Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

Business Analytics | the way we see it

Big Data

Next-Generation Analytics

People matter, results count.

Table of contents

Executive summary

Big data is the big message that technology vendors have been pushing in the business intelligence and data market in the last two years. The response from the market has been one of ambiguity, and to a degree of misalignment of definitions and understanding. Clearly the three Vs (Volume, Velocity, and Variety) have begun to emerge as a common theme, if not as the defining characteristics of big data. By focusing too much on the shape of big data, though, we are in danger of thinking purely in technology terms, when we should be concentrating on the business outcomes that it can deliver.

A clear message from the market has been an acceptance that big data is a good thing. It allows businesses to analyze a much broader set of data about aspects of their business as an integrated whole. This ability is providing a new level of insight and opportunity, and ultimately a new source of business value. The game-changing aspect of big data lies in using data from beyond the firewall to derive new insights, often in real time or near real time. This is often, but not exclusively, unstructured data such as multimedia and social network content. The enterprise's internal data may well be necessary to enable those insights, but cannot, on its own, produce the game-changing moment.

Because big data can transform the way your front office interacts with the customer, investments in big data here are likely to produce better returns than those in back-office process improvement. However, there are many opportunities to derive value from big data across the business spectrum. The most obvious relate to social media, with applications like reputation management but there are many more opportunities, ranging from better-informed credit checks and exploitation of digital assets such as commercial multimedia to equipment self-monitoring and replenishment. With the proliferation of new sources of data, the opportunities can only increase. Organizations need to be selective: the strategy for exploiting big data must support the overall business strategy.

A range of new technologies – for example Hadoop – have emerged to deal with the technical aspects of managing large volumes of data. Whilst these technologies represent a significant departure from what has gone before, traditional disciplines of information management still remain essential. In fact big data concepts have been around for a while, and since the 1990s vendors such as Thinking Machines, Teradata, Informix, and Sybase have all offered solutions to the volume issue. Recent entrants have focused specifically on the complexity (Variety) and Velocity aspects thrown up by social media and multimedia customer interaction.

Even more important than the technology is the governance and organization-wide co-ordination required to "master" (i.e. structure) data as far as possible, so that it can be used in a consistent and meaningful way. Equally critical is the approach to analyzing the newly available data – often in conjunction with more conventional data – to obtain insight and drive action.

Big data will bring about the next wave of performance improvement by leading to more effective, fact-based decision making and optimizing business processes. It's very likely that your competitors are already deriving benefit from big data, or at least planning how they can do so. In recent research conducted for Capgemini[1], respondents estimated that, on average, they have seen business performance improvements of 26% in processes where big data analytics have been applied, and that they expected those improvements to accelerate rapidly. Clearly, any company that wants to maintain competitive advantage must start getting to grips with big data.

[1]The Deciding Factor: Big Data & Decision Making", June 2012.

Introduction: What is big data and why is it different?

Big data has emerged as a key topic for CEOs and CIOs as a result of new business drivers and opportunities that make it necessary to use data as a corporate asset in order to become more competitive and to create real business value.

There is no unanimity as to what big data actually is. The classic definition is in terms of Volume (the sheer size of the data), Velocity (the speed with which the data is collected and needs to be processed), and Variety (the different formats of data) – but there are other definitions. Most discussions suggest that some or all of the following features are found in big data:

- Data volumes are higher than a given organization is accustomed to processing.
- Data volumes are larger than can be handled by traditional database technology.
- External data that is brought into the business from third-party or public sources.
- Some of the data may come from social media.
- A significant amount of data may be highly unstructured (e.g. voice or video).
- Various data sets of different types are integrated together for analysis.
- Real-time or near real-time analysis is sometimes required.

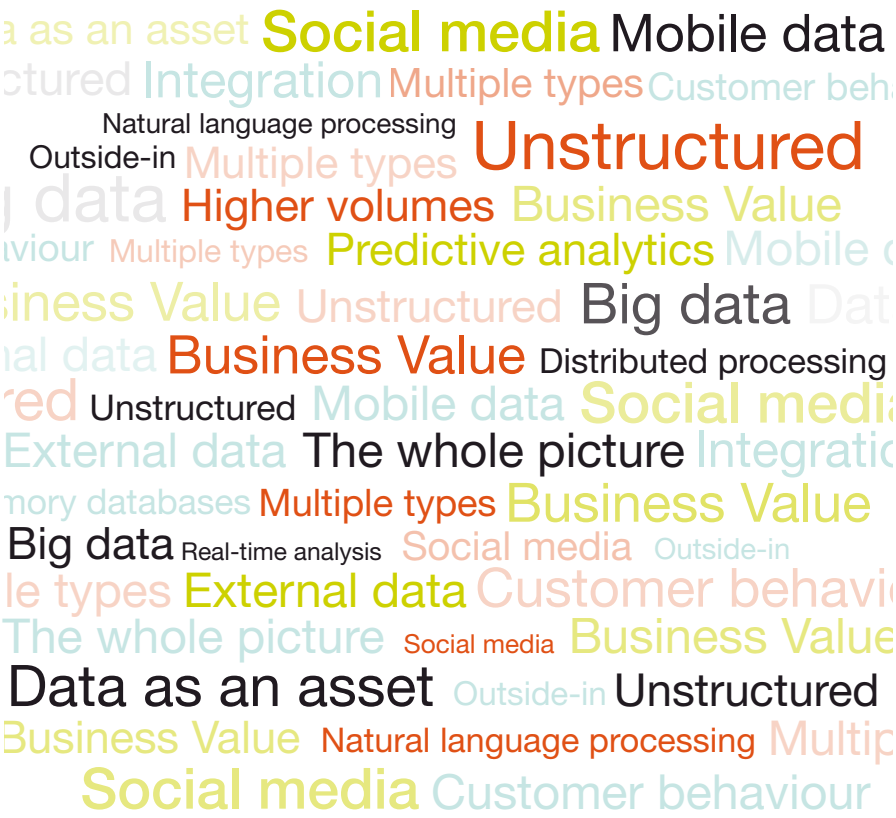

Figure 1: Big data incorporates many different features

However, in our view, none of these features is necessary or sufficient to define big data. We find it more useful to view the concept in terms of three elements:

- The data itself.
- The process for dealing with the data.
- The holistic view that it can enable.

While identifying characteristic tendencies of big data in each of these areas, we'll argue that the data and process aspects are, in fact, not not so different from what has gone before; the real novelty of big data lies in the opportunity it provides to see aspects of the organization's business in new and more holistic ways, and importantly as a source of business value.

The data

Recent years have seen exponential growth of data in businesses and society at large. Contributors to this growth, which is enabled by low-cost storage, include:

- Online business, mobile computing, and social media.
- Adoption of digital sensors and RFID tags, connected to ever larger sensor networks, together with IP-addressable objects leading to the "internet of things".
- Digitization of voice and multimedia.
- Automation of processes like client interactions.

Large data volumes mean different things to different organizations. At the very large end, where Google and eBay are dealing with petabytes of data, these activities clearly fall into the "big data" category. The grey area comes lower down the scale: many large corporations are already managing data warehouse applications with 100s of terabytes of data, but for smaller organizations the same volumes would be seen as "big".

The lack of a clear threshold in terms of volume means that we are drawn to a definition in terms of technology: "big data" could refer to instances where new distributed architectures (e.g. Hadoop) are needed to achieve affordable storage of larger volumes of data. However, traditional databases are evolving quickly and coming down in cost, and so even this definition is marginal. So we need to look to other factors for a definition.

More characteristic of big data is the importance of external data from beyond the firewall. Organizations are moving from an inside-out orientation – where they analyze and report on data from within the organization – to an outside-in one, where data from outside the enterprise is brought inside to provide new value. This is the really game-changing feature of the big data concept: data from outside the organization, or from its edge, joining with data from inside to provide a new, more holistic, view.

Insights derived from big data typically inform the front office's interactions with the customer. That means an investment in big data is likely to produce more worthwhile returns than a comparable investment in back-office processes, which have been undergoing improvement for years. But big data still offers significant benefits in the back office too.

The process

The big data concept is also associated with a process or approach. Clearly, it is not enough to collect huge volumes of data. Big data initiatives have to identify the right data, organize it into a form that can be explored with analytics, and then use those analytics to derive insights – to allow the business to "thrive on data", as we have called it elsewhere (see Capgemini's Technovision: Thriving on Data)[2].

Big data builds on existing technologies like data warehousing, business intelligence, data mining, enterprise content management, and search (although it combines them in new ways), and extends or replaces them. In many cases this is about data streaming and real-time analysis. These aspects are covered in more depth in later sections.

The holistic view that big data can enable

Increasing volumes and complexity of data have led to the emergence of new technologies and processes. However, big data's real differentiator is its ability to provide radically new, and far more complete, views of many aspects of the business, including customers, processes, supply chains, and products. Organizations are achieving previously unattainable levels of insight by bringing in data from sources that either did not exist before or were seen as out of reach because they were external to the business – and, crucially, by combining that new data with data from within the enterprise. Organizations are starting to view all this data as an integrated whole, recognizing that to get "the whole picture" of their business they need to bring data together regardless of source, format, or type.

Exploiting big data entails dealing with:

- A multiplicity of sources – everything from internal ERP systems and document management systems, to external sources like the internet and business partner systems, or those of third-party suppliers of market or demographic data.
- Both structured and unstructured formats – structured data from database tables, semi-structured data in forms and XML files, and unstructured free text, voice, and video data.
- A range of customer data types – from "exact" client multi-channel interaction and sales data, to volatile online behavior data and social media data.
- A range of technical data types – from formally approved and managed product design and engineering data to "fuzzy" sensor data.

Combining and analyzing all this data provides a richness and depth of understanding never achievable before. Consequently, more and more organizations are viewing their huge data collections as a primary source of business value.

This is a realistic view because for the first time, we now have both the business drivers and technological capabilities to turn data from a by-product of automation into a first-class corporate asset for a competitive business.

[2]www.capgemini.com/services-and-solutions/technology/technovision/clusters/thriving-on-data

Big data brings new challenges

The truly distinctive feature of big data, then, is the business's expectation that it can treat large volumes of diverse data as an integrated whole, and a source of value. This expectation creates a number of new organizational, as well as technological, challenges.

Organizational challenges mostly relate to the integration of data. In Capgemini's recent survey[1] , business users were asked to identify their biggest impediment to using big data for effective decision-making. Most (54%) cited the existence of "silos" preventing data from being pooled for the benefit of the entire organization. These barriers need to be broken down, and data shared as the corporate asset it is.

The technological challenges are well documented. There is a need for technology that can integrate and handle data regardless of source, format, and type, along with the ability to achieve real-time or near real-time processing. A further requirement is for analytics to transform huge volumes of data into relevant information and practical insights.

The more data gets shared, the more its quality depends on mature data management, governance, and stewardship. An effective master data management strategy is critical to the data integration challenge. As organizations increasingly automate decisions based on big data analytics, data quality will grow ever more critical. A simple master data error could result in the wrong response to a customer, with all the damage to the relationship which that entails.

Analyzing more customer data, social media, and web transactions also means that organizations need to be conscious of privacy and security requirements, and of the significant variations between countries. Overenthusiastic use of personal data can land organizations in court – not to mention the impact on corporate image.

[1]The Deciding Factor: Big Data & Decision Making", June 2012.

The business opportunity

Big data provides opportunities across the spectrum of business activity, but the prime benefits can be grouped under three headings:

1. Improving interaction with the ecosystem, particularly with customers.
2. Improving business processes.
3. Risk mitigation.

We'll look at each of these in turn before contrasting strategic opportunities with action-oriented ones.

Improving partner interaction with the ecosystem, particularly with customers

Many of the sources of big data are external to the enterprise (the outside-in view mentioned above) and generated by business partners (mainly customers, but also vendors and third parties) via social media, web transactions, and goods movements. These are allowing a much more intimate relationship – a better understanding of behavior – be it how they consume electricity or their preferences on pretzels.

For example a mobile phone operator might group individual account holders into households, so that it can target offers based on knowledge of the whole household's product mix (such as whether the household has a broadband connection). This will enable step changes in targeted marketing, tailored service offerings, and customer retention.

This will bring a whole new level of customer experience, where your service or product provider will treat you as an individual and will have a level of knowledge about you that might seem frightening. But today's new consumers are embracing this type of experience. They understand that it will mean a much better level of personal service, where they are presented only with offers and services that are relevant to them, often at preferential prices. An example which is emerging in the smart energy solutions market is a home energy management application. This could record power usage for home devices, identify energy-inefficient equipment (e.g. a freezer), and share the data with partner device manufacturers who then prompt the customer with the cost savings case for buying a more efficient replacement.

Improving business processes

Many of the benefits here overlap with improving partner interaction, as they aim to improve customer or vendor business processes. However, improving business processes is a broader category.

Improving a business process depends on having the data to understand it in more detail. The data might be smart grid information, telemetry from aircraft or other transport systems, or readings from electronic devices (such as medical equipment or factory machines). Combined with analytical tools, these types of data are enabling better prediction of future activity and performance and allowing organizations to adjust processes to achieve the best outcomes.

For example, monitoring plant and equipment lets you know in advance when parts are likely to fail, and take preventive action. RFID can tell you how a product is running through the supply chain and enable you

to optimize the flow dynamically. In healthcare, providers are using information about patient experience and outcomes to inform product development decisions.

Risk mitigation

Big data is enabling organizations to understand and quantify risk. Understanding performance of components in a large manufacturing or processing plant will alert operators to risks early on, allowing corrective action to be taken.

Big data also provides external risk mitigation – for instance sentiment analysis allows organizations to find out what is being said about them in discussion forums in order to pre-empt problems. In the financial sector, risk is high on the list and big data analysis is becoming a key part of managing risk compliance.

Strategy versus action

Some opportunities are longer-term, helping to define a more informed strategy or forecast, while others are action-oriented, for example automated up-sell based on customer profiles in web and call center interactions.

Big data plays in both areas. Often it is the longer-term strategy element that determines what immediate actions are feasible. For example, analysis of buying patterns will provide a framework for determining which cross-sell opportunities are likely to be effective in an individual interaction. The intended use determines the timescale in which data is required, as discussed in our section on business analytics.

Your competitors are already seizing the opportunity

In our recent survey of over 600 business leaders, 57% agreed or strongly agreed that most of their competitors are already using big data to their strategic advantage. Many were already seeing significant benefit themselves from their own use of big data, and expected this trend to accelerate in the next three years.

An attitude of "if it ain't broke, don't fix it" will not work when your competitors are being more efficient in their marketing because they are better targeted, or when they are improving their call response by optimizing their mobile workforce based on better analysis. Those who do not take advantage of this new source of business insight will be left behind.

To realize the opportunities, big data has to be used in ways that release its value. Later sections of this report discuss how best to do that. First, however, we review the technologies and techniques that are available to help organizations deal with big data.

Traditional information management techniques remain essential

We have seen that, as well as conventional structured data, big data solutions often need to deal with a variety of unstructured data such as social media content, documents, and streaming audio and video. There may also be instrument data from devices like RFID, sensors or smart meters, logs from databases and firewalls, and more. Content is dynamic and may need real-time or near real-time processing.

The next section discusses some of the new and emerging technologies designed to deal specifically with these big data challenges. As we have noted, however, big data is not about technology, either new or old: it is about data management, leading to business value. As business value from data is a traditional goal, it is not surprising to find that traditional information techniques and disciplines are often called for, either in conjunction with or instead of the newer ones. Furthermore, many of the vendors of traditional technologies are also scaling up their solutions to deal with these new challenges.

To see how these traditional solutions address big data requirements, let's consider the four major steps required to implement a big data solution: acquisition, marshaling, analysis, and action. For each step, we will indicate which traditional techniques are useful and we will point out where traditional techniques complement newer technologies; this mainly applies to data marshaling and to a lesser extent analysis.

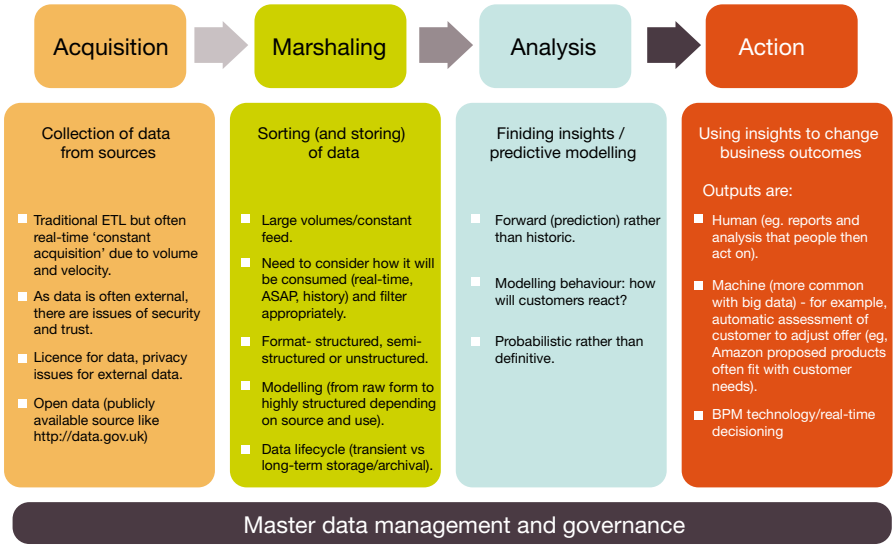

Figure 2: Big data process model

Step 1: Data acquisition

This includes two elements: Extraction/transfer and integration:

- **Extraction/transfer.**
  Large volumes of data must be obtained from a range of sources, including external and mobile ones. Source format and content often change – for example, when smart meter or network antenna software is upgraded, or when the hardware is physically replaced.

  Extracting and transferring data is therefore likely to require a capability to manage complex projects and multiple teams, constraints and technologies, always with an eye on budgets.

  There are other non-technical points to address: legal licenses for data storage, for example, and privacy issues in the case of external data. These issues become particularly important when you are accessing sensitive private data like geographic localization, or content from social networks, for instance. Publicly available data, and data from suppliers or partners, raise complexities of their own.

- **Integration.**
  Extract, transform and load (ETL) tools can integrate non-structured as well as structured data. To achieve continuous data acquisition, they can work in a quasi-continuous mode, being configured to spool data every five minutes, or every time 10GB of data is waiting to be loaded, for instance.

  Classic enterprise application integration (EAI) technologies such as the Enterprise Service Bus (ESB) are also frequently used as a complementary solution to integrate messaging data flows.

  Complex Event Processing (CEP) tools sit somewhere between ETL and ESB, providing more transformation and rule capabilities than ESB but less ability than ETL to manage high-volume database integration. They often have the ability to keep some historical data in memory and compare it with incoming data to detect exceptions and produce alerts. These are classic technologies but are now being used in a new mode.

  Integration is also the right time to compute metadata – for example, the metadata needed to extract what has been said from a recorded sound, or to take real text content from a web page. This metadata can then be stored once and for all, to save computing it again for every new analysis.

  Both aspects of data acquisition can be handled by classic tools for extraction/transfer and integration.

Step 2: Data marshaling

Not all big data has the same destination. Most data will go to a single destination, some may go to more than one destination, and some nowhere at all. The choices depend on intended use, i.e. on whether the data is wanted for historical, real-time, or "as soon as possible" analysis, as the record of truth, or as a source of signal data.

Much data storage will be in traditional architectures:

- Structured data storage on very large databases of the type associated with classic business intelligence (BI) data warehouses.
- Large enterprise content management (ECM) solutions – big data doesn't mean ECM solutions do not work any more.
- Dedicated big data solutions like Apache Hadoop complement these traditional architectures by providing a low-cost option for storing large volumes of more or less free-form data (in fact, there does not have to be a data model at all).
- Search engine indexes. Some content will directly feed the indexes to speed up future searches (metadata generated during integration is often used to feed these indexes efficiently).
- Archiving solutions may be used, for instance, to guarantee (for legal purposes) that stored data has not been altered since it was stored, or to store at low cost old data from databases that is not used any more but still needs to be kept.
- Transient data. Not all data needs to be held. It may just pass through, being acted on in the moment or leaving some summary information.
- Garbage. Not all data has to be stored – and big data tends to contain a high proportion of garbage.

Another relatively recent addition is specialist in-memory databases, which can be used as in the CEP example discussed above, or to launch a fast analytics procedure of the type needed for fraud analysis.

Generally, even if new big data technologies are used, it is likely to be in conjunction with traditional architectures. Whatever combination of old and new is chosen, the choice of storage architecture should be made with the entire lifecycle in mind.

Step 3: Analysis

When it comes to the quest to turn data into insights, once again classic tools will exist alongside, and sometimes in place of, specialist ones. SQL, improved SQL, extraction to SAS, and SPSS are all options here.

Traditional BI is often talked about as a "rear-view mirror" on the business: it tells you where you have been. With big data there is an emphasis on forward analytics (i.e. prediction) rather than historic ones. Modeling emphasizes behavior – for example, predicting how customers will react to a change. Predictions tend to be probabilistic rather than definitive, e.g. trying to work out the optimum time to replace parts.

Sometimes these requirements are best tackled using the new generation of big data-oriented tools – for example, as will be discussed in the next section, MapReduce and R can be used along with Hadoop. But once again, these are very likely to be found alongside the classic tools.

Step 4: Action

To get value from big data, analysis must lead to fast action – the action has to be an integral part of the process. Actions can be carried out by three types of agent:

- A human acting on a report or analysis. However, a human can readily understand only certain types of content, like figures reported against key performance indicators, documents such as PDFs or spreadsheets, and outputs from search engines. Providing big data analytics, without further interpretation, to humans is therefore of limited use.
- A computer – which is more likely to be the case with big data. For example, a site like Amazon's may automatically assess a customer's characteristics in order to propose an offer that fits with that customer's needs. In this case, the content needed to make the right decisions is much larger and more complex than for humans. It is likely to include predictive models, full list of products and probabilities per type of customer, and so on.
- A mixture of human and computer. For example, a contact center agent may use intelligence about spending patterns to decide whether to advance credit. In this case, statistical models are applied to give to the agent some further interpretation through pre-computed solutions, but the final choice is left to the agent.

The action step is where big data, with its need for immediate action, diverges most from traditional approaches, where action tends to be slow and "cold". Nonetheless, the disciplines are once again those of traditional IT. Business process management (BPM) techniques and real-time decisioning, for example, can be used to integrate big data results into an existing process, as required by the credit scoring example above.

Hadoop: A new technology for big data

All the major software vendors have solutions out in the market for big data. These include extensions to data warehousing and content management technologies, in-memory solutions, parallel processing and advanced analytics. Often, one or more of these solutions will be the right answer for a client's big data needs.

However, this paper does not aim to compare and contrast these solutions. Instead, we will talk about one new technology – Hadoop – because it represents a departure from previous technologies, and brings with it new challenges and opportunities.

Hadoop is designed to deal with very large datasets by distributing the data over many servers. To understand this technology it is useful to consider it in relation to the four major steps identified earlier – acquisition, marshaling, analysis, and action. Hadoop and its related technologies really only address the marshaling and analysis steps. Within these steps, they perform the following functions:

1. Marshaling

- Storage and data management technologies to deal with terabytes, even exabytes, of complex data in a wide range of forms.
- Processing this high volume of complex data in a distributed manner.
- Administration of the big data environment.

2. Analysis

- Data mining and predictive analysis to identify patterns, find the insight, and obtain value from the data.

Each of these functions is discussed below.

There are many players competing in this space, but we will illustrate with reference to the Apache Hadoop framework. This is because most big vendors have invested in Hadoop, and there are signs that it is becoming a *de facto* standard:

1. Storage and data management. Before big data can be used for business purposes, it must be stored and managed efficiently. Distributed file systems, elastic storage systems, and distributed and massively parallel processing databases enable storage of much larger volumes at low cost.

Open source frameworks like Apache Hadoop have had a major impact on the storage of high-volume and complex data in a distributed manner across large numbers of servers.

The Hadoop File System (HDFS) is the heart of the Hadoop Framework and has two major components: NameNode and DataNode. NameNode manages the file system metadata, while DataNode stores the data. The entire Hadoop framework is built using Java and hence all the major components of the Hadoop framework can interact with each other using Java.

Other important storage mechanisms in the Hadoop framework include:

- **HBase.** An open source, distributed, versioned, column-oriented store that is in a true sense a NoSQL database. HBase provides BigTable-like capabilities on top of Hadoop.
- **Hive.** Structures data into well-understood database concepts like tables, columns, rows, and partitions.

- **Further (non-Hadoop) NoSQL databases.** These include Cassandra (multi-column store); CouchDB and MongoDB (document databases); Redis, Riak, and Membase (Key Value (KV) stores); and Neo4j (graph database).

2. Processing and data integration. Before meaningful analysis can be performed, there has to be a robust mechanism to churn and crunch a high volume of data, and integrate a variety of types of data, all in an acceptable time frame. Important elements here include:

- **MapReduce.** A programming model for efficient distributed processing, designed to perform computation reliably on large volumes of data spread across the HDFS, in parallel.
- **Pig.** A high-level data processing language which analyses datasets with Hive. Pig is an abstraction layer on top of MapReduce.
- **Hive SQL.** An interface to access the Hive structure.
- **JAQL.** A scripting language for large-scale, semi-structured data analysis.
- **Cascading.** An API which makes it easier to perform complex operations such as grouping and aggregation.

3. Administration. Because big data solutions usually involve massively parallel and distributed processing, a dedicated effort is needed to manage the entire environment from a performance, optimization, and load balancing perspective. Relevant tools include:

- **Hadoop On Demand (HOD).** A system for provisioning and managing independent Hadoop MapReduce and HDFS instances on a shared cluster of nodes.
- **Zookeeper.** Cluster management, load balancing, etc.
- **Chukwa.** Data collection system for monitoring distributed systems.

It is worth noting that administration is the least developed of the new technology categories, and it is an important part because large distributed networks need a lot of administration. This is a fast-evolving part of the market.

4. Data mining and analysis. Thanks to the new capabilities for processing and integrating data, organizations are now in a position to mine all the available data, instead of just a sample data set as would have been the case in the past. The result is more accurate analysis and better predictions.

Although traditional predictive analytical tools used against relational databases still play an important role when the data is stored in a distributed landscape such as Hadoop, some additional tools are needed for this environment. These include:

- **MapReduce.** This has data mining and machine learning algorithms which can help to drill into huge volumes of data, and help to mine the data using previously developed models. MapReduce is always necessary to access the data in the Hadoop infrastructure and transform it into a "big grid" computing solution to enable analytical tools.
- **Apache Mahout.** A Java library of machine learning and data mining algorithms, many (but not all) of which are designed to run on Hadoop. The algorithms are categorized into four main use cases: recommendation mining; clustering; classification; and frequent itemset mining.
- **R.** Although a standard analytical language used against relational databases, this is also the language used on Hadoop environments, as it can generate MapReduce code.

New and old must co-exist

We have shown that in implementing big data solutions, it is necessary to combine traditional tools and techniques with new technologies designed especially with big data requirements in mind.

It makes no sense to retrieve large amounts of data without being able to manage it and link it to business meaning. Master data management (MDM) is not an option for big data – it is part of it. We therefore turn next to the management strategies that companies need to adopt in order to derive business advantage from big data.

Mastering big data

To get value from big data, it's necessary to access and use large and disparate information sets to gain insights into markets and opportunities. That means, above all, being able to link different types of data, internal and external, together.

Before you can do that, you have to "master" the data – and that requires not only the right techniques but also the right organizational structure to be in place. This can be regarded as an additional aspect of "marshaling" data.

Importance of data quality

The garbage in, garbage out (GIGO) rule applies to any system. The difference with big data is that the consequences of getting the inputs wrong can be catastrophic. The larger the data volumes, the greater the impact of poor quality becomes – and the smaller the opportunity to correct problems by hand.

The consequences of poor inputs are particularly serious when trying to make sense of customer data. If you fail to identify a customer correctly in different transactions, you could form the mistaken impression that you are looking at five customers instead of one customer buying five products. That can impact not only the company's internal analysis but also the customer's experience.

Structuring big data: the POLE framework

The secret of mastering big data is to structure it correctly, which above all means correctly structuring its "core" entities. The core of the data consists of the common reference points – people, organizations, accounts and so on – that link together the mass of data such as transactions or social media interactions forming the really big part of big data.

Get the core right and you ensure consistency and correct interpretation (so that you know you are dealing with one customer and not five in the example above). With tight control of the core, analytics work better, and more value can be derived from big data.

In thinking about how to structure the core data, it is useful to think in terms of POLE: an acronym that stands for Parties, Objects, Locations and Events. The Parties, Objects and Locations are core data; the Events are the mass of transactions and interactions. If you are analyzing consumer behavior, a party might be a customer or a retailer, an object might be a product, a location might be a customer's home, a retail outlet or a social media channel.

Correctly identifying the P, O and L of the POLE gives a clear structure for the core entities, which can in turn be used to structure event information. This type of framework can accommodate new dimensions and information sources that may be required in future without massive re-engineering of the solution.

Loading Events into the POL structure

After defining the framework and implementing it on your chosen technology platform, it becomes possible to load the Events – the transactions. Each event should get linked to the POL elements in such a way that information is consistent across channels and interpreted consistently in all parts of the organization.

External data may need significant manipulation to ensure it too is consistent, which given that data volumes are also very large, implies a need for significant processing power. This is particularly true of unstructured data which can be linked to POL to provide the first stage of business comprehension but needs much more work to turn it into information that can be analyzed.

Consistency is, however, essential if the result is to be of sufficient quality to enable adaptable and accurate analysis. Therefore, it is important to choose a technology platform that has the capabilities and the scalability to carry out the transformation in the required timescale.

Not everything can be mastered

What is optimal from a standards perspective may well be suboptimal, or impossible, from an operational policy perspective. Forcing every individual to identify themselves by inputting 10 key attributes will probably reduce the amount that you sell.

To get the best out of big data, the key is for a business to understand both what it would ideally like to achieve and what is practical. If there is no value in mastering or controlling a particular piece of information, don't. In practice there will be a lot of data that cannot be mastered: it may not be as useful as mastered data, but it can still be valuable.

Data governance

Before defining the structure of big data, two types of governance need to be in place:

1. Governance of standards which define the structural format and definitions of information:
   - When are two objects, locations or parties considered equivalent?
   - What core information is required to identify these elements?
   - What rules govern the relationships between them?
2. Governance of policies which define how the standards will be enforced:
   - Is there a central cleansing team?
   - At which point is a relationship considered valid?
   - What will be done with 'possible' matches?
   - Who gets to decide when a standard needs to change?

Governance of this type should not apply only to big data, or to a single information area, but to all the common elements that occur across the enterprise, and in its external interactions.

Iterative approach

Mastering data is not about achieving perfection instantly: it's about improving quality and control over time, in a way that creates incremental value for the business.

Therefore, the mastering process is usually an iterative one, with areas prioritised for the POLE treatment according to what analysis the business needs at a given time. This iterative approach works well because the organisation learns more about big data techniques at each stage, and can put its new knowledge to use in subsequent stages.

Adding value – business analytics for big data

Predictive analytics (PA) hold the key to generating value from big data. There are three ways in which big data can be used by PA: it can be used to enhance existing analytics, create new analytics, and enable better decisions:

Enhancing existing analytics

Big data techniques make it possible both to examine additional dimensions and patterns in traditional data, and to analyze it in conjunction with new types of data. When applied to customer behavior, these techniques can be extremely useful in formulating strategy.

An illustration is provided by the history of the telecom industry and its attempts to predict and manage churn (see Figure 3).

Companies have traditionally relied on analyzing customer data and product information, together with call detail record (CDR) data. To this they have added contact center data, and then data about network behavior – the latter allows a company to correlate its own performance with customer attrition. The recent availability of social media data has provided detailed information about customer preferences.

Applying new techniques, algorithms, and software to traditional data such as CDR makes it possible to explore additional dimensions of that data, as does combining it with new data. For instance, companies can use social network analysis (SNA) to find out more about customer usage patterns, identifying influencer/follower relationships in social networks that may influence churn, and relating those to the CDR data they already have.

Text mining applied to unstructured data – such as call center notes, blogs, or customer feedback on websites – can further increase insight into customer preferences and behavior.

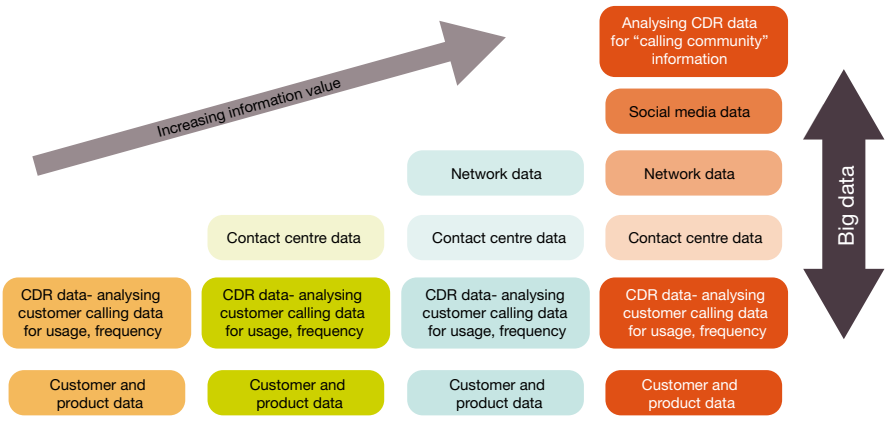

Figure 3: Developments in analysing telecoms customer data

Integrating big data techniques with existing data mining and advanced analytics can not only strengthen a company's ability to predict churn, but also build a retention strategy that targets the right customer segments. A similar approach can build and enhance marketing, cross-sell and up-sell models.

Creating new analytics

New types of analytics are becoming possible as a direct result of the additional types of data now available. These analytics can help a company to sell more products and provide a better service and overall experience for customers.

For example, supermarkets can tailor special offers for customers currently in the store, based not only on their profile, preferences and purchase history but also on the items they currently have in the trolley. (This can be identified by an intelligent trolley that scans barcodes or RFID tags.) For example, if the customer has picked up baby toiletries, a baby food coupon could be offered.

Again, with the advent of smart meters, power companies will be able to collect detailed data about customer energy consumption. They will be able to offer customers guidance about the best times to use appliances, and can generate alerts to reduce wastage. They can also forecast usage with greater accuracy, both for individual consumers and to synchronize their power generation supply processes.

Enabling better decisions

Big data technology enables analytics to be implemented in (near) real time. If modeling data is collected more frequently, it can be used to improve business rules on the fly, so that those in customer-facing roles (for example, salespeople and contact center agents, or their virtual counterparts) can achieve real-time decisioning.

Consider a credit decision by a contact center agent (or "virtual agent"). Traditional scoring methods like customer segmentation may assign a high credit score to someone who has just lost their job, for instance. Now, customers can be scored on the basis of recent payment patterns, or additional credit taken up in the past few days, or even today, and the results relayed to contact center agents.

Big data analysis can also yield competitive intelligence about other companies' offers, which should again result in better decisions by customer-facing personnel. For example, a television company's call center agent talking to a customer who is threatening to leave should make different offers depending on whether a better deal is available in the marketplace.

Practicalities of analyzing big data: need for collaboration

To date, predictive analytics and data mining have depended on a combination of science and art. Data analysis has tended to be iterative, discovery-based, and not completely automated. In a typical scenario, analysts extract data from the data warehouse into the analytical platform where they perform data discovery, model development, and model evaluation. The final model is handed back to the IT team to implement and integrate it into various BI and operational applications.

IT and analytics teams now need to change their work processes. It is only by integrating big data with existing enterprise data that you can get the "whole picture", and that requires collaboration. IT people must become more adaptable in their role as custodians of data and processing power/memory in order to enable a new, less structured, and more discovery-oriented style of analytics. At the same time, predictive analytics teams need to harness their inventiveness so that the new results and models can be seamlessly integrated with business processes and applications.

If the functions fail to collaborate, organizations will not get the whole picture and therefore they are likely to miss out on some of the intended benefits.

Speeding up response

Prevalent techniques tend to rely on analyzing historical data, while many applications of big data need predictive analytics to access current information, and to provide insights or modify models in (near) real time. This necessitates new PA techniques that are in some sense "self-correcting" or "self-learning", such as Bayesian methods or machine learning algorithms.

There are also ways to speed up traditional analytics using newer methods such as in-database or in-memory analytics. Here, certain phases of the model development lifecycle, such as data visualization and recovery, are (within limits) pushed into the database or memory instead of being performed on the analytical platform. This not only makes the analytical process faster, but also eliminates the time taken for data extraction and loading.

Making sure big data really adds value

Analyzing big data will cost money, and so it is important to measure the benefit gained. To do that accurately, you need to start from a reliable baseline: for example, the telecom company seeking to reduce churn needs to have a reliable picture of current churn before it starts, together with a measurement of the accuracy of its existing churn model. By continuing to measure churn, the organization can create a feedback loop, evaluating the success of the new techniques and changing them as necessary.

Strategic perspective

Most of all, it is important to approach this type of analytics from a strategic perspective, rather than starting with the technology or the data. The strategy should be developed to suit the enterprise as a whole, to avoid conflict between, or duplication of effort by, sales, marketing, and so on.

Decide what you need to find out first, and then work out whether, and how, big data can help you do it. A small pilot to find out what is achievable is often a key first step to realizing value from big data.

Making the most of social media

The social media explosion has seen consumers publishing their lives online via tweets, photo uploads, "likes", status updates, and so on. Much of this material is unstructured, has limited relevance to brands and exists in vast quantities, making it challenging to get value from the data. In working with social media, many organizations are encountering these big data challenges for the first time.

Companies that have invested in a social media monitoring (SMM) tool tend to be disappointed with the results. A common complaint is inaccuracy. Many SMM tools have access to only a fraction of available data, and struggle to classify content into (for example) positive, negative, neutral, or mixed categories. To compensate, organizations (or their agencies) often rely on humans to classify content manually – hardly a scalable approach.

So what is the best way to obtain business advantage from social media? We describe three important steps below: improve accuracy, work out what is relevant, and focus on action. We then discuss the need to incorporate social media into an overall data strategy.

Improve accuracy

100% accuracy is not achievable, but there are ways to improve on SMM tools – for example with natural language processing. Some human intervention will almost certainly be required, for example to "train" the software to filter out noise and interpret slang and sarcasm.

Semi-structured sources of social data are easier to analyze accurately. Numerical product ratings and "likes" can be aggregated to provide insight into how customers feel about products, and can be tied to a structured piece of data like a digital asset.

Work out what is relevant

To know which items in the sea of available data are relevant to your business, you need to integrate SMM with other data-related activities. If you are applying text analytics to social media data, then why not use it on other unstructured "verbatim" sources such as surveys and call center notes?

Add structured data to the mix and you are positioned to identify, for example, whether a spike in social media complaints is a sign of a new problem or relates to a known one. Once again, integrating different types of data together to get the whole picture is the key to value.

Human intervention is still required, and must be fast enough to keep up with social media storms or viral campaigns. For B2C organizations where reputation is vital, the aim should be to create something like an air traffic control room, with screens displaying and blending data from multiple sources to allow fast, informed decisions on what action is needed: product recall, competitive campaign.

Focus on action

Insights must generate actions. Empower your "air traffic controllers" to make decisions, and make sure those decisions are relayed to the parts of the organization that must act on them: customer service, product development, sales and so on.

For example, using a mixture of technology and people, Dell's social media command center monitors tens of thousands of conversations daily. According to Adam Brown, Dell's executive director of social media, its remit is to "help us understand which of these are relevant... which of these are things we really, really need to respond to"[3].

Incorporate social media into your overall data strategy

Seeing that a product is about to dive is of little value unless you can also see the impact on profitability, manufacturing plans, and other aspects of the business. Social media therefore needs to be linked to enterprise data and subject to the same data governance. Earlier sections of this report discussed tools and techniques like mastering and analysis that can integrate social media into enterprise data.

[3] See Dell's social media hub monitors and responds to online chatter, Ragan's HR Communication, 18 March 2011 – www.hrcommunication.com/Main/Articles/Dells_social_media_hub_monitors_and_responds_to_on_3608.aspx

Conclusion

Big data is a (probably overused) buzzword that means different things to different people. For some, the term immediately suggests Hadoop-type technology; however, the draw of big data for organizations is not technology but the opportunity to change their businesses. Recent research for Capgemini shows that, for processes where big data analytics has already been applied, organizations have seen an average performance improvement of 26% in the past three years.

Big data is about exploring and exploiting data in new ways. Often this will be data external to the organization – a largely untapped source of information that can radically change business performance.

One enabling factor for working with this data is the advent of distributed technologies for storage and sorting, or "marshaling" as our model calls it. The concept here is simple: think of SETI, an application devised in the 1990s to harness otherwise idle time on home PCs to assist astronomers with their search for extra-terrestrial life by processing data from telescopes[4]. Similarly, Hadoop and its like store vast volumes of data on hundreds of commodity hardware platforms, achieving new levels of scalability at relatively low cost. Managing large "data farms" presents some administrative headaches, but software vendors are finding solutions.

As we saw earlier, marshaling is just one of four major process steps in obtaining value from big data; the others are acquisition, analysis, and action – the last two being arguably the most critical to business advantage.

The primary technologies for acquisition come from the traditional business intelligence (BI) and enterprise content management (ECM) toolset. Long-established ETL solutions are being used for acquisition of big data, and will continue to play an important part. There are also new demands, such as the need to deal with real-time or near real-time feeds, and with semi-structured and unstructured data like voice, text, and video.

For the analysis step, familiar tools such as SAS and SPSS continue to play a major role. Newer high-end analytics have been around for some time, and have enormous predictive potential when applied to big data. The financial sector uses them to tailor offers to customers based on individual propensity to pay or default; online retailers' increasingly personalized marketing is perhaps a more conspicuous example.

As regards the final step – action – big data brings not only a high volume of new information but also a requirement to act on it instantly, which will often mean automatically. Changing onscreen offers based on the customer's last three clicks is not something that can wait for human interpretation and intervention, nor can re-routing of power supplies based on predicted local demand spikes. Business process management (BPM) and real-time decisioning help with these requirements.

Beyond the four process steps, big data will have far-reaching effects on the way we work. With limited

[4]http://setiathome.berkeley.edu

information and analysis, we used to "know" the answer, but with big data we, or our machines, will make decisions on the basis of the most probable answer out of a range of possibilities: a black-and-white world will need to be rethought in shades of grey. Organizations will need skills, particularly in statistics, that barely exist today.

Applying analytics to big data offers you the ability to see, or at least glimpse, the future. If you do not position your organization to do it, your competitors will. A company that gets ahead will find that its offer is far more effective, that it has greatly improved ability to provide the right service at the right place and time, and that its supply network is much cheaper. And that's the company that will be around tomorrow.

Big data should not be seen as an IT problem, but as proof that technology is no longer the limiting factor that it was 10 years ago. Now that we have the technology to handle exabytes of data, IT can do more for your business than you may realize.

It is time to encourage your key users to rethink the design of their business processes as if there were no technology limitations. Then you can set about using big data solutions to realize their ideas.

Together with cloud and mobility, big data is changing the rules of 21st century business. Invent it. Create it. Use it.

Capgemini Business Analytics

Capgemini's Business Analytics global practice network is a core unit within the Business Information Management (BIM) global service line and operates in 25 locations across the world, drawing on a database of over 100 analytics client credentials and analytical models. It provides high-function analytics-based solutions to all major industry sectors and business functions.

Capgemini has over 7,000 consultants working in BIM across the world. We work with all the leading big data and analytical technologies, and provide services to support business analytics, from high-level strategy to managed outsourced services. We recognize that analytics are specific to industry sector and sub-sector, and have experts and solutions across all of them.

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

www.capgemini.com

About Capgemini and the Collaborative Business Experience

With around 120,000 people in 40 countries, Capgemini is one of the world's foremost providers of consulting, technology and outsourcing services. The Group reported 2011 global revenues of EUR 9.7 billion.

Together with its clients, Capgemini creates and delivers business and technology solutions that fit their needs and drive the results they want.

A deeply multicultural organization, Capgemini has developed its own way of working, the Collaborative Business Experience™, and draws on Rightshore®, its worldwide delivery model.

More information about our services, offices and research is available at www.capgemini.com

For more information contact us at :
bim@capgemini.com or visit www.capgemini.com/bim