Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

BIM | the way we see it

Mastering Big Data

Why taking control of the little things matters when looking at the big picture

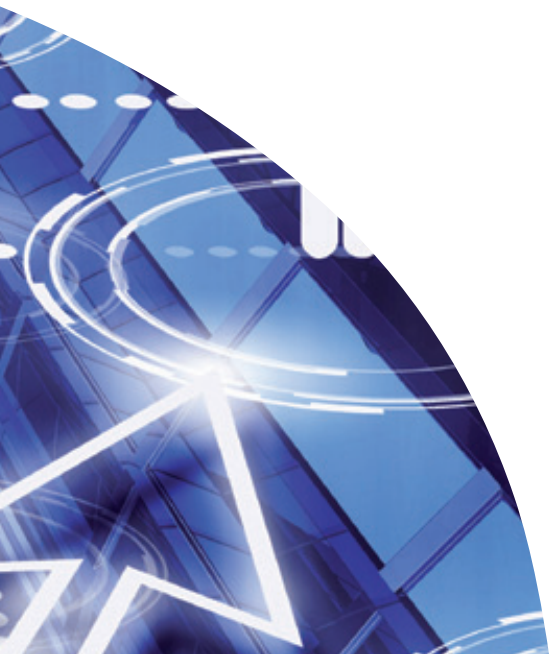
People matter, results count.

Big Data represents a big opportunity and a big reality

Many industry analysts and advisors are looking at Big Data as being the next frontier for competition and innovation. The first question, of course, is "what is Big Data?" The definition is hugely variable, but tends to refer to the massive explosion of information now available to organizations: the shift, for instance, of a retailer from tracking transactions to tracking the journey a customer takes around a store, and when they choose to buy. The breadth and depth of this data means that it can't simply be stored and queried within a traditional database solution.

The amount of data being created by organizations is increasing exponentially every year and is not something that companies can opt out of. The challenge is therefore to quickly identify what should be retained, avoid duplication where possible, and make use of the information that is being generated. Big Data is not about acquiring data from outside of an organization; it's about combining the Big Data being created internally with external Big Data sets.

Big Data is, therefore, about being able to access and leverage much larger information sets than ever before in order to gain greater insight into markets and opportunities. McKinsey talks about opportunities worth trillions for the overall exploitation of Big Data. At the heart of Big Data is the ability to ask questions about what will happen and receive more accurate answers than has been possible until now.



[1] McKinsey Global Institute, "Big Data: The next frontier for innovation, competition and productivity", May 2011

The Bigger the Data, the harder they fall

Let's imagine the information requirements of the future. We want to find clear trends on how people buy products, and we also want to add new information every day to keep our trends up-to-date.

We will have a few large data sets that gather information from the US over the last ten years, including:

1. Every retail transaction, obtained from Point of Sale (PoS) information
2. Weather history, by hour and city
3. Local sports team results by city
4. TV and other media advertising by region
5. Every online advert view by individual and location

Clearly these data sets are bigger than those being considered today, but it's this scale of challenge that companies will be looking to address in coming years. Before we leap into the world of distributed storage cloud computing, there are some basics that need to be realized.

Garbage In, Garbage Out rules

The first fact to realize is that any analytical model is only as good as the quality of information that you put into it. This means that you need sources of information that are trusted and known to be reliable. If you simply want a big data set and aren't worried about accuracy, then a random number generator will give you more than enough information.

Build the Islands of Certainty

The second key fact is that these differing data sets need common information to link them together so that, for example, a retail receipt for Lincoln, Nebraska can be linked to the weather report for that town. Within the data sets we also need accuracy. When someone is buying beer in New Jersey, are they buying the same brand of beer as someone in Connecticut or Maryland? And which products can be defined as "beer"; how well are the terms defined to ensure consistent results across locations?

These are the Islands of Certainty, a concept that Capgemini introduced in the "Mastering the Information Ocean" paper. Islands of Certainty are provided by understanding the points of reference before you create a large data set.

[2] Available from http://www.capgemini.com/insights-and-resources/by-publication/

Hanging Big Data from the POLE

The key to mastering Big Data is understanding where these data reference points are located. In our previous example, we have several reference points:

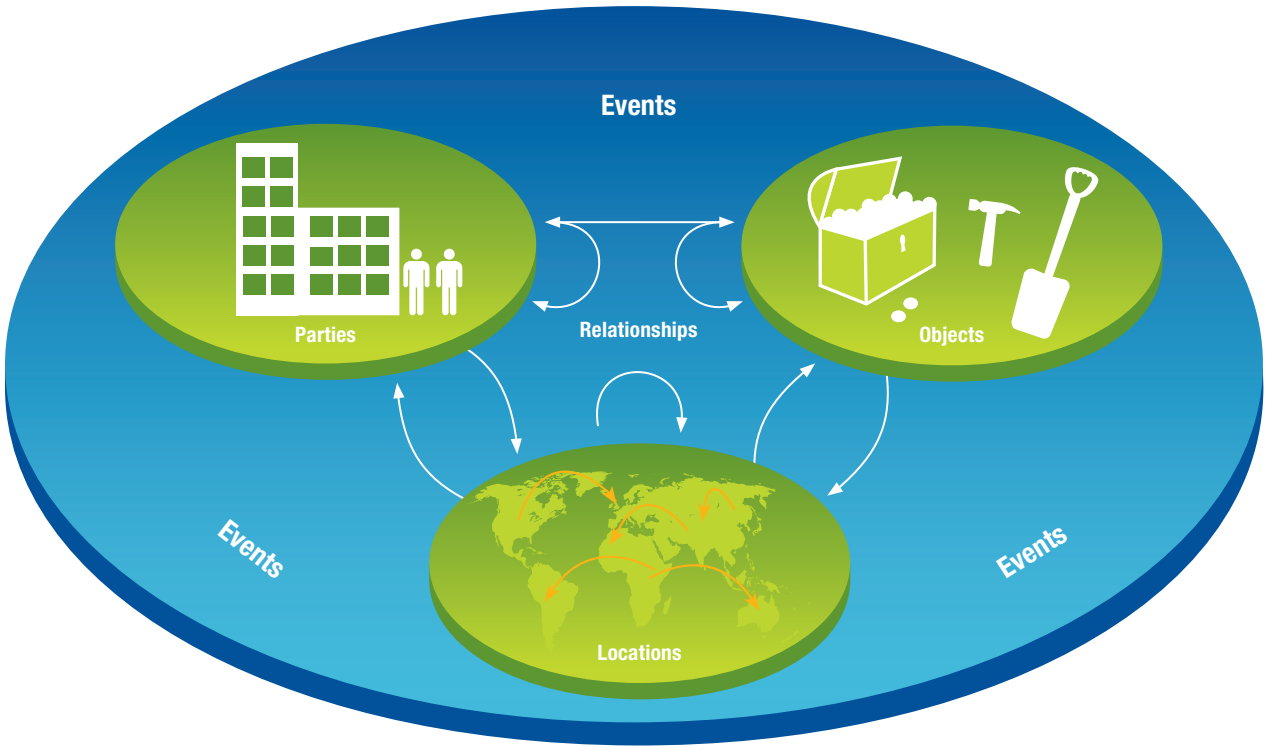
1. Customers – the individuals in the various interactions
2. Locations – where the customers are based and where they are buying
3. Products – what is being advertised and what is being sold
4. Team – the sports teams
5. Channels – the advertising media
6. Adverts – a linking of a product to a channel
7. Times – the calendar items when events happened

In addition to these key elements, there are the specific events that create the mass of data. We can structure this information as firstly a core of Parties, Objects and Locations, and secondly a volume of transactions or Events. We call this information model The POLE. The model is shown in figure 1.

It's by focusing on the P, O and L of the POLE that we can ensure Big Data actually delivers value. The POLE, therefore, acts as the structure around which Big Data is hung, in the manner of a skyscraper, from its internal infrastructure. Capgemini's blog post "MDM: Making Information Dance around the POLE" outlines this approach.

This means that there is a clear structure, based on the POLE, for our core entities, and from here it becomes possible to add the event information. Before we consider how to handle event information, we will take a look at the question of governance.

Figure 1 The POLE

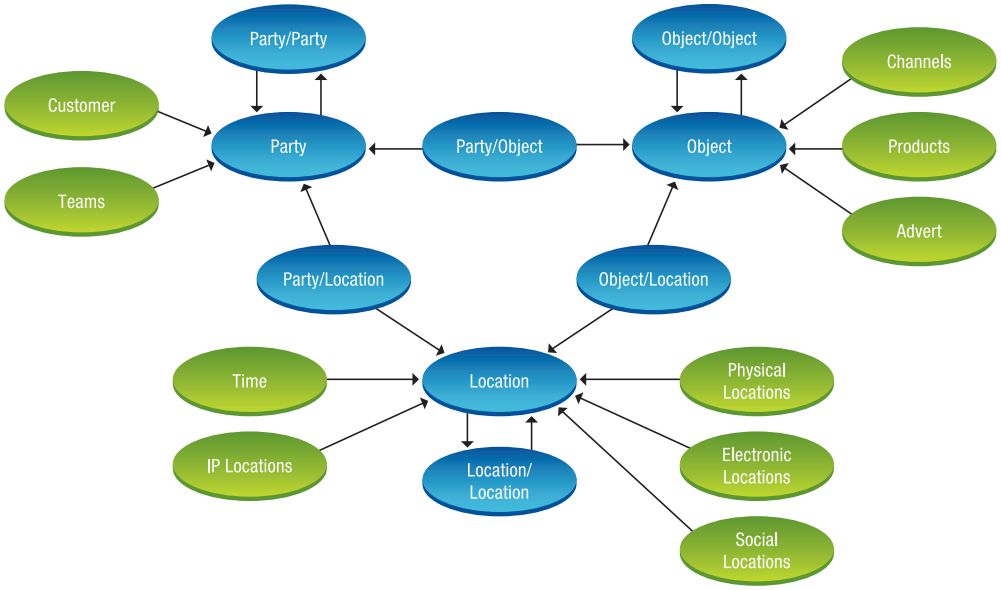


Governing the POLE

When looking at the task of mastering the POL for Big Data (figure 2), the challenges are exactly the same as those for mastering data generally within the business and between business partners. The number of entities is similar to those that an organization should already be mastering, and the practices are exactly the same.

However, the impact of not mastering information becomes exponentially greater in the case of Big Data. In our example, if we are unaware that cornflakes are sold in different size packets, for instance, then it is impossible to compare volumes sold between regions, or to compare prices accurately. If we don't even know that two different products are both cornflakes, then there is no ability to understand substitutions, and if we don't know that cornflakes are breakfast cereals, then we can't understand the broader category buying trends.

Figure 2 The POL of the POLE



With customer information it is important to be able to marry customers' online presence with their physical location and experiences, so as to be aware of the adverts they have seen; both traditional and digital. This means understanding the individuals across all channels, the products being advertised or placed on shelves, and the locations where all of these interactions happen.

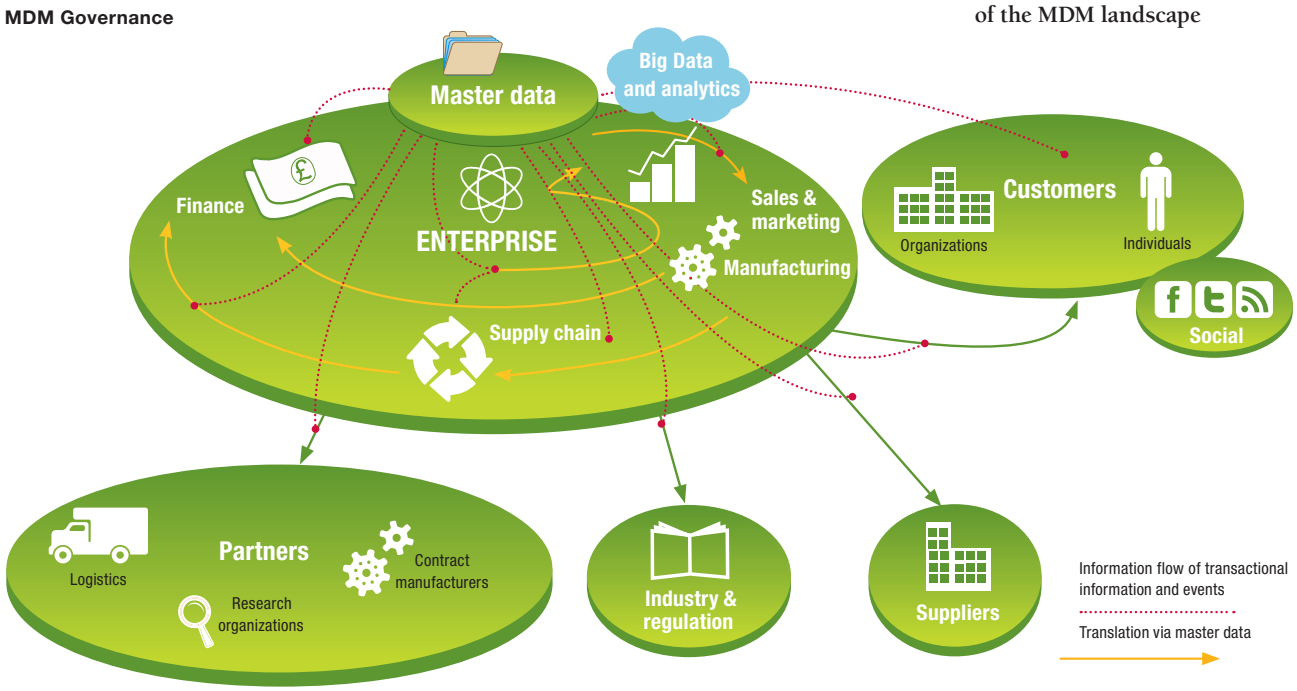
To achieve this you need organizational governance, which comes in two distinct types:

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

Governance is about taking control of the POLE in order to drive consistency across Big Data so it can deliver real insights (figure 3). Setting up this governance not only drives value within the analytical world of Big Data, but also enables the insights of Big Data to be directly tied to the operational aspects of the business.

Governance is not simply something for Big Data: it is something for the whole of the business that enables operations and analytics to work in sync. Governance applies not to a single information area but to the consistent elements that occur across all of the enterprise, including its external interactions.

Figure 3 Governance is at the center of the MDM landscape



Setting up governance for the POLE is standard practice in any sophisticated MDM organization. In our example there would probably be two groups of standards and policies:

1. **Customer-centric**
   a. Customer – handling the definition and standards around customers across all channels
   b. Product – handling the definition and standards around products across all channels
2. **Enterprise-centric**
   a. Locations – handling the definition and standards for locations, including how sports teams are assigned to regions and how advertising channels are defined within regions or across regions
   b. Weather – how weather information will be managed and tied to regions

Broad engagement across the business is needed in order to create standards that will work effectively and policies that can be implemented operationally on a broad basis.

Using the structure to load Big Data

Once we have defined the P, O and L of the POLE, and set up a clear governance structure that enables the business to match across the different information channels we can load our Big Data set-up, based around this framework.

Events – the individual transactional elements – are added to the model at this stage and their links to the core POL are used to associate information in a consistent way across channels (figure 4).

This is a reasonably standard approach within well-managed data warehouses, but with Big Data this level of rigor is essential. The challenge of Big Data is that many sources might be external to the organization during the transformation, so the matching processes might require significant processing power owing to the size of data.

Big Data scaling is often overlooked. It is necessary to scale the actual load as well as the analytics. Too often the focus is simply on the data-shifting challenge, rather than on ensuring that what is contained within the Big Data environment is of sufficient quality to enable any questions to be effectively answered.

The POL structure has now been augmented with events, giving us the full map of what we wish to query within Big Data, all hung from a single consistent structure. This framework-driven approach to Big Data management ensures that new dimensions and information sources can be rapidly adopted without requiring massive re-engineering of the solution.

Figure 4 Mastering Big Data

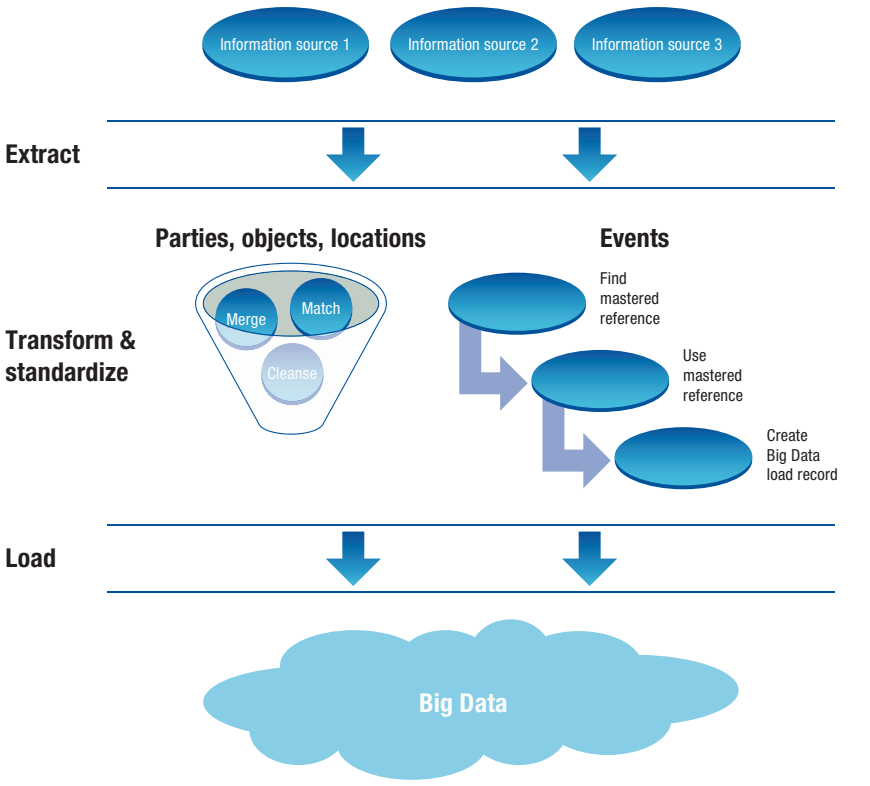
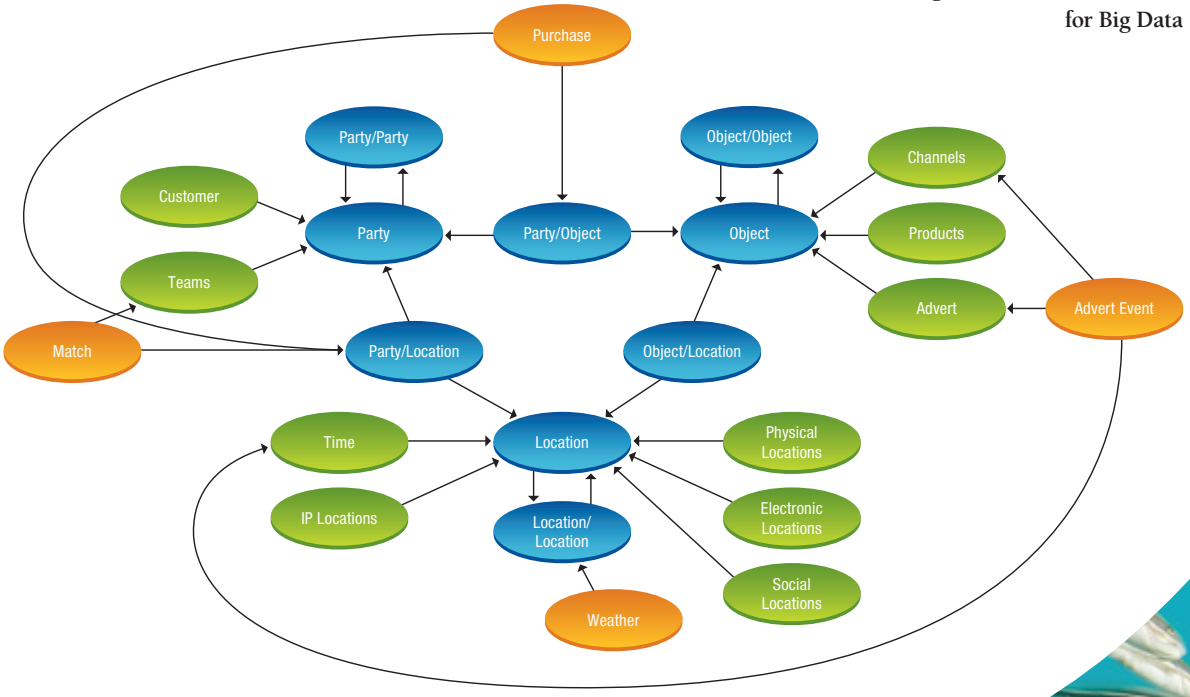


Figure 5 Cultural structure for Big Data



Securing Big Data – anonymizing information

One of the risks with Big Data, particularly when looking at customer information, is that any information leak or breach would be massive in scale. Using master data during loading gives a simple way of anonymizing the information.

Rather than loading the full customer profile information with information such as names and ages, the information can be anonymized to include the demographic segmentation of the individual but omit anything that could be deemed sensitive. In order to map this information back to the individual later, the Master Data Management (MDM) "key" could also be supplied for that record.

By employing these strategies it becomes possible for Big Data to answer challenging queries without loading sensitive information into the data set. Using MDM as a filter helps to deliver the value, control and security required to leverage Big Data effectively.

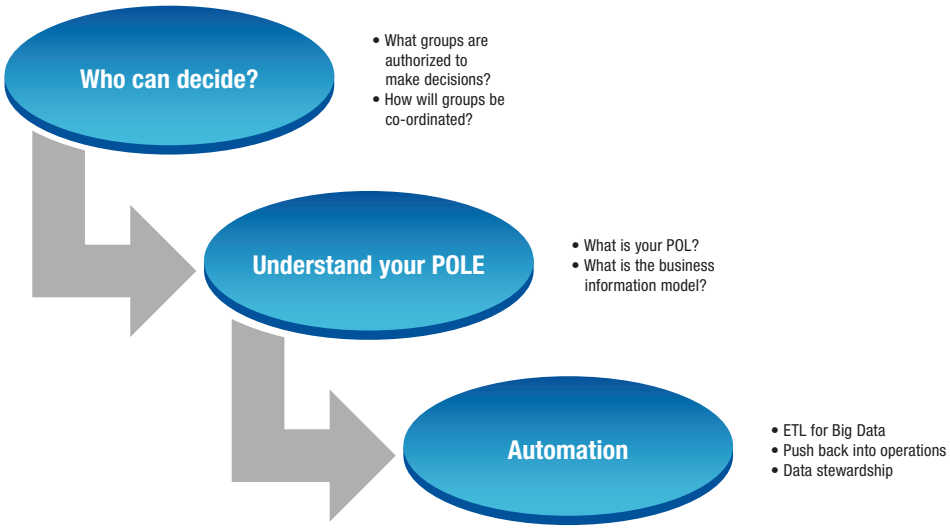
A simple plan to turn Big Data into Big Information

To get value out of Big Data, you need to turn the mass of disparate data into easily accessible, measurable information. This transformation can be achieved through a simple three-step approach, shown in Figure 6.

The first two of these steps are critical. The first step ensures that the correct entities have the authority to define the mappings, the second defines those mappings, and the third realizes them within both Big Data and operations.

Thirdly, technology is at the heart of automating this solution, ensuring a consistent level of quality by providing a hub that automates the distribution of information between the sources and Big Data.

Figure 6 Three steps to turn Big Data into Big Information



By starting with governance and business information, it becomes possible to construct a Big Data environment that is able to answer specific questions on historic data, as well as answer questions based on projected data by analyzing future trends.

In practical terms, this means that we would have first set up a governance team that could authorize any data standards and policies. This group works across the multiple organizations that are providing information to ensure there is consistency of definition and agreement about where any decisions about new entities, attributes or policies may be taken. Having an active group that is constantly reviewing and updating policies and standards ensures that, as Big Data sets expand and change, the organization is ready to adapt.

Plan for Big Data by taking control

A simple philosophy lies at the heart of this paper: that the complexities of BIG Data can be easily managed and exploited by dividing the project up into a series of disciplines and recompiling it around a clearly defined structure. Information becomes easier to access, manage, and far more secure. By setting up the governance of master data, you provide Big Data with the framework required to drive its long-term success, and do so in a way that ensures both the accuracy of results and their direct correlation to operations. There is a clear case for effective MDM as the first step towards leveraging Big Data, especially given its added benefits for information security. Central to an organizations approach to mastering Big Data is which parts of the POLE model deliver the benefits to master at a given point in time.

For most organizations, the mastering of the POLE is an iterative process, often driven by a business function that understands the benefits that mastering the information will deliver. As a basic rule, it is only worth mastering information at the point at which you wish to have quality information for analytical purposes. If it is sufficient for information to simply provide trends rather than an accurate holistic view then mastering should be delayed until it has a clear ROI. By taking an iterative approach to mastering the POLE it becomes possible for organizations to continually realize more and more benefits from Big Data as the ability to leverage and understand the complex analytical models improves.

Big Data is a powerful tool, but power is nothing without control. MDM provides business with the tools and ability to realize the power of Big Data.

"By taking an iterative approach to mastering the POLE it becomes possible for organizations to continually realize more and more benefits from Big Data as the ability to leverage and understand the complex analytical models improves."

Capgemini
CONSULTING.TECHNOLOGY.OUTSOURCING

www.capgemini.com/bim

About Capgemini

With more than 115,000 people in 40 countries, Capgemini is one of the world's foremost providers of consulting, technology and outsourcing services. The Group reported 2010 global revenues of EUR 8.7 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 is available at **www.capgemini.com**

Rightshore® is a trademark belonging to Capgemini

To find out more about Master Data Management and Big Data visit **www.capgemini.com/bim** or email: **bim@capgemini.com**

People matter, results count.