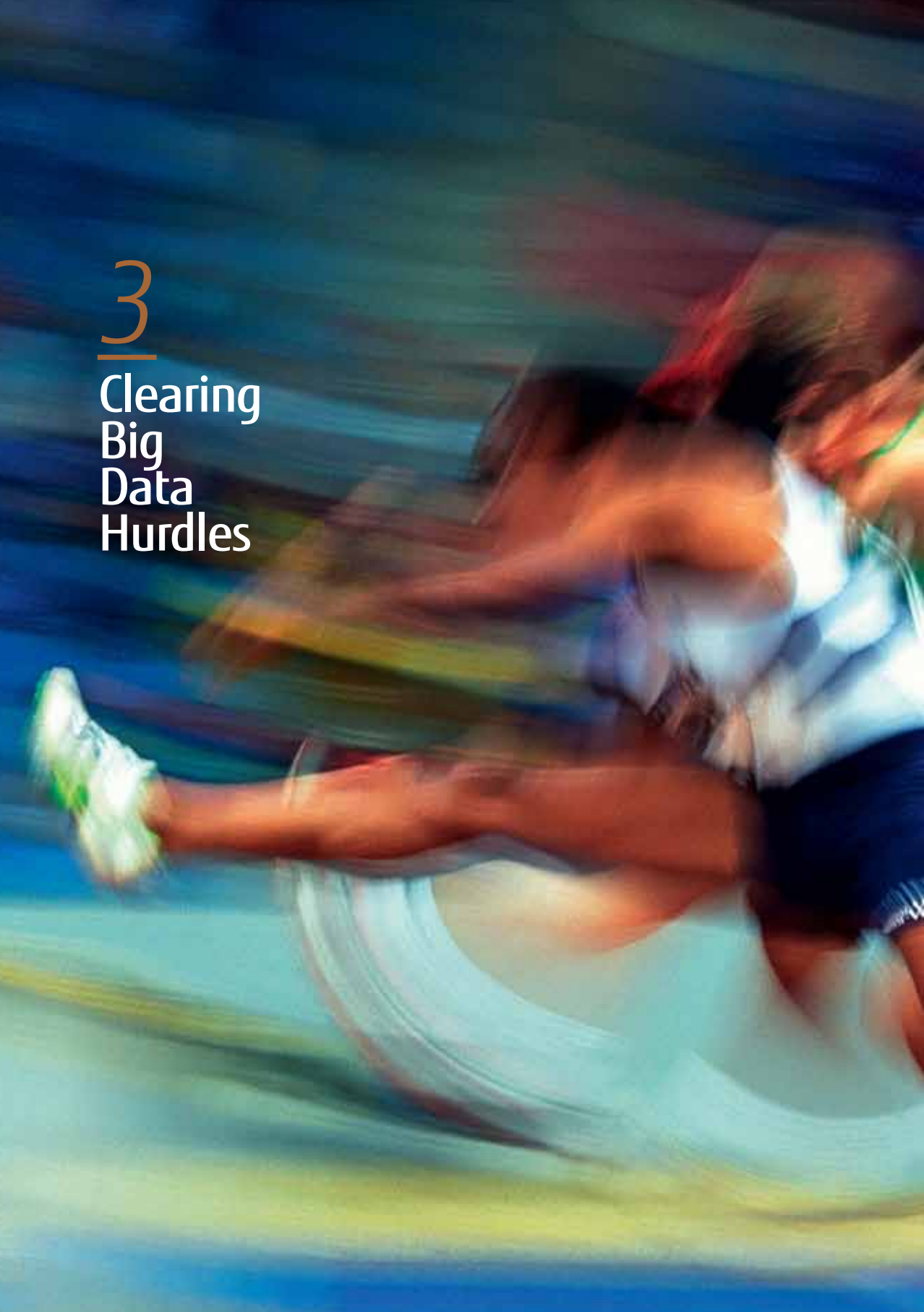
Alternative perspectives on an organization's data can open new pathways to success.

- **Structured data is no longer** the only data that can be analyzed. This leads to new opportunities and possibilities. Unstructured social media data is a gold mine, for example.

- **Creating Big Data programs** focused on the customer are good, but businesses shouldn't forget to track the competition as well.

70% of senior managers believe Big Data has the potential to drive competitive edge.

Survey of 200 senior managers
by Coleman Parkes Research for
Fujitsu UK & Ireland (2012)



3 Clearing Big Data Hurdles

To realize the advantages of Big Data, organizations must first tackle a number of obstacles that potentially stand in the way of their success. Broadly speaking, these can be grouped into business, technology and legislative challenges. This chapter explores these three areas in detail.

The business challenges

Big Data can uncover hidden insights that can generate previously impossible-to-realise value.

Questions before answers

Big Data holds the potential to offer answers to many business problems. But, depending on how data is queried (i.e. the algorithms used), the same problem can throw up very different answers. As the previous chapter notes, it is therefore vital that businesses spend time working out the right questions to ask of the data.

Know the unknowns

Businesses also need to be able to quantify the latent value within the data. There are many unknowns in Big Data analysis – it often uncovers hidden insights that can generate previously impossible-to-realize value. For example, Big Data can provide more acute market and competitive analyses that might signal the need for fundamental changes to a company's business model.

Don't trust all sources equally

The increasing use of third-party data sources is creating a requirement for platforms that can guarantee their data can be trusted. This is essential to enable the safe trading of information with appropriate checks and balances (just as with long-established credit reference systems used in the financial services sector). Businesses generally trust their internal data, but when dealing with external sources it is vital to understand the provenance and reputation of those sources. It is useful to consider data sources as sitting at different points on a continuum from 'trusted' (e.g. open government data) to 'untrusted' (e.g. social networks). The level

of trustworthiness can also (but not necessarily) equate to whether the source is internal or external, paid or unpaid, the age of the data and the size of the sample.

Data source dependency

If a business model relies on a particular external data source, it is important to consider what would happen if that source were no longer available, or if a previously free source started to levy access charges. For example, GPS sensor data may provide critical location data, but in the event of a war it might become unavailable in a certain region or its accuracy could be reduced. Another example is the use of (currently free) open data from government sources. A change of policy might lead to the introduction of charges for commercial use of certain sources.

Avoid analytical paralysis

Access to near real-time analytics can offer incredible advantages. But the sheer quantity of potential analyses that a business can conduct means there's a danger of "analytical paralysis" – generating such a wealth of information and insight
(some of it contradictory) that it's impossible to interpret. Organizations need to ensure they are sufficiently informed to react without becoming overwhelmed.

Manage the information lifecycle

Senior managers need to ensure Big Data initiatives are not undermined by employee resistance to change.

While some of the concerns around handling information at different stages in its lifecycle are technical (see 'Data lifecycle management' under 'Technical challenges', below), there are also business issues to consider. For example, how should a record containing personal information be processed and what needs to be done when that record expires? Businesses need to decide, for instance, if such records are stored in an anonymized format or removed after a time.

Overcome employee resistance

In common with many business change projects, senior managers need to ensure Big Data initiatives are not undermined by employee resistance to change. For example, one utility company's Big Data project identified a large number of customers who weren't on the billing system despite the fact they'd received services for months (and, in some cases, years). While this should have been an opportunity to increase revenues, the news was met with a combination of disbelief, messenger-shooting and protective behavior as some employees believed the discovery of the error had cast them in a poor light. Such resistance might have been avoided had the company paid more attention in advance to pre-empting staff concerns, assuaging their fears and communicating the positive aims of the project. Another potential cause of employee resistance is

the fear that advanced predictive analytics undermines the role of skilled teams in areas such as forecasting, marketing and risk profiling. If their fears aren't comprehensively addressed at the outset, such employees may attempt to discredit the Big Data initiative in its early stages – and could potentially derail it.

Technical challenges

Many of Big Data's technical challenges also apply to data it general. However, Big Data makes some of these more complex, as well as creating several fresh issues. Chapter 1 outlined the technical elements of a Big Data solution (see "The IT bit", page 11). Below, we examine in more detail some of the challenges and considerations involved in designing, implementing and running these elements.

Data integration

Since data is a key asset, it is increasingly important to have a clear understanding of how to ingest, understand and share that data in standard formats in order that business leaders can make better-informed decisions. Even seemingly trivial data formatting issues can cause confusion. For example, some countries use a comma to express a decimal place, while others use commas to separate thousands, millions, etc – a potential cause of error when integrating numerical data from different sources. Similarly, although the format may be the same across different name and address records, the importance of "first name" and "family name" may be reversed in certain cultures, leading to the data being incorrectly integrated.

Organizations might also need to decide if textual data is to be handled in its native language or translated. Translation introduces considerable complexity – for example, the need to handle multiple character sets and alphabets.

Further integration challenges arise when a business attempts to transfer external data to its system. Whether this is migrated as a batch or streamed, the infrastructure must be able to keep up with the speed or size of the incoming data. The selected technology therefore has to be adequately scalable, and the IT organization must be able to estimate capacity requirements effectively. Another important consideration is the stability of the system's connectors
(the points where it interfaces with and "talks" to the systems supplying external data). Companies such as Twitter and Facebook regularly make changes to their application programming interfaces (APIs) which may not necessarily be published in advance. This can result in the need to make changes quickly to ensure the data can still be accessed.

Data transformation

Another challenge is data transformation – the need to define rules for handling data. For example, it may be straightforward to transform data between two systems where one contains the fields "given name" and "family name" and the other has an additional field for "middle initial" – but transformation rules will be more complex when, say, one system records the whole name in a single field.

Organizations also need to consider which data source is primary (i.e. the correct, "master" source) when records conflict, or whether to maintain multiple records. Handling duplicate records from disparate systems also requires a focus on data quality (see also "Complex event processing" and "Data integrity" below).

Complex event processing

Complex event processing (CEP) effectively means (near) real-time analytics. Matches are triggered from data based on either business or data management rules. For example, a rule might look for people with similar addresses in different types of data. But it is important to consider precisely how similar two records are before accepting a match. For example, is there only a spelling difference in the name or is there a different house number in the address line? There may well be two Tom Joneses living in the same street in Pontypridd – but Tom Jones and Thomas Jones at the same address are probably the same person.

Organizations need to consider which data source is primary – their master data.

IT professionals are used to storing data and running queries against it, but CEP stores queries that are processed as data passes through the system. This means rules can contain time-based elements, which are more complicated to define. For example, a rule that says "if more than 2% of all shares drop by 20% in less than 30 seconds, shut down the stock market" may sound reasonable, but the trigger parameters need to be thought through very carefully. What if it takes 31 seconds for the drop to occur? Or if 1% of shares drop by 40%? The impact is similar, but the rule will not be triggered.

Semantic analysis

Semantic analysis is a way of extracting meaning from unstructured data. Used effectively, it can uncover people's sentiments towards, for example, organizations and products, as well as unearthing trends, untapped customer needs, etc. However, it is important to be aware of its limitations. For example, computers are not yet very good at understanding sarcasm or irony, and human intervention might be required to create an initial schema and validate the data analysis.

Historical analysis

Historical analysis could be concerned with data from any point in the past. That is not necessarily last week or last month—it could equally be data from 10 seconds ago. While IT professionals may be familiar with such an application its meaning can sometimes be misinterpreted by non-technical personnel encountering it.

Search

As Chapter 1 outlined, search is not always as simple as typing a word or phrase into a single text input box. Searching unstructured data might return a large number of irrelevant or unrelated results. Sometimes, users need to conduct more complicated searches containing multiple options and fields. IT organizations need to ensure their solution provides the right type and variety of search interfaces to meet the businesses' differing needs.

Another consideration is how search results are presented. For example, the data required by a particular search could be contained in a single record (e.g. a specific customer), in a ranked listing of records (e.g. articles listed according to their relevance to a particular topic), or in an unranked set of records (e.g. products discontinued in the past 12 months). This means IT professionals need to consider the order and format in which results are returned from particular types of searches. And once the system starts to make inferences from data, there must also be a way to determine the value and accuracy of its choices.

Data storage

As data volumes increase storage systems are becoming ever more critical. Big Data requires reliable, fast-access storage. This will hasten the demise of older technologies such as magnetic tape, but it also has implications for the management of storage systems. Internal IT may increasingly need to take a similar, commodity-based approach to storage as third-party cloud storage suppliers do today—i.e. removing (rather than replacing) individual failed components until they need to refresh the entire infrastructure. There are also challenges around how to store the data—for example, whether in a structured database or within an unstructured (NoSQL) system—or how to integrate multiple data sources without over-complicating the solution.

Data integrity

For any analysis to be truly meaningful it is important that the data being analyzed is as accurate, complete and up to date as possible. Erroneous data will produce misleading results and potentially incorrect insights. Since data is increasingly used

to make business-critical decisions, consumers of data services need to have confidence in the integrity of the information those services are providing.

Data lifecycle management

In order to manage the lifecycle of any data, IT organizations need to understand what that data is and its purpose. But the potentially vast number of records involved with Big Data, and the speed at which the data changes, can give rise to the need for a new approach to data management. It may not be possible to capture all of the data. Instead, the system might take samples from a stream of data. If so, IT needs to ensure the sample includes the required data, or that the sampled data is sufficiently representative to provide the required level of insight.

Data replication

Generally, data is stored in multiple locations in case one copy becomes corrupted or unavailable. This is known as data replication. The volumes involved in a Big Data solution raise questions about the scalability of such an approach. However, Big Data technologies may take alternative approaches. For example, Big Data frameworks such as Hadoop (see Chapter 1, page 15) are inherently resilient, which may mean it is not necessary to introduce another layer of replication.

The vast number of records involved in Big Data, and the speed at which the data changes, can give rise to the need for a new approach to data management.

Data migration

When moving data in and out of a Big Data system, or migrating from one platform to another, organizations should consider the impact that the size of the data may have. Not only does the 'extract, transform and load' process need to be able to deal with data in a variety of formats, but the volumes of data will often mean that it is not possible to operate on the data during a migration – or at the very least there needs to be a system to understand what is currently available or unavailable.

Visualization

While it is important to present data in a visually meaningful form, it is equally important to ensure presentation does not undermine the effectiveness of the system. Organizations need to consider the most appropriate way to display the results of Big Data analytics so that the data does not mislead. For example, a graph might look good rendered in three dimensions, but in some cases a simpler representation may make the meaning of the data stand out more clearly. In addition, IT should take into account the impact of visualizations on the various target devices, on network bandwidth and on data storage systems.

Data access

The final technical challenge relates to controlling who can access the data, what they can access, and when. Data security and access control is vital in order to ensure data is protected. Access controls should be fine-grained, allowing organizations not only to limit access, but also to limit knowledge of its existence.

One issue raised by advanced analytics is the possibility that the aggregation of different data sources reveals information that would otherwise be deemed a security or privacy risk. Enterprises therefore need to pay attention to the classification of data. This should be designed to ensure that data is not locked away unnecessarily (limiting its potential to provide actionable insights) but equally that it doesn't present a security or privacy risk to any individual or company.

In addition, open-ended queries (searching with wildcards) may have performance implications—or cause concerns from a data extraction perspective. For example, organizations may invest significant resources in consolidating data to give them a 'single view' of customers or other key competitive information which could become a target for hacking, theft or sabotage.

If a business provides an external interface to its data by means of an API, this needs to be maintained, echoing the challenge referred to above (under Data integration), but this time as a provider of data rather than as a consumer. Finally, application developers need to be aware that the move from serial to parallel processing may affect the way that applications are designed and implemented.

Legislative challenges

From a legal standpoint, many of the challenges relate to data ownership, privacy and intellectual property rights. Over time, we can expect a societal shift in attitudes towards data handling, but currently organizations have to take into account that:

- **Depending on where data originates,** there may be issues around ownership, intellectual property and licensing—all of which will need to be resolved before data can be used
- **As data is aggregated** even anonymized data may contain identifiable information, which may place a business in breach of data protection regulations
- **With data being stored** on various systems across the globe, there may be issues of data sovereignty and residency (i.e. questions over which country or countries can claim legal jurisdiction over particular data) depending on the type of data being stored and processed.

4 Adoption Approaches

There has never been a better time for organizations to begin their Big Data journey. The Fuitsu 2012 Global Megatrends Survey interviewed CIOs across multiple regions and sectors to uncover the technological developments shaping the future of global business.

The research confirmed that four "megatrends" currently dominate the IT landscape:

- **Cloud computing**
- **IT consumerization**
- **Social and collaboration**
- **Big Data.**

Although all of these trends have been growing for some time, they are now beginning to converge and accelerate. For example, powerful smart mobile devices in the hands of consumers feed the uptake of social media, creating reams of potentially valuable unstructured information. Cloud computing solutions are making the analysis of this information economically viable, and this in turn is fuelling business demand for Big Data solutions.

Big Data, cloud, consumerization , social and collaboration: these trends, which have been growing for some time, are now beginning to converge and accelerate.

There are early adoption case studies in which retail companies, for example, are using Big Data solutions to bring together weather information and logistics data to deliver the right products, to the right place, just in time. However, the most obvious and predominant use of Big Data solutions is in the area of customer analysis and product cross-selling.

External and open data

Another major trend driving business interest in Big Data is the growing availability of external and open data sets. As previous chapters have noted, businesses will increasingly need to consume and aggregate data from outside their own organizations. That will not only extend to unstructured data from social networks, sensors, etc, but also to public data sets and private databases offering

Photograph: Alamy

Free sources
There are many free sources of data to be exploited, not all of them within your organization.

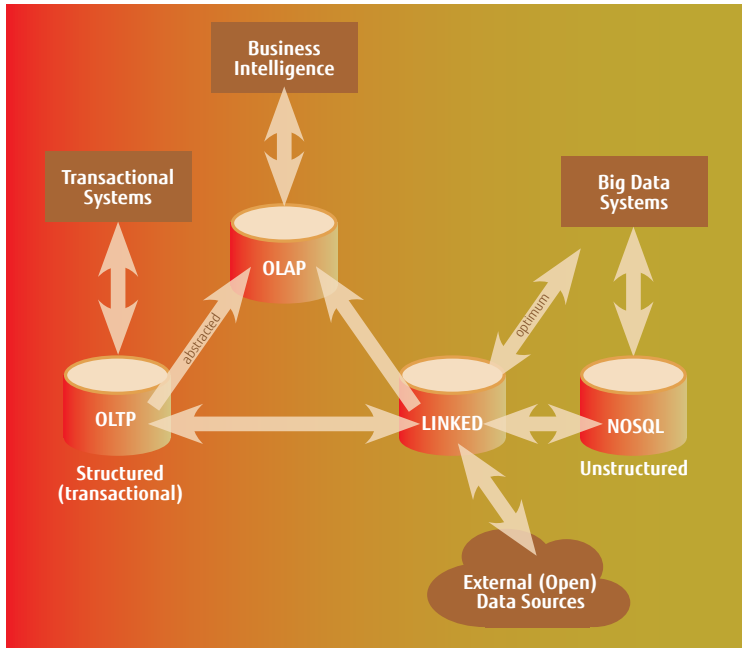
access on a subscription basis (e.g. credit records), which could be structured or unstructured. In addition to established government data sources (such as the USA's data.gov, the UK's data.gov.uk and Ireland's opendata.ie), the European Commission has launched an open data strategy to drive the release of public authority data sets. The theory is that giving away data already created using public money will provide an economic stimulus as companies build new products and services around the data.

Other public initiatives across the world are similarly attempting to free up public data for commercial use. In the UK, for example, the government is supporting the Open Data Institute—a body set up to help businesses exploit open data, spearheaded by respected technologists Sir Tim Berners-Lee and Professor Nigel Shadbolt.

From why to what

All these trends help to answer the question "why now?", but the next question for many businesses is "what next?". How should a business interested in adopting a Big Data solution approach the task? The last chapter outlined some of the

Linked Data: a new model for connecting and exploiting data



challenges organizations can face along the road, and questions they need to ask. This chapter will look in more detail at how organizations can overcome some of those challenges and find the most appropriate and successful adoption approach for their particular business.

Lose the silos by linking data

Many proposed approaches to the management of Big Data could potentially result in organizations creating new "silos" (separate, self-contained islands of information) for transactional systems, business intelligence (data warehouse) systems and unstructured Big Data solutions. However, this need not be the case. By linking data (instead of creating copies of the same data in separate databases), organizations can index and feed information from a variety of sources in close to real time (see diagram opposite).

Identify the problem

Before adopting a Big Data solution, an organization needs to know the problem it wants to solve. For example, it might want to understand the relationship between customers' buying patterns and their online influence (on social networks, etc). Equally, it also needs to understand how best to represent the results (in a table, graphical format, textually, etc).

Many proposed approaches to the management of Big Data could result in organizations creating new information 'silos'.

Data lifecycles

Organizations already have to determine the length of time for which information should be retained (most notably for reasons of legal and regulatory compliance). But as they manage increasing volumes of data, IT departments have a second lifecycle to consider – that of the systems in which the data is held.

For example, social media data may be combined with internal customer data to uncover new insights. This new data could be discarded once it has served its immediate use, or it could be retained for a period. If an organization decides to retain the data, it needs to consider not only the regulatory implications, but also the technical question of how and where to store that data:

- **Should it remain in an unstructured,** Big Data solution (e.g. a NoSQL database) or be moved to a data warehouse?
- **Should it remain online** or be made accessible from an archive on an as-needed basis?
- **Should it be retained in full,** aggregated with other data or anonymized for more general use?

There are no definitive answers to these questions. They will vary depending on the contents of the data, an organization's policies and any legal or regulatory restrictions. However, these considerations underline the need for robust data lifecycles when adopting Big Data solutions.

In effect, data is created (arrives), is maintained (exists), and then is deleted (disappears) at some point in the future. That data needs to be managed in a way that ensures it is kept for the right length of time and no longer. But Big Data's potential to find new insights and generate value from old data prompts another question – shouldn't organizations be keeping it forever, as long as they can do so securely, cost-effectively and legally?

Choosing the right tools

When selecting tools for Big Data analysis, organizations face a number of considerations:

- **Where will the data be processed?** Using locally-hosted software on a dedicated appliance, or in the cloud? Early Big Data usage is likely to focus on business analytics. Enterprises will need to turn capacity on and off at particular times, which will result in them favoring the cloud over private solutions.
- **From where does the data originate** and how will it be transported? It's often easier to move the application than it is to move the data (e.g., for large data sets already hosted with a particular cloud provider). If the data is updated rapidly then the application needs to be close to that data in order to maximize the speed of response.
- **How clean is the data?** The hotch-potch nature of Big Data means that it needs a lot of tidying up – which costs time and money. It can be more effective to use a data marketplace. Although the quality of the data provided via such marketplaces can vary, there will generally be a mechanism for user feedback/ratings which can give a good indication of the quality of the different services on offer.
- **What is the organization's culture?** Does it have teams with the necessary skills to analyze the data? Big Data analysis requires people who can think laterally, understand complicated mathematical formula (algorithms) and focus on business value. Data science teams need a combination of technical expertise, curiosity, creative flair and the ability to communicate their insights effectively.
- **What does the organization want to do with the data?** As well as honing the choice of tools, having an idea about the desired outcomes of the analysis can also help a business to identify relevant patterns or find clues in the data.

Key for prioritising Big Data initiatives*

- High value
- Medium value
- Low value
- No value

* Readers are invited to add their own shading to the diagram

Identifying and acting on Big Data priorities

The chart above has been designed to help readers of this book identify the key Big Data priorities for their businesses. At the heart of business is the drive for increased profit, represented here in the centre of the target. Working outwards, businesses can either increase profit by increasing revenue or reducing costs. Both of these methods can be achieved through either direct or indirect actions, but by combining the two we move outwards towards the appropriate segment.

The outer circle shows the various actions a business can take. The left hemisphere contains revenue-increasing actions, while the right side contains cost-reducing actions. The diagram also splits horizontally to show direct actions (in the top half) and indirect actions (bottom half). From this it is easy to see the possible actions a business can take to increase revenues or reduce costs, either directly or indirectly. These are also listed below with examples:

Direct actions to increase revenues:

- **Develop new products** or services (e.g. to address new opportunities)
- **Increase existing customer revenue** (e.g. by raising prices)

- **Acquire new customers** (e.g. by running a sales campaign).

Direct actions to reduce costs:
- **Improving productivity** (e.g. automating processes)
- **Displacing costs** (e.g. outsourcing non-core functions)
- **Reducing capital requirements** (e.g. moving to an operational expenditure model).

Indirect actions to increase revenues:
- **Increasing brand awareness** (e.g. by running a marketing campaign)
- **Increasing customer loyalty** (e.g. by improving account management)
- **Increasing customer satisfaction** (e.g. by improving customer service).

Indirect actions to reduce costs:
- **Reducing fulfilment errors** (e.g. putting in place additional checks in the dispatch process)
- **Decreasing time to market** (e.g. by shortening the product development lifecycle)
- **Reducing customer support dependencies** (e.g. by introducing a self-service element to the support process).

By shading the outer gray segments of the diagram on the previous page – green (for high value), yellow (for medium value), blue (for low value) or leaving them blank (for no value) – it would be straightforward for organizations to identify their business priorities. They can then use the table on the pages overleaf to identify the actions that need to be taken.

Ensuring success for a Big Data project

In common with any business change initiative, a Big Data project needs to be business-led and (ideally) executive-sponsored – it will never work as an isolated IT project. Even more importantly, Big Data is a complex area that requires a wide range of skills – it spans the whole organization and the entire executive team needs to work together (not just the CIO).

In addition, there is a dearth of data scientists, and it may be necessary to fill gaps with cross training. The next two chapters of this book look at the changing role of the executive team and the rise of the data scientist.

Over 70% of organizations rank better marketing and responding to changing needs of the customer/citizen as the areas where Big Data could have most impact.

Survey of 200 senior managers
by Coleman Parkes Research for
Fujitsu UK & Ireland (2012)

Big Data action plan

Problem	Solution	Value	Next step
Acquire new customers	Acquire and integrate marketing data, using a Big Data solution to better identify potential prospects.	Reduced cost of acquisition and improved conversion rate.	Identify relevant external/ internal data sources and implement the historical analytics part of a Big Data solution, ensuring that operational sales data is fed into the algorithms.
Increase existing customer revenue	Use a Big Data solution to monitor customer shopping habits, online or in-store, and alert employees to respond to faltering sales. Also use a Big Data solution to better understand customers and their desires.	Increased sales.	Implement the alerting part of a Big Data solution.
Develop new products or services	Use a Big Data solution to interact with your customers in a social context, as well as more traditional channels, to better understand their needs, wants and desires.	Quicker and more accurate understanding of your customers' needs and a more intimate relationship with existing customers.	Integrate your Big Data solution with social media, etc and use the Big Data solution to analyze discussion regarding your products.
Increase brand awareness	Use a Big Data solution to identify your 'market penetration' heat map, across multiple geographies, demographics, etc.	A clearer understanding of where your brand is strong or weak, allowing you to target specific sectors, with near real-time feedback of campaign effectiveness.	Implement a Big Data solution as a pilot on one dimension, e.g. brand visibility within region X, once the concept is proven and confidence gained extend to other regions, demographics or brand awareness questions.
Increase customer loyalty	Use a Big Data solution to better understand your customers, preferred engagement approach. Use this information to both personalize the service for the customer and to improve the overall process, etc.	Customer feels the relationship is more personal and therefore better than the one they have with the competition. Customers also more likely to engage in an improvement dialogue.	Integrate a Big Data solution into your customer relationship systems, social media and other customer related data sources. Use semantic analysis tools to understand sentiment, ensure a continual feedback loop is employed to improve the analysis process.

Problem	Solution	Value	Next step
Increase customer satisfaction	Use a Big Data solution to monitor and alert on customer complaints, etc.	Quicker identification of customer issues and comments, analyzed to identify trends, etc.	Implement a Big Data solution, specifically the alerting element.
Reduce fulfilment errors	Use a Big Data solution to connect consumer comments re. your products on social media to their purchases, identifying product problems.	Identify consumer problems quicker. Address bad word-of-mouth marketing quickly.	Integrate Big Data solution to social media and your fulfilment systems. Use CEP to identify connections and raise alerts.
Decrease time to market	Use a Big Data solution to gather feedback from customers by analyzing the product usage.	Quick determination of what is good/bad in a product and therefore what needs improvement. Leads to a reduction in user testing of products.	Develop products that provide Big Data. Install Big Data solution to monitor usage, etc.
Reduce customer support dependencies	Use a Big Data solution to monitor and analyze customer support requests, responses, votes and industry blogs, etc.	Receive and publish better support responses.	Implement a Big Data solution that monitors usage and votes on support sites, etc.
Reduce capital requirements	Use a Big Data solution to better understand your value chain, including the timings around customer demands and provision of supplier products. Additionally exploit a Big Data solution to better predict and manage human resources.	Less capital depreciation of bought components. Less wasted capital through no sales. Improved cash flow across the value chain.	Integrate a Big Data solution into your logistics/manufacturing/ CRM systems, as well as external systems, e.g. weather forecasts. Exploit real-time analytics to manage the value chain more efficiently.
Displace costs	Use a Big Data solution to monitor and alert on spending trends.	Identify cost trends as they happen, rather than several months later on.	Implement a Big Data solution and integrate to the finance and procurement systems.
Improve productivity	Use a Big Data solution to monitor processes, including the use of RFI tags, etc.	Rapid identification of process weaknesses, which can be optimised/improved.	Implement a Big Data solution and start by selecting one key process to pilot. Then expand to other key processes.

5 Changing Role of the Executive Team

Fujitsu CIO research has found that some sectors (e.g. utilities, telecoms and manufacturing) typically view IT as business-enabling, but others (notably, retail finance and government) see IT largely as a cost centre. So the attractiveness of exploiting Big Data for many organizations will come down to culturally how IT is used to generate competitive advantage.

Big Data is not an IT-owned decision it's a business-owned decision, i.e. what does an organization need to know to be more effective in executing its strategy and business model to maintain competitive advantage in an ever-more complex and competitive business landscape?

"To out-compute is to out-compete." This oft-quoted maxim – first coined almost a decade ago by Bob Bishop, ex-CEO of former market-leading workstation supplier Silicon Graphics – is no less true today than it ever was. The next generation of leading organizations will achieve their success by making the best use of IT to exploit an ever-growing mountain of Big Data. Executive teams that don't understand what is possible with the technology will not be able to lead their businesses effectively.

The next generation of leading organizations will achieve their success by making the best use of IT to exploit an ever-growing mountain of Big Data.

In some areas, the nature of key leadership roles has changed dramatically in modern times. For example, 30 years ago, few people would have imagined that statisticians and behavioral scientists would shape political parties' strategies, or that elections would be won through targeted online campaigns rather than blanket mass-media coverage.

A successful executive team is more likely to take a calculated risk on a passionate individual with a potentially winning idea. For example, at the turn of the century, the board of Oakland Athletics Baseball Club in the US supported its general manager Billy Beane's (successful) plan to improve the team's performance through the use of computer analysis – at the time a completely off-the-wall idea. Today, Beane is credited with having revolutionized the sport

Photograph: Getty Images

and his story was recounted in the 2003 book *Moneyball* by Michael Lewis (made into a movie starring Brad Pitt in 2011).

These examples illustrate how today and tomorrow's executives will need to ensure they are receiving a timely and accurate flow of information, which they can use to make better decisions and give their organizations an edge over the competition. Only time will tell if they can make the transition successfully, but those seeking a head start should certainly be sure they are aware of the value that Big Data initiatives can potentially bring to their businesses.

And it will take strong leadership, particularly since organizations are likely to experience various forms of resistance to Big Data. Executive teams should pinpoint initial projects to sponsor and look for early success stories they can exploit to ensure their organizations remain excited and positive about the opportunities ahead.

Executives will need to ensure they are receiving a timely and accurate flow of information, which they can use to make better decisions and give their organizations an edge over the competition.

55% of companies are already changing their business processes to make best use of Big Data.

Survey of 200 senior managers
by Coleman Parkes Research for
Fujitsu UK & Ireland (2012)

6 Rise of the Data Scientist

In 2006, executives at international customer data analysis company Dunnhumby began using the phrase: "Data is the new oil". They were right in more than one way. Like oil, data is often buried in extremely hard-to-reach places. Similarly, in its raw form it is of little use – it needs to be refined and may need additional ingredients to realise its value. In addition, its scale, flow and variety increase the challenges associated with extracting and exploiting it.

Just as the oil industry requires people from diverse disciplines to work together to maximise returns, so does Big Data. For a Big Data venture to succeed, a business needs three key things:

1. **IT capability** to capture, store, process and publish the data
2. **Business knowledge** to define and articulate the desired business outcome, sponsor the initiative, provide business insight and ensure resources are effectively deployed
3. **Data scientists**.

Data is the new oil: it needs discovery, extraction and refining to realize its value.

The data scientist's role is to couple an understanding of the business's challenges with all the potential data that can be brought to bear on those challenges, working out how best to use, refine and process that data. This often requires a combination of complex mathematics and creativity, as well as good communications and visualisation skills to ensure the right information is presented to the right people in the right way – i.e. a way that allows them to make better decisions and optimize business outcomes.

Unfortunately, such people are currently rare – and command very high salaries. Today, they are often concentrated in specialist areas such as financial trading (where they are known as quantitative analysts or 'quants') and simulation modeling for aircraft manufacturing. However, organizations should not despair. Supported by the right software tools, work that would once have required a doctorate in mathematics can increasingly be done by numerate and creative non-specialists. This will lead to the democratization of data science.

Photograph: Corbis

7

The Future of Big Data

Big Data is an emerging discipline, therefore most of what is discussed in this book is about the future. But what developments can organizations expect beyond the short term, and what implications are these likely to have on their business?

Big data for all

Currently Big Data is seen predominantly as a business tool. Increasingly, though, consumers will also have access to powerful Big Data applications. In a sense, they already do (e.g. Google, social media search tools, etc). But as the number of public data sources grows and processing power becomes ever faster and cheaper, increasingly easy-to-use tools will emerge that put the power of Big Data analysis into everyone's hands.

Increasingly tools will emerge that put the power of Big Data analysis into everyone's hands – consumers, business, government.

Data evolution

It is also certain that the amount of data stored will continue to grow at an astounding rate. This inevitably means Big Data applications and their underlying infrastructure will need to keep pace.

Looking out two to three years, it is clear that data standards will mature, driving up accessibility. Work on the 'semantic web' – a collaborative project to define common data formats – is likely to accelerate alongside the growth in demand among organizations and individuals to be able access disparate sources of data. More governments will initiate open data projects, further boosting the variety and value of available data sources.

Linked Data databases will become more popular and could potentially push traditional relational databases to one side due to their increased speed and flexibility. This means businesses will be able to develop and evolve applications at a much faster rate.

Photograph: Corbis

Data security will always be a concern, and in the future, data will be protected at a much more granular level than it is today. For example, whereas users may currently be denied access to an entire database because it contains a small number of sensitive records, in future access to particular records (or records conforming to particular criteria) could be blocked for particular users.

As data increasingly becomes viewed as a valuable commodity, it will be freely traded, manipulated, added to and re-sold. This will fuel the growth of data marketplaces – websites where sellers can offer their data wares simply and effectively and buyers will be able to review and select from a comprehensive range of sources.

Dawn of the databots

As volumes of stored data continue to grow exponentially and data becomes more openly accessible, "databots" will increasingly crawl organizations' linked data, unearthing new patterns and relationships in that data over time. These databots will initially be small applications or programs that follow simple rules, but as time moves on they will become more sophisticated, self-learning entities. Potentially, they will be able to ascertain the complexity of a query, call on the help of however many databots are needed to answer it and draw processing power from the cloud as needed. The artificial intelligence programs they employ will continue to grow more effective due to the fact that they can operate over time and learn from ever larger data sets.

As data increasingly becomes viewed as a valuable commodity, it will be freely traded, manipulated, added to and re-sold.

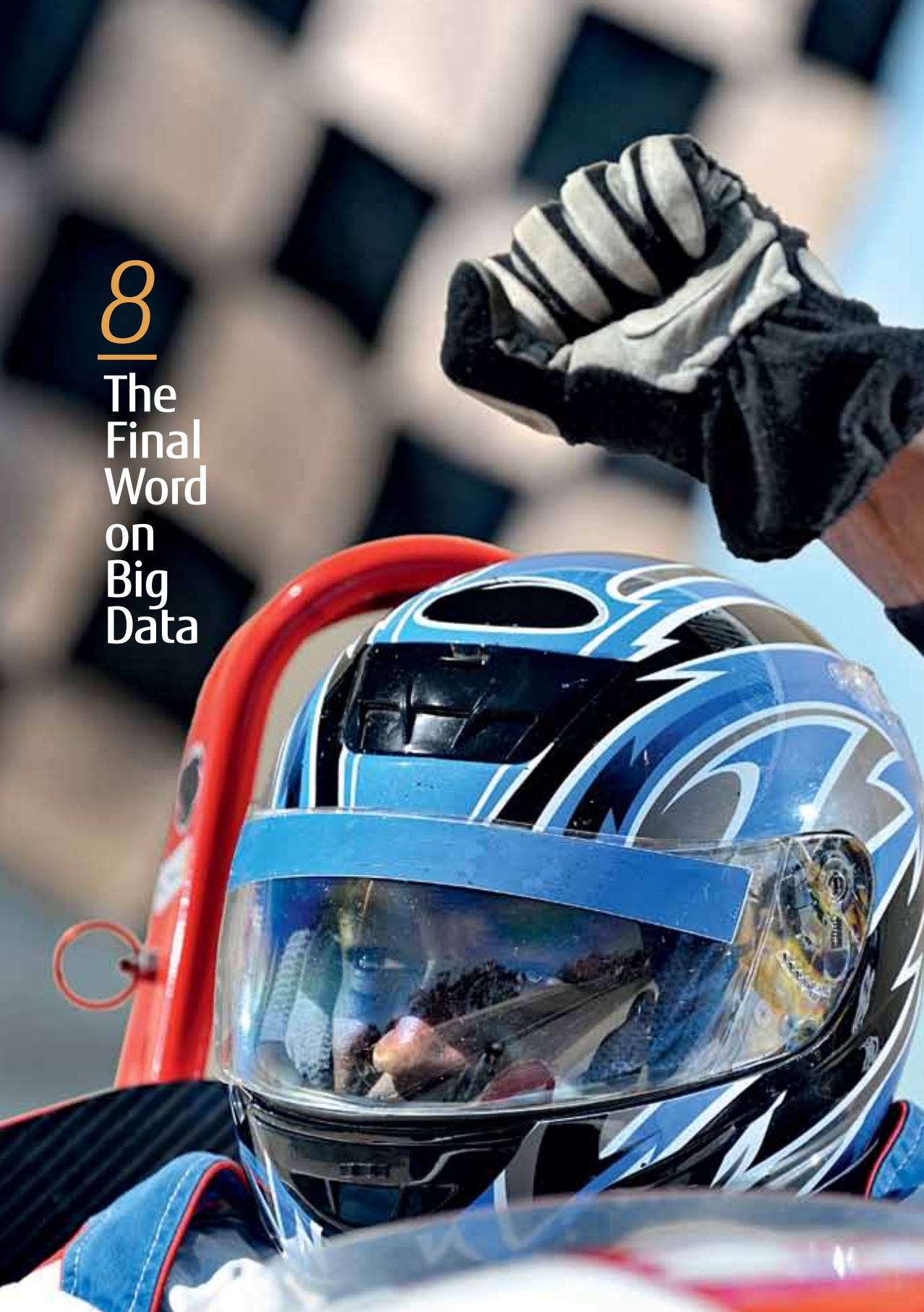
The power of linked data unleashed

Data will become increasingly connected, with the potential to unleash huge power. The more of it that is connected, the more powerful it will become – creating new job opportunities (and roles) and giving people the ability to better understand their work and make more informed decisions. And because Linked Data is "machine readable", increasingly less human input will be required to make sense of it.

In short, Linked Data will be at the heart of the world's computing. People and organizations will no longer have to worry about connecting devices or accessing documents. Instead, they will be able to focus on the information they really need to make decisions.

49% of senior managers believe that, by 2015, Big Data will have fundamentally changed how businesses operate.

Survey of 200 senior managers
by Coleman Parkes Research for
Fujitsu UK & Ireland (2012)



08 The Final Word on Big Data

Fujitsu is driven by a vision which the company refers to as "the human-centric intelligent society." This is focused on building a prosperous society through the use of information and communication technologies. Big Data plays a critical part in this vision, with business and wider society able to make use of the new opportunities both to provide real-time insight and to enable complex simulations and modeling.

Such solutions need to be delivered on a planetary scale. The ability to manage, process and act on Big Data will require a highly integrated global capability – whether that is to deliver services to consumers across every geography or to carry out large-scale simulations of oceans, weather systems, new drugs, or, indeed, an organization's business and manufacturing processes.

The vision is about linking people, products and services with information. In this context, the digital world offers mechanisms that allow individuals and organizations to share and collaborate on a planetary scale.

The vision is about linking people, products and services with information, allowing individuals and organizations to share and collaborate on a planetary scale.

In this fast-moving, connected world, intuition, experience and training will not be enough to give businesses the insight they need. They need to apply scientific and data analysis to their questions – and find the answers in the time frame demanded so they can make the right decisions. And that all requires scalable, global solutions.

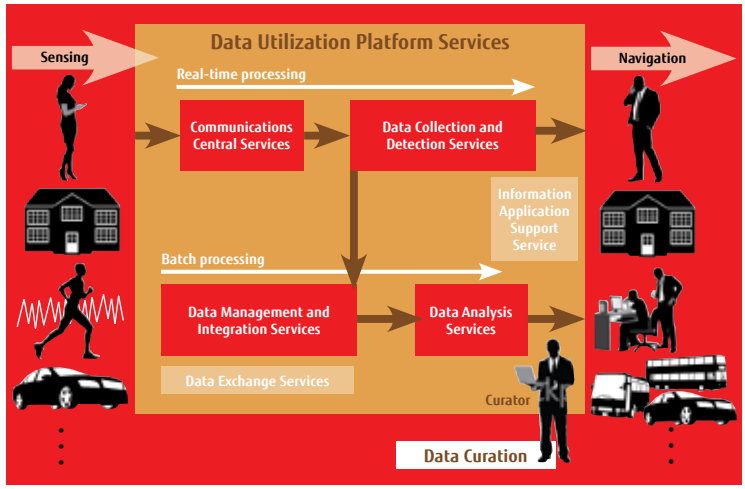
The "intelligent society" aspect of the Fujitsu research program is seeking to develop "social intelligence technology" that generates insights from a huge variety of sources including sensors, human activity and all sorts of machines, with intelligent optimization technology that allows society and businesses to understand and react to those changes. The research is wide-ranging and far-reaching. Among other aspects, it involves social sentiment and trends analysis, risk mining, natural disaster simulations, next-generation power grids and high-speed distributed parallel processing.

But this is by no means something for the distant future. Today, Big Data is rapidly moving from being a specialist field, open only to a few, large organizations, to an widely available, everyday service.

Photograph: iStockphoto

The current Fujitsu Big Data platform offerings have been build on the back of a long-standing R&D commitment in the critical area of data management, and feature a broad array of products and services.

Big Data Services



Data management and integration

A huge volume of diverse data in different formats, constantly being collected from sensors, is efficiently accumulated and managed through the use of technology that automatically categorises the data for archive storage.

Communication and control

This comprises three functions for exchanging data with various types of equipment (e.g. home appliances over networks: communications control, equipment control and gateway management.

Data collection and detection

By applying rules to the data that is streaming in from sensors, it is possible to conduct an analysis of the current status. Based on the results, decisions can be made (with navigation or other required procedures performed in real time.

Data analysis

The huge volume of accumulated data is quickly analyzed using a parallel distributed processing engine to create value through the analysis of past data or through future projections or simulations.

Data curation

To generate new sources of value, Fujitsu can offer techniques it developed for using data to make its operations more efficient or develop new businesses. This is done using "curators" – specialized analytical tools that feature mathematical and statistical data modeling, multivariate analysis and machine learning.

Big data solutions in action

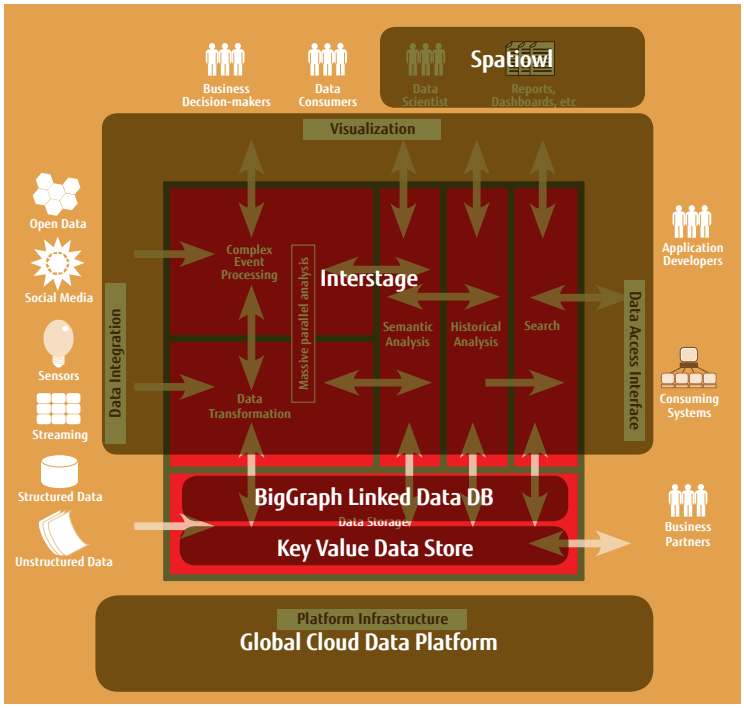
One example of Big Data in use today is TC Cloud, a Fujitsu service for the Japanese market which supports non-linear structural analysis, electromagnetic wave analysis and computational chemistry. Another is the company's Akisai Cloud, for the food and agricultural industry, which (for example) analyses environmental and other data to ensure crop yields are maximized.

Fujitsu Big Data Products

Fujitsu offers a broad range of Big Data products, and the diagram below summaries how these map onto the model of a Big Data solution outlined in Chapter 1 (page 11).

The Fujitsu Big Data offerings are built on the back of a long-standing R&D commitment into the critical area of data management.

How Fujitsu Products Deliver a Big Data Solution



At its foundation is the Fujitsu **Global Cloud Data Platform**, providing all the required data integration, distribution, real-time analytics and manipulation capabilities. This can be coupled with **Interstage®,** the Fujitsu integration and business process automation engine. Additionally, Fujitsu offers its **Key Value Store** for high-volume data storage and retrieval, supplemented by the **BigGraph Linked Data** database.

Alongside that, **Spatiowl®** is one of the Fujitsu business-focused Big Data solutions, that has been applied to resolving location data service problems.

Reshaping Business, Reshaping ICT

Fujitsu is changing technology to reshape business, bringing together a global cloud platform that enables organizations to conduct sophisticated simulations and Big Data analytics anytime, anywhere. How will this reshape business? By creating new relationships and collaborations that no one could have envisaged, by unearthing insights that revitalize existing businesses and industries, by improving the management of global resources and by accelerating innovation.

And that potential for Big Data to generate big value – for business and society – is only just being realized.

For more information on Fujitsu Big Data capabilities and to learn how we can assist your organization further, please contact us at AskFujitsu@us.fujitsu.com or contact your local Fujitsu team (see page 62).

Big Data Speak: Key terms explained

Access control	A way to control who and/or what may access a given resource, either physical (e.g. a server) or logical (e.g. a record in a database).
Application programming interface	An interface for separate computer systems to communicate in a defined manner. For example, many social networks have an API in order to allow data to be queried, uploaded and extracted by different systems and applications.
Architectural pattern	A design model documenting how a solution to a design problem can be achieved and repeated.
Availability	The proportion of time a system is live (working as it is supposed to), based on a number of performance measures such as uptime.
Big Data	(1) The application of new analytical techniques to large, diverse and unstructured data sources to improve business performance (2) Data sets that grow so large that they become awkward to work with using traditional database management tools (3) Data typically containing many small records travelling quickly (4) Data characterised by its high volume, velocity, variety (or variability) – and ultimately its value.
Business intelligence	A term used to describe systems that analyze business data for the purpose of making informed decisions.
Cloud architecture	The architecture of the systems involved in the delivery of cloud computing. This typically involves multiple cloud components communicating with one another over a loosely-coupled mechanism (i.e. one where each component has little or no knowledge of the others).
Cloud service provider	A service provider that makes a cloud computing environment – such as a public cloud – available to others.
Cloud service buyer	The organization purchasing cloud services for consumption either by its customers or its own IT users.
Cloud service stack	The different levels at which cloud services are provided. Commonly: Infrastructure-as-a-Service (IaaS); Platform-as-a-Service (PaaS); Software-as-a-Service (SaaS); Data-as-a-Service (DaaS); and Business Process-as-a-Service (BPaaS).
Complex event processing (CEP)	High-speed processing of many events across all the layers of an organization. CEP can identify the most meaningful events, analyze their impact and take action in real time.
Context-sensitive	Referring to a system, exhibiting different behaviour depending on the task or situation – for example, presenting data differently on different types of device like big-screen PCs and small-screen smartphones.
Data access interface	A way of allowing external systems to gain access to data.
Data integration	The processes and tools related to integrating multiple data sources.
Data integrity	In the context of data security, integrity means that data cannot be modified without detection.
Data residency	The location of data in terms of both the legal location (the country any related governance can be enforced) and the physical location (the systems on which it is stored).

Data scientist	An emerging and increasingly important job role involving an understanding of business challenges, the data available to address those challenges, and how best to refine and process the data to achieve desired business outcomes. This will often require mathematical, creative, communications and visualisation skills.
Data storage	The processes and tools relating to safe and accurate maintenance of data in a computer system.
Data transformation	The processes and tools required to transform data from one format to another.
Esper	A complex event processing engine available for the Java (Esper) and Microsoft .NET (NEsper) programming frameworks.
Hadoop	An Apache open-source framework for developing reliable, scalable, distributed computing applications. Hadoop can be considered as 'NoSQL data warehousing' and is particularly well-suited to storing and analysing massive data sets.
Historical analysis	The processes and tools used to analyze data from the past – either the immediate past or over an extended period.
Interoperability	The ability of diverse systems and organizations to work together.
Key value stores	A means of storing data without being concerned about its structure. Key value stores are easy to build and scale well. Examples include MongoDB®, Amazon Dynamo® and Windows® Azure® Table Storage. These can also be thought of as 'NoSQL online transaction processing'.
Linked Data	A concept (famously championed by Sir Tim Berners-Lee) in which structured data is published in a standard format so that it can be interlinked and queried, or read by both humans and machines. This facilitates the widespread use of multiple, diverse data sources in the creation of services and applications.
Map/Reduce	A programming model and software framework, originally developed by Google, for writing applications that rapidly process vast amounts of data in parallel on large clusters of compute nodes, taking a simple functional programming operation and applying it, in parallel, to gigabytes or terabytes of data.
Non-repudiation	A service that provides proof of the integrity and origin of data together with authentication that can be asserted (with a high level of assurance) to be genuine.
NoSQL	An alternative approach to data storage, used for unstructured and semi-structured data.
Open data	Data which is made freely available by one organization for use by others, generally with a license attached.
Personally identifiable information	Data that, by its nature, is covered under privacy and data protection legislation. This applies to information about both employees and consumers.
Real-time	Providing an almost instantaneous response to events as they occur.
Search	The processes and tools that allow data to be located based on given criteria.
Semantic analysis	The processes and tools used to discover meaning (semantics) inside (particularly unstructured) data.
Semantic web	A collaborative movement led by the World Wide Web Consortium (W3C) that promotes common formats for data on the web.

Service level agreement (SLA)	Part of a service contract where the level of service is formally defined to provide a common understanding of services, priorities, responsibilities and guarantees.
Shadow IT	A term often used to describe IT systems and IT solutions built and/or used inside organizations without formal approval from the IT department.
SOLR	An Apache open-source enterprise search platform which powers the search and navigation features of many of the world's largest Internet sites.
Structured data	Data stored in a strict format so it can be easily manipulated and managed in a database.
Structured query language	A commonly used language for querying structured data, typically stored in a relational database.
Tokenization	The process of replacing a piece of sensitive data with a value that is not considered sensitive in the context of the environment in which it resides (e.g. replacing a name and address with a reference code representing the actual data, which is held in another database hosted elsewhere).
Unstructured data	Data with no set format, or with a loose format (e.g. social media updates, log files, etc).
Uptime	A measure of the time a computer system has been available (working as intended). Not to be confused with overall system availability, which will depend on a number of measures, including the uptime of individual components. (See also Availability)
Visualization	The processes and tools for presenting the results of data analysis in a manner that enables better decisions to be made.
World Wide Web Consortium (W3C)	An international community that develops open standards to ensure the long-term stability and growth of the web.

Also in this series...

The White Book of...

Cloud Adoption
The definitive guide to a business technology revolution

Even in an industry hardly averse to talking up the "Next Big Thing", there is a phenomenal amount of hype and hot air surrounding cloud computing. But cloud is real; it is a huge step change in the way IT-based services are delivered, and one that will provide substantial business benefits through reduced capital expenditure and increased business agility. The key issue that IT decision-makers must address is how and where to adopt cloud services so they maximize the benefits to their organizations and their customers.

This Fujitsu *White Book*, produced in consultation with some of the UK's leading CIOs, cuts through the market hype to clearly explain the different cloud models on offer. It also provides a mechanism to determine which IT applications and business services to migrate into the cloud, setting out best practice and practical approaches for cloud adoption.

Cloud Security
The definitive guide to managing risk in the new ICT landscape

The journey to cloud is no longer a question of "if" but rather "when", and a large number of enterprises have already traveled some way down this path. However, there is one overwhelming question that is still causing many CIOs and their colleagues to delay their move to cloud: Is cloud computing secure? A simple answer is: Yes, if you approach cloud in the right way, with the correct checks and balances to ensure all necessary security and risk management measures are covered.

By providing a clear and unbiased guide to navigating the complexities of cloud security, this book will help to ensure your cloud computing journey is as trouble-free and beneficial as it should be.

To order these books, and for more information on the steps to cloud computing, please contact: AskFujitsu@us.fujtisu.com

Fujitsu Regional Offices

Europe, Middle East, Africa, India
FUJITSU (UK & IRELAND)
+44 (0) 870 242 7998
askfujitsu@uk.fujitsu.com
uk.fujitsu.com

FUJITSU TECHNOLOGY SOLUTIONS
(CONTINENTAL EUROPE, MIDDLE EAST, AFRICA & INDIA)
+49 1805 372 900
(14ct/min; mobile devices are limited to 42ct/min)
cic@ts.fujitsu.com
ts.fujitsu.com

FUJITSU (NORDIC REGION)
+358 45 7880 4000
info@fi.fujitsu.com
www.fujitsu.com/fi

North America
FUJITSU AMERICA , INC
+1 800 831 3183
globalcloud@us.fujitsu.com
www.fujitsu.com/us

Asia Pacific
FUJITSU HEADQUARTERS
+81 3 6252 2220
Shiodome City Center, 1-5-2 Higashi-Shimbashi
Minato-ku, Tokyo, Japan,105-7123
www.fujitsu.com

FUJITSU CHINA HOLDINGS CO LTD
+86 5887 1000
www.fujitsu.com/cn

FUJITSU (AUSTRALIA)
+61 9113 9200
askus@au.fujitsu.com
www.fujitsu.com/au

FUJITSU (NEW ZEALAND)
+64 4 495 0700
askus-nz@nz.fujitsu.com
www.fujitsu.com/nz

FUJITSU (KOREA)
+82 (080) 750 6000
webmaster@kr.fujitsu.com
www.fujitsu.com/kr

FUJITSU (SINGAPORE)
+65 6512 7555
fujitsucloud@sg.fujitsu.com
www.fujitsu.com/sg

Ref: 3402

FUJITSU AMERICA, INC.
Address: 1250 East Arques Avenue Sunnyvale, CA 94085-3470, U.S.A.
Telephone: 800 831 3183 or 408 746 6000
Website: http://solutions.us.fujitsu.com
Contact Form: http://solutions.us.fujitsu.com/contact
Have a question? Email us at: AskFujitsu@us.fujitsu.com