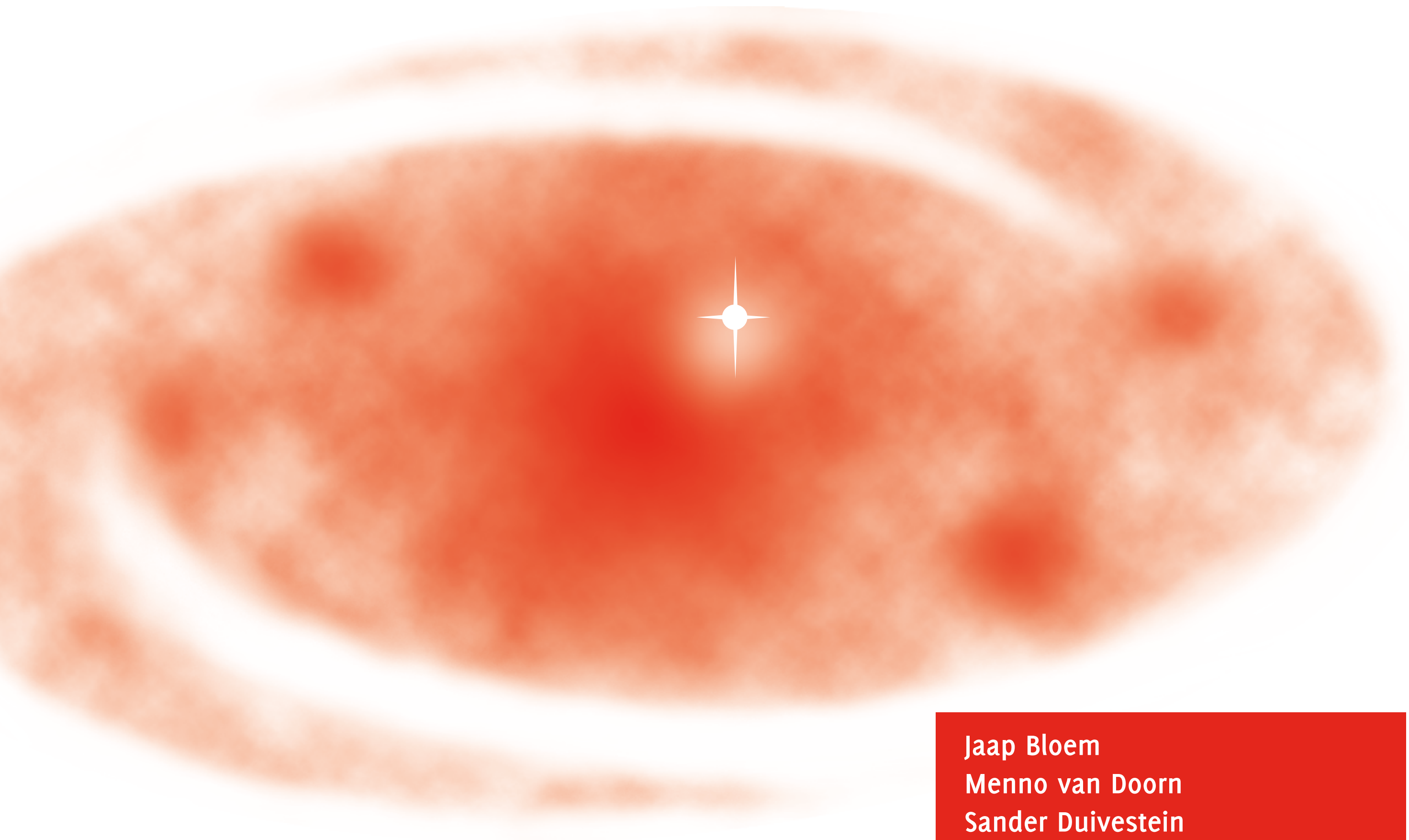
VINT research report: ① of 4
VINT research report: ② of 4
VINT research report: ③ of 4
VINT research report: ④ of 4

Creating clarity with Big Data



Jaap Bloem
Menno van Doorn
Sander Duivestein
Thomas van Manen
Erik van Ommeren

Participate in our Big Data discussion at
www.sogeti.com/vint/bigdata/questions

SOGETI

VINT | Vision • Inspiration • Navigation • Trends

vint.sogeti.com
vint@sogeti.nl

Table of contents

© 2012 The Sogeti Trend Lab VINT | Vision | Inspiration | Navigation | Trends
LINE UP boek en media bv, Groningen, the Netherlands

SOGETI
VINT | Vision • Inspiration • Navigation • Trends

The VINT Big Data research reports

Since 2005, when the term “Big Data” was launched – by O’Reilly Media remarkably enough, which had issued Web 2.0 a year previously – Big Data has become an increasingly topical theme. In terms of technological development and business adoption, the domain of Big Data has made powerful advances; and that is putting it mildly.

In this initial report on Big Data, the first of four, we give answers to questions concerning what exactly Big Data is, where it differs from existing data classification, how the transformative potential of Big Data can be estimated, and what the current situation (2012) is with regard to adoption and planning.

VINT attempts to create clarity in these developments by presenting experiences and visions in perspective: objectively and laced with examples. But not all answers, not by a long way, are readily available. Indeed, more questions will arise – about the road map, for example, that you wish to use for Big Data. Or about governance. Or about the way you may have to revamp your organization. About the privacy issues that Big Data raises, such as those involving social analytics. And about the structures that new algorithms and systems will probably bring us.

The new data focus is a quest involving many issues, at the start of and certainly during the entire journey. That is why we will be pleased to exchange thoughts with you: online, at www.sogeti.com/vint/bigdata/questions, and in personal conversations, of course. By actively participating in the discussion, you can sharpen your (and our) ideas with respect to Big Data and come to progressive insights for taking lucid and responsible decisions. In this way, we jointly determine the concrete substantiation of the coming three research reports after this kick-off on the topic of Big Data.

Join the conversation

For inspiration, we have included seven questions to which we would very much like to receive a response or, rather, your opinion. You can click on the relevant buttons in the PDF version of this document. You will then be directly transported to the discussion in question.

1 Digital data as the new industrial revolution

In 2012, approximately forty years after the beginning of the information era, all eyes are now on its basis: digital data. This may not seem very exciting, but the influx of various data types, plus the speed with which the trend will continue, probably into infinity, is certainly striking. Data, data and more data: we are at the centre of an expanding data universe, full of undiscovered connections. This is not abstract and general, but rather specific and concrete, as each new insight may be the entrance to a gold mine. This data explosion is so simple and fundamental that Joe Hellerstein of Berkeley University speaks of 'a new industrial revolution': a revolution on the basis of digital data that form the engine of completely new business-operational and societal opportunities.

At the beginning of May 2012, at the Cloud Computing Conference arranged by Goldman Sachs, Shaun Connolly from Hortonworks presented data as "The New Competitive Advantage." Connolly articulated seven reasons for this statement, two of which were business-oriented, three were technological, and two were financial:

Business reasons

1. New innovative business models become possible
2. New insights arise that give competitive advantages

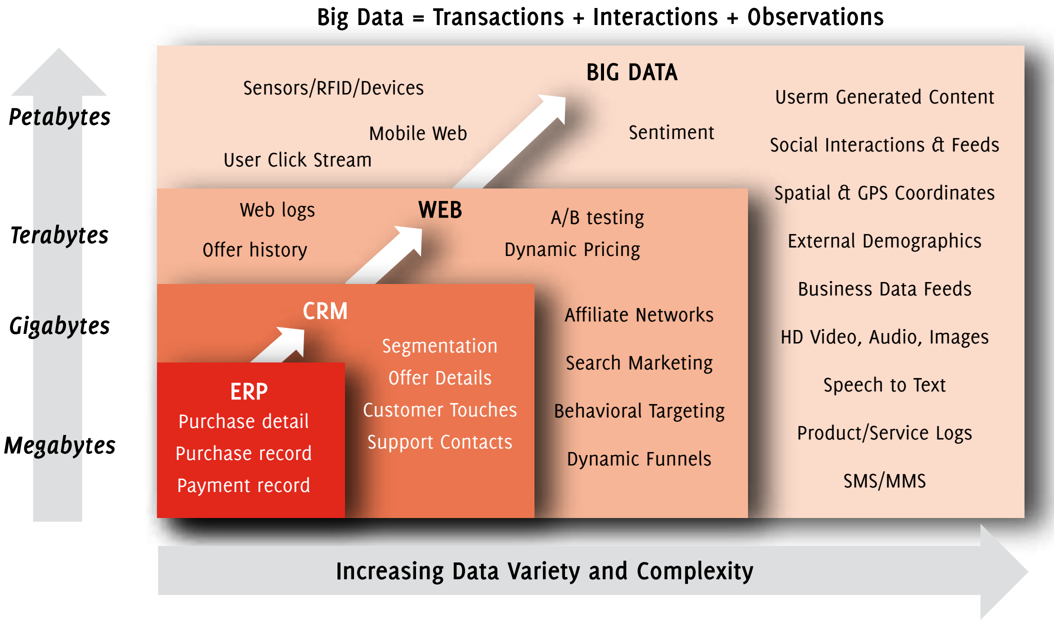
Technological reasons

3. The generation and storage of data continue to grow exponentially
4. We find data in various forms everywhere
5. Traditional solutions do not meet new demands regarding complexity

Financial reasons

6. The costs of data systems continue to rise as a percentage of the IT budget
7. New standard hardware and open-source software offer cost benefits

Connolly believes that, as a consequence of this combination, the traditional data world of business transactions is now beginning to merge with that of interactions and observations. Applying the formula *Big Data = Transactions + Interactions + Observations*, the goal is now: more business, higher productivity and new commercial opportunities.



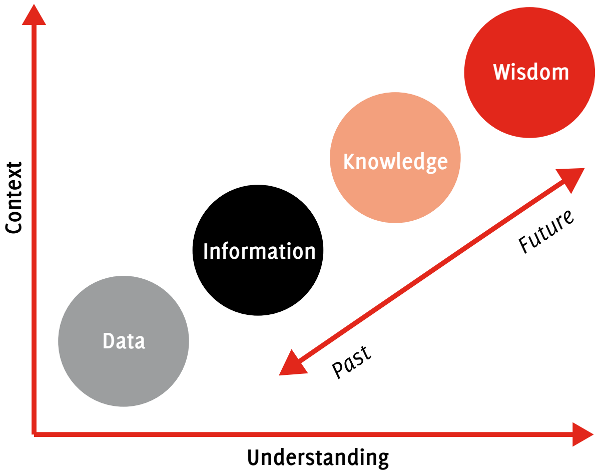
Source: Contents of above graphic created in partnership with Teradata, Inc.

Digital data as the basis

At present we are living in at least three periods with digital data as their basis: the information era, the social era, and the Big Data era. This is what is stated in Wikipedia's *List of Periods*, which covers the entire history of humankind. The explosive growth of data genuinely comes from all corners: from business transactions, mobile devices, sensors, social and traditional media, H-D video, cloud computing, stock-and-share markets, Web-clicks, etc., etc. All data is generated in the interaction between people, machines, applications and combinations of these. Those who have difficulty in grasping all this ought to take a look at a publicly accessible corner of our new data universe: the Linked Open Data domain at http://lod-cloud.net. The visualization of that data network and its components immediately clarifies what is going on in the world, in all sectors of the economy, society and science, and also in a combination of these.

Everything is information

Organizations exist thanks to information, and within the realm of science nowadays there is a movement that claims that, in fact, everything is information. Data forms the fundament of this information, and the more relevant facts we have, the better we can understand the most diverse issues, and the better we can anticipate the future. This is necessary in order to be able to take the correct decisions, certainly in these times of hyper-competition and crisis. The unprecedented data intensity of the Big Data age that we have just entered, ironically at this crisis-ridden moment, is nevertheless a blessing, say the proponents. After all, analysis of complete datasets is, by definition, the only real way to be able to fully comprehend and predict the situation.



Join the conversation

Question 1
Will facts now definitively defeat intuition?

www.sogeti.com/vint/r1q1

This argument has no leaks, and thanks to modern and affordable IT – hardware, software, networks, algorithms and applications – analysis of complete datasets can now genuinely take off.

Big Data case: loss of clients

Until recently we were compelled to take random samples and analyze them. But how do you sample a network or a collection of sub-networks? If a telecom provider wishes to have insight into the circumstances under which a sub-network of friends and acquaintances suddenly switches to a rival company (it "churns"), we are probably dealing with a total of more than 10 million existing and recent subscribers, with information on their habits, their expenditures on services, and who their friends are: in other words, the number of times the phone is used for calls or SMS messages, for example. We are dealing with tipping points: a part of the sub-network churns and the rest follow after a (short) time. In itself, this is rather predictable: if colleagues or friends have switched and are better off or cheaper out under a rival, then there is a social and economic stimulus to switch as well. A provider will, of course, attempt to prevent this situation arising and must take a hard look at all the data. For example, if a random sample is taken from a million clients, the circles of friends that formed the basis of the switch can no longer be seen as a unit, and therefore in this case the basis for accurate prediction crumbles. Therefore, sampling is not the appropriate method here. In order to obtain a good view of the tipping points we must examine all the data in their proper context and coherence. Then, on the basis of developing patterns, we can anticipate their churn at an early stage and apply retention actions and programs.

Detection of fraud

Another area for which we require a complete dataset is fraud detection. The signal is so small that it is impossible to work with random samples until the signal has been identified. Accordingly, all data must be analyzed in this field as well. It can justifiably be referred to as an evident case of Big Data when the possibility of 'collusion' is being examined: illegal collaboration that is directed toward impeding others as much as possible and of sabotaging them, as occurred in the casino world. Churn and fraud detection are examples of the application possibilities of Big Data Analytics (see also Section 7).

Big Data Success Stories

In October 2011, under the title *Big Data Success Stories*, IBM published an illustrative reader with twelve different case studies, to demonstrate what Big Data actually signifies. We shall also respond to that issue here, in the following section and in Section 7, "Big Data in organizations in the year 2012." For the moment we shall simply work on the assumption that Big Data Analysis goes further than what traditional relational databases can offer, and that current trends are moving toward the use of an increasing number of new data types. With all the empirical data that are there for the taking, it seems as if, in the future, we will only need to examine the facts in a smart way so that theory and model-forming, as intermediate steps, can ultimately be skipped. This Big Data promise was articulated as far back as 2008, in an article entitled "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete".

2 Total data management in each organization

Big Data, the enormous data expansion that is spreading rapidly in all respects, demands total data management in every organization. This fact has been underlined by many experts, including The 451 Group.

An increasing quantity of data is arriving from all kinds of sources: from traditional transactional data to sensors and RFID tags, not forgetting the social media, Internet, clouds and mobile devices. It no longer matters whether data is structured, semi-structured or unstructured, as current IT possibilities for data acquisition and processing, and their affordability are thriving at the same time.

Data growth surpasses Moore's Law

Although the flood of data now exceeds Moore's Law – every 18 months a doubling of the number of transistors per square inch on integrated circuits takes place against increasingly lower cost – we are still able to deal with this in a meaningful way. This is possible due to advanced hardware, software, networks and data technologies. In short, we are capable, along with everyone in the organization, of exploiting the entire data field. Anyone who can do this well, stated Gartner in their presentation entitled

Information Management Goes "Extreme": The Biggest Challenges for 21st Century CIOs, can achieve 20% better than competitors who do not do so:

> *Through 2015, organizations integrating high-value, diverse new information types and sources into a coherent information management infrastructure will outperform their industry peers financially by more than 20%*

The rules of the game remain the same, but the tactics have changed. Just as in bygone days, we wish to process information from raw data and extract intelligent new insights that enable better and faster business decisions. Big Data is a kind of appeal to organizations to elevate their Business-Intelligence efforts to a radically higher level: on the basis of the appropriate technology, the proper processes, the right roles and the relevant knowledge and expertise, called 'Data Science'. This ought to occur throughout the entire organization and constantly.

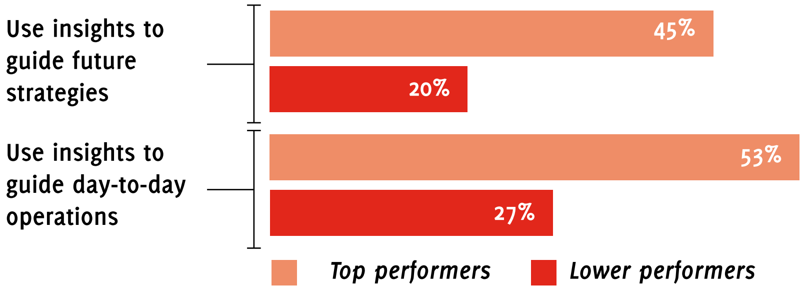
Big Data is a new phase

This being the case, Big or Total Data constitutes a new phase in the trend that was quantified by MIT Sloan Management Review and the IBM Institute for Business Value in 2010, in their study entitled *Analytics: The New Path to Value*. Almost half of the best-achieving organizations, it turned out, used digital data for their long-term strategies, in contrast to only a fifth of the underperformers. With regard to daily operations, this was even more than half of the top-performers against slightly more than a quarter of the poorly achieving organizations. The conclusion must be drawn that priority must be given to an analysis of the full array of digital data.

Join the conversation

Question 2
How do you link real-time Big Data to the operational supervision of your company?

www.sogeti.com/vint/r1q2



Of course, organizations do not wish to discard this type of advice, all the more because it builds logically upon existing Business Intelligence and the target of economic profit. But various demands and requirements must be dealt with and put in place. In addition to the potential and promise of Big Data, we shall also cover this aspect in our research report. The ambition of all Big Data reports is to exchange thoughts with you on the topic of this important subject matter, and to jointly explore the possibilities.

3 Participate in our Big Data discussion at www.sogeti.com/vint/bigdata/questions

The Big Data issues about which we would like to exchange ideas and experiences, on the basis of the research report before you, are threefold:

A. Your Big Data profile: what does that look like?
B. Ten Big Data management challenges: what are your issues?
C. Five requirements for your Big Data project: are you ready?

Note:

Interaction on this and related matters occurs on our website, and also face-to-face as far as we are concerned. We shall share new research insights with you on a weekly basis, via blog posts, e-mail and Twitter alerts. The accompanying video material, presenting leading experts, is intended as inspiration to think through and discuss the entire theme of Big Data from various angles.

A. Your Big Data profile: what does that look like?

Big Data is concerned with exceptionally large, often widespread bundles of semi-structured or unstructured data. In addition, they are often incomplete and not readily accessible. "Exceptionally large" means the following, measured against the extreme boundaries of current standard IT and relational databases: petabytes of data or more, millions of people or more, billions of records or more, and a complex combination of all these. With fewer data and greater complexity, you will encounter a serious Big Data challenge, certainly if your tools, knowledge and expertise are not fully up to date. Moreover, if this is the case, you are not prepared for future data developments. Semi-structured or unstructured means that the connections between data elements are not clear, and probabilities will have to be determined.

B. Ten Big Data management challenges: what are your issues?

1. How are you coping with the growing quantities of semi-structured and unstructured data? It has been estimated that 80 per cent of the data most valuable to organizations are located outside the traditional relational data-warehousing and data-mining to which Business Intelligence has been primarily oriented until now.
2. Those new valuable data come from a range of different data types and data sources. Do you know which of these are important for your business and do you have a plan to apply them strategically?

3. Do you have an overall view of the complexity of your data, either independently or in combination? And do you know what exactly you want to know in which order of sequence. Now and in the future?
4. New insights obtained from the combination of structured and unstructured data may have an imminent expiry date. Are you aware of the desired speed of processing and analyzing various data and data combinations? Which issues that you might wish to solve require a real-time approach? Please keep in mind that Real-time processes are needed to enable Real-time decisions.
5. Have you thought about the costs of your new data management? How are they structured: according to data domains, technology and expertise, for instance?
6. The storage of all data that you wish to analyze and stockpile will probably make new demands upon your IT structure. Do you have any kind of plan to deal with this, and are you also watching performance?
7. What is the state of your data security system?
8. The storage and security of Big Data is of major importance with regard to your data governance, risk management and compliance. Are you involving the appropriate departments and functionaries in your Big Data activities?
9. Generating new business insights from large quantities of data requires an organization-wide approach. New knowledge and expertise are needed for this. Are they available in your organization and how can these be guaranteed and further developed?
10. Do you know what your Big or Total Data efforts mean for your energy use?

C. Five requirements for your Big Data project: are you ready?

On the basis of the above-listed management challenges, we now summarize five fundamental conditions that are collectively needed in order for you to begin confidently on a concrete Big Data project:

1. Your organization has at its disposal the right mindset and culture. There is no doubt, throughout the whole organization, about the usefulness of a Big or Total Data approach, you know where you want to begin, and what the targets for the future are.
2. There is sufficient management support and it is evident who the executive sponsors are.
3. The required expertise and experience with regard to Data Science and Big Data frameworks and tools are available and guaranteed.
4. Sufficient budget has been allocated for the necessary training, in order to ensure that the required expertise and experience, mindset and culture will bond.
5. There are adequate resources and budget for the development of Big Data applications, and you have selected the right partners and suppliers in this context.

Join the conversation

Question 3

What is the best approach to capture positive attention among the management?

www.sogeti.com/vint/r1q3

4 Why the word "Big"?

We refer to something as "big" – Big Mac for example – to draw attention to its volume. But if we supply no relevant image, the word "big" immediately evokes fundamental questions. That is also exactly the case with Big Data, and also with the related Big Science. How large is Big Data actually, and in relation to what?

"Big" is not a particularly handy term

Accordingly, the analysts at Forrester and at Gartner agree completely with this statement: in retrospect, "big" is perhaps not a convenient name for the flood of data that is increasing at an enormous pace. Both offices, and others with them, prefer to use "extreme" rather than "big." That term also has a longer history in the field of statistics.

In everyday life, "big" refers to very concrete oversize phenomena. But inconceivably high quantities of digital data are not perceived by the eye. In addition, more is happening than "quantity" alone.

Big Data and Web 2.0

It is no coincidence that O'Reilly Media introduced the term "Big Data" a year after Web 2.0 appeared, as many valuable Big Data situations are indeed related to consumer conduct. Web 2.0 provided the impulse to rethink the interaction that was taking place on Internet, and to push it somewhat further. In much the same way, the qualification "Big Data" calls attention to the business possibilities of the flood of data on the one hand, and the new technologies, techniques and methods that are directed toward these, on the other.

A simple answer

As mentioned, the increase in data has now exceeded Moore's Law. Various types of data in combination with the necessary speed of analysis now form the greatest challenge, although we must not forget the limited number of people who can deal proficiently with Big Data. In 2020, there will be 35 zettabytes of digital data. That represents a stack of DVDs that would reach half way from the Earth to Mars. Face-book has 70 petabytes and 2700 multiprocessor nodes. The Bing search engine has 150 petabytes and 40,000 nodes. But what does Big Data exactly signify for organizations? We can approach Big Data from the standpoint of the issues, but also from the standpoint of the solutions. The simplest response comes from Forrester Research and is as follows:

> *Big Data: Techniques and Technologies that Make Handling Data at Extreme Scale Economical.*

Just like The 451 Group and Gartner, Forrester also makes no distinction between Big and Little Data. Compared to bygone times, many new and different data have arrived on the scene, and this is an ongoing process; but data remain data. They go hand in hand, and we can only truly advance further if there is well-thought-out integration of the whole spectrum of various orders of magnitude. We are dealing with a single data spectrum, a single continuum, and that is what organizations ought to be strategically exploring step by step.

One large data continuum

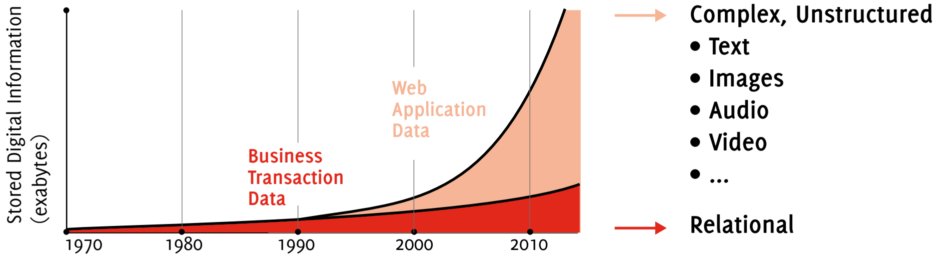
Around thirty years ago, this also applied to the growth of scientific activity: large and small. In his book entitled *Reflections on Big Science* (1967), the atomic scientist Alvin Weinberg wrote:

> *The scientific enterprise, both Little Science and Big Science, has grown explosively and has become very much more complicated.*

This observation referred to science at that time, and it now refers precisely to what is happening in the realm of data. Check what Chirag Metha has to say. As a Technology, Design & Innovation Strategist, Metha was associated with the Office of the CEO at SAP:

> *Today, technology — commodity hardware and sophisticated software to leverage this hardware — changes the way people think about small and large data.* ***It's a data continuum.*** *[...] Big Data is an amalgamation of a few trends – data growth of a magnitude or two, external data more valuable than internal data, and shift in computing business models. [...]* ***Big Data is about redefining what data actually means to you.*** *[...] This is not about technology. This is about a completely new way of doing business where data finally gets the driver's seat.*

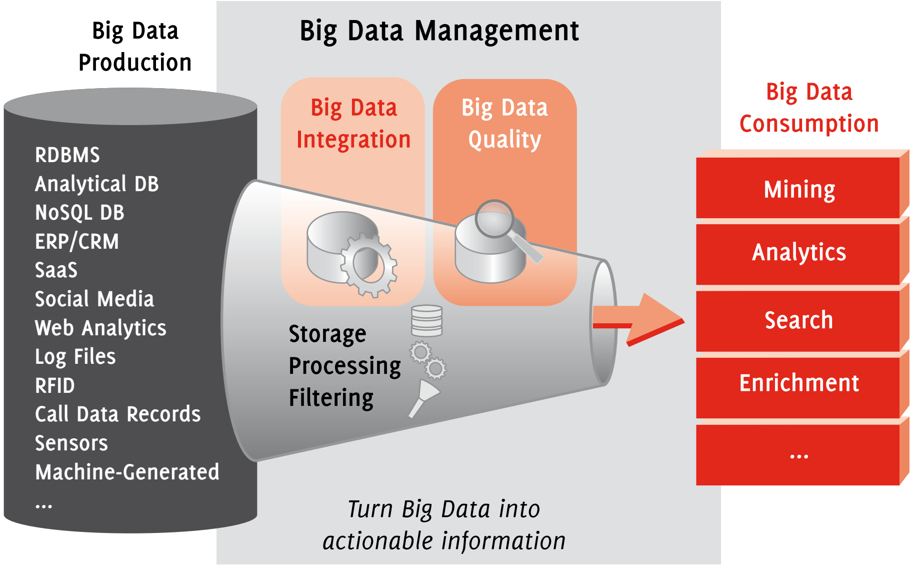
Big Data does not at all mean to say that we ought to forget Little or Small Data, or Medium, Large etc. On the contrary, it is important that we can and must review all the data in all their forms. It is possible technologically, and desirable, if not essential, businesswise.



This is particularly the case because 80 per cent of all new data is not relational or is unstructured and, in combination with transaction data, contains the most valuable information for organizations. In the view of some people, not all data that initially seem unstructured need to remain so, not by a long way, and indeed such data can be accommodated within a structure with relatively little difficulty.

5 The importance of Big Data

The reason why we should wish to have and examine all that data is evident. The social media, web analytics, logfiles, sensors, and suchlike all provide valuable information, while the cost of IT solutions continues to drop and computer-processing power is increasing. With developments like these, the surplus of information seems to have largely vanished: in principle, organizations are now capable of managing the flood of data and to use it to their own (financial) advantage. Those who excel in acquiring, processing, and managing valuable data, says Gartner, will be able to realize a 20% better result, in financial terms, than their competitors.



Within organizations, the share of unstructured data, such as documents, e-mail and images, is around 60 to 80 per cent. Of all data analyses that currently take place in organizations, 5 to 15 per cent contain a social component that enriches the structured data. This number must increase, not least because of all the external data that can be included in the analyses.

The Internet of Things is also becoming an increasingly rich source of data. At this moment, says Cisco CTO Padmasree Warrior, there are 13 billion devices connected to the Internet and that will be 50 billion in 2020. IDC expects more than 1 billion sensors to be connected to the Internet by that time. All the accompanying data flows can supply interesting insights that can aid better business decisions.

Join the conversation

Question 4
For organizations, what is the most important new rule of play with regard to Big Data?

www.sogeti.com/vint/r1q4

We are at Big Data's point of departure

Banks belong to the top of the organizations that are engaged with Big Data but, in the report with the eloquent title *Big Data: Harnessing a Game-changing Asset* by the Economist Intelligence Unit, Paul Scholten, COO Retail & Private Banking at ABN AMRO, candidly admits that the bank is in an exploratory phase when it comes to making good use of unstructured social data in particular:

> *We are used to structured, financial data. [...] We are not so good at the unstructured stuff. [...] The company is just beginning to understand the uses of social media, and what might be possible in terms of improving customer service.*

Mark Thiele states that it is interesting to compare Big Data in the year 2012 with the start of the World Wide Web. Thiele is the Executive VP Data Center Technology at Switch, the operator of the SuperNAP data center in Las Vegas, the largest and most powerful of its type in the world:

> *Big Data today is what the web was in 1993. We knew the web was something and that it might get Big, but few of us really understood what "Big" meant. Today, we aren't even scratching the surface of the Big Bata opportunity.*

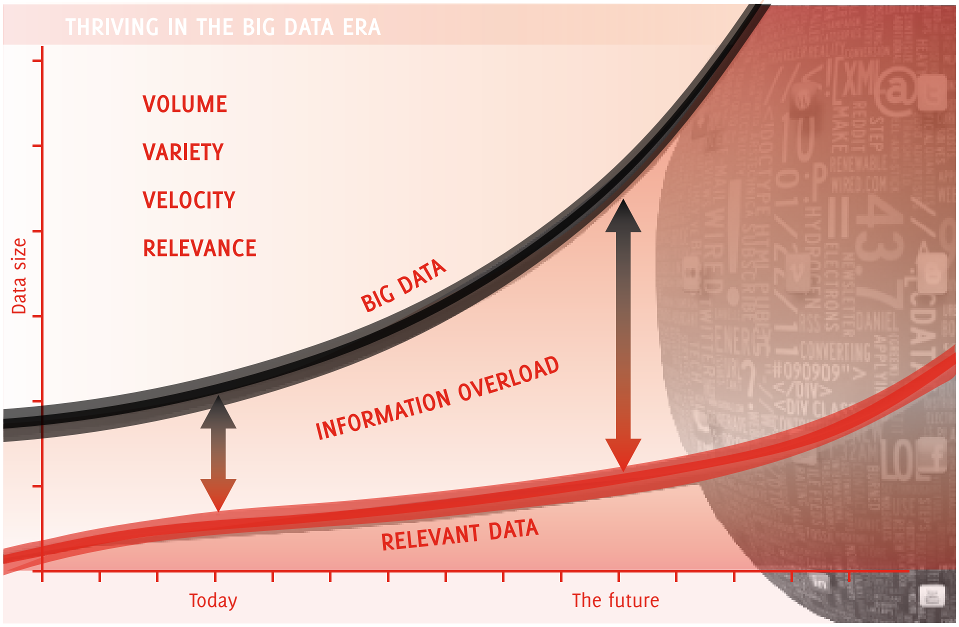
No isolated phenomenon

If there is one thing that has become clear, that is the fact that Big Data is not an isolated phenomenon. Moreover, the word "big" emphasizes the quantitative aspect. Fortunately, this immediately raises the necessary questions, so that we are compelled to think more profoundly about Big Data.

In March 2012, Credit Suisse Equity Research published the report entitled *The Apps Revolution Manifesto, Volume 1: The Technologies*. The authors regard, in particular, the convergence of Service-Oriented Architecture, Cloud, Fast Data, Big Data, Social and Mobile as being determinative of the value that new enterprise applications can provide. Credit Suisse Equity Research estimates this development to be just as transformative as the client/server and web applications were in the past.

Volume, Variety, Velocity

As far back as 2001, Doug Laney made clear – then at the META Group and nowadays at Gartner – that three factors can influence one another in the growth of data flow: the quantity of data *(Volume)*, the nature of the data type: structured, semi-structured and unstructured *(Variety)* and the desired analysis rate *(Velocity)*. Nowadays we often add *Complexity*, *Value* and *Relevance* to this list. The last two are included because we would like to know what we can and want to do with all the data, so that we are not investing time, money and effort for no return.

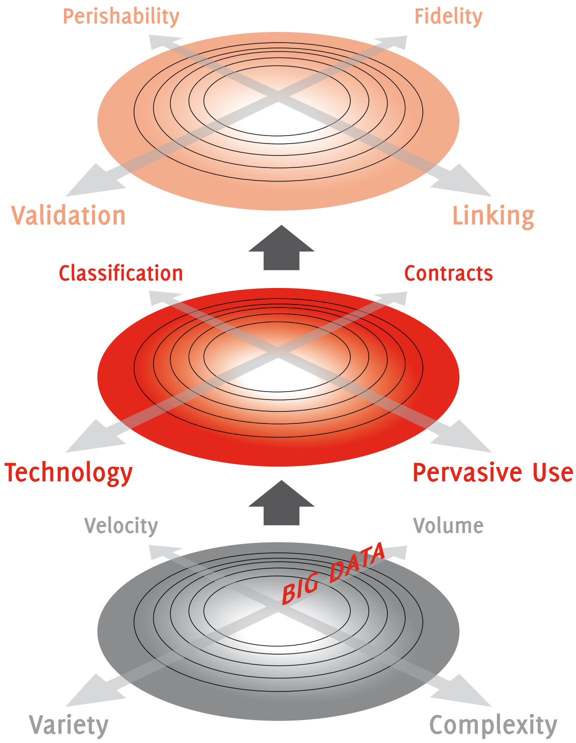


Big Data as the Next Frontier

On that basis, predicts the McKinsey Global Institute in its report entitled *Big Data: The Next Frontier for Innovation, Competition and Productivity*, the right exploitation of Big Data can produce hundreds of billions of dollars for various sectors of the American economy. McKinsey underlines the great sectoral differences (see Section 11) with respect to the ease with which Big Data can be acquired, set against the value that the use of Big Data can be expected to produce. It further emphasizes the necessity of eradicating the knowledge gap in organizations, with regard to dealing with (Big) Data (see Section 10).

6 Big Data is Extreme Information Management

Gartner has now elaborated the basic model of Volume, Variety and Velocity into the three interactive layers, each with four dimensions (as shown in the illustration below). The resulting twelve factors dovetail together and must all be purposefully addressed in the information management of the 21 st century: separately and as a whole.



In short, here we have, moving from the bottom to the top, the following: departing from the variety and complexity, in particular, of an increasing amount of data – often also in real-time – it is very possible to express validated statements and to establish connections on the basis of correct technological applications in combination with intensive input of all data, in order to elevate business decision making to a qualitatively higher level.

If we take Big Data as the point of departure, this should be on the volume side, as the name indicates. Variety and speed are the other dimensions at that level, in Doug Laney's view. An extra addition is the complexity of not only the data but also of the 'use cases': the way in which all data is brought into association by means of relevant and constructive questioning. We have already presented a concrete typology on the basis of the formula *Big Data = Transactions + Interactions + Observations* in Section 1.

The intermediate level is concerned with access and control. To start with, there are always agreements (*Contracts*) about which (*Classification*) information should be recorded and how it can be used. The social media and cloud computing provide splendid opportunities, but new technology (*Technology*) is needed to ensure that the data can be used everywhere and at any time (*Pervasive use*).

The top layer covers the reliability of information (*Validation, Fidelity*). It must be not only relevant and accurate when acquired (*Perishability*), but also in the use case. It is also important whether or not enrichment occurs in combination with other information (*Linking*).

Altogether, in a Big Data context, organizations must respond to the six well-known standard questions: what, when, why, where, who and how? The first four cover the structure of your Enterprise Information Architecture and the last two that of your Enterprise Information Management.

What? What are the correct data and information?
When? What are their ideal lifecycle stages?
Why? What are the right characteristics?
Where? What are the proper interfaces for interaction?
Who? What are the right roles in the organization?
How? What are the right information activities?

This is the concretization that belongs to the standard questions, in a nutshell. These questions serve as a guideline for the further structuring of Big Data, Total Data or Extreme Information Management processes.

EIM and Big Data Governance

IBM's Big Data Governance Maturity Framework provides reliable handholds for Extreme Information Management. The accompanying checklist contains more than 90 points of interest in 11 sub-areas. This elucidating material can be accessed via:

ibmdatamag.com/2012/04/big-data-governance-a-framework-to-assess-maturity

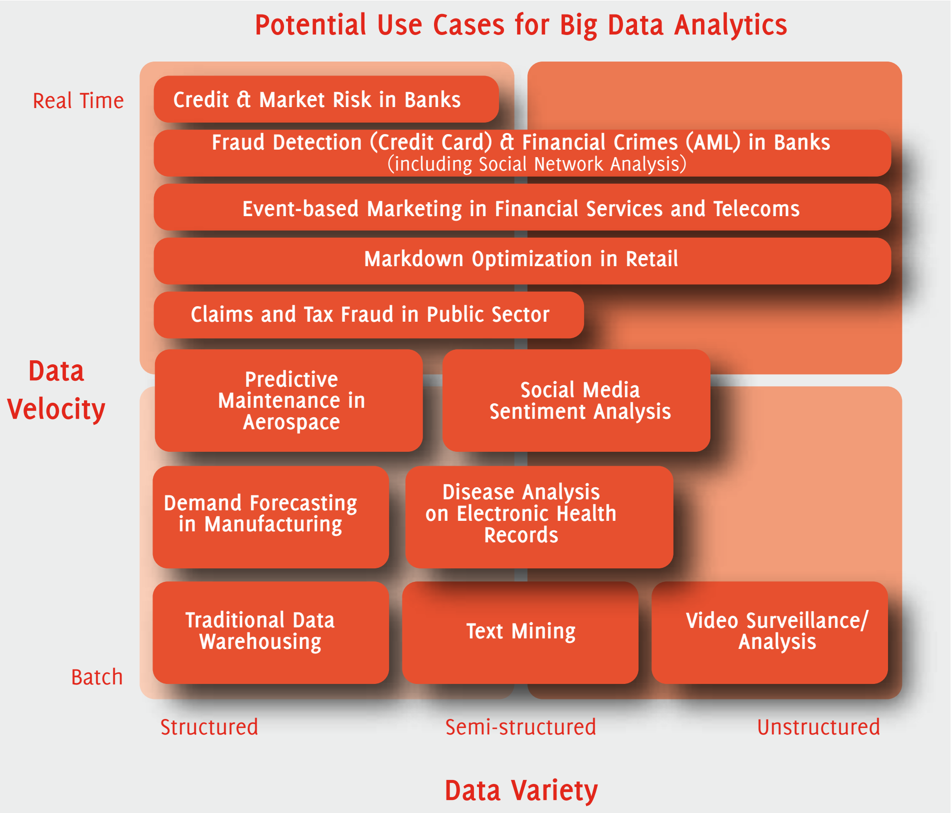
Join the conversation

Question 5
To what extent is Big Data a solution looking for a problem?

www.sogeti.com/vint/r1q5

7 Big Data in organizations in the year 2012

Along the axes of speed (Velocity) and data types (Variety) – thus deliberately abstracting from data quantities (Volume) – SAS and IDC formulated the following self-evident potential of Big Data Analytics for organizations in the year 2012.

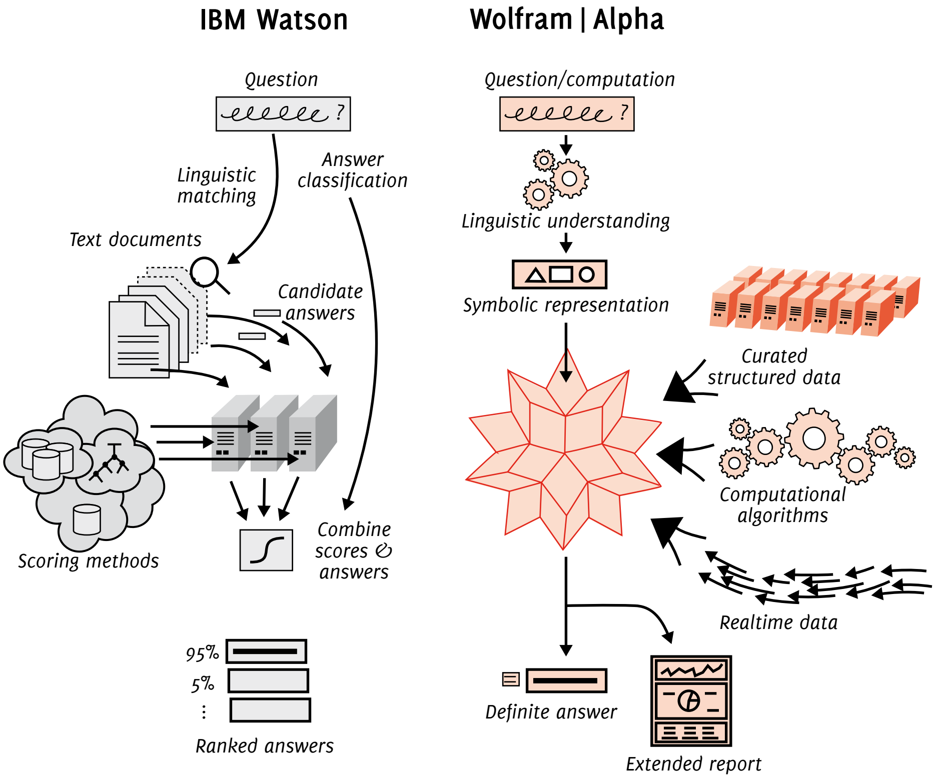

Potential Use Cases for Big Data Analytics
Real Time
Credit & Market Risk in Banks
Fraud Detection (Credit Card) & Financial Crimes (AML) in Banks (including Social Network Analysis)
Event-based Marketing in Financial Services and Telecoms
Markdown Optimization in Retail
Claims and Tax Fraud in Public Sector
Data Velocity
Predictive Maintenance in Aerospace
Social Media Sentiment Analysis
Demand Forecasting in Manufacturing
Disease Analysis on Electronic Health Records
Traditional Data Warehousing
Text Mining
Video Surveillance/ Analysis
Batch
Structured
Semi-structured
Unstructured
Data Variety

Data Science as a sport

The desired intensive interplay between staff members in the field of Big Data and the current shortage of expertise and experience within organizations allow scope for the Web 2.0 approach called 'crowdsourcing'. The Australian Kaggle.com is one example of this kind of online initiative in Big Data service-provision. It makes a sport of Big Data challenges: "We're making data science a sport." In their online arena, as Kaggle calls it, data cracks can participate in diverse competitions. Organizations offer their data and questions, which are subsequently and skillfully analyzed right down to the finest details by experts affiliated with Kaggle. The best solution is the winner and is awarded the stated prize. Fame, prize money and game enjoyment are what the gladiators are seeking:

> *Kaggle is an arena where you can match your data science skills against a global cadre of experts in statistics, mathematics, and machine learning. Whether you're a world-class algorithm wizard competing for prize money or a novice looking to learn from the best, here's your chance to jump in and geek out, for fame, fortune, or fun.*

Developments such as Kaggle are very interesting because the potential of innovations and/or innovative entrepreneurship on the basis of Big Data are highly valued. State-of-the-art computer systems such as Watson by IBM and Wolfram|Alpha play a major role here. These and other intelligent computers are applied in an increasing number of Big Data challenges: from banks to the Smart Grid and healthcare.



The Social Business Analytics example of churning, the erosion of a client stock, which occurs all too frequently in the telecom business, was dealt with at the beginning of this report, in Section 1.

The Smart Grid

All over the world, a great number of pilot projects are currently taking place at the interface of Big Data and the so-called 'Smart Grid'. Grid monitoring is one of the major areas of interest, as is now happening in the Tennessee Valley Authority project, in which 9 million households and more than 4 billion measurements a day collectively supply 500 terabytes of data. Typical applications include the tracing of interruptions and the monitoring of energy use. There are smart meters for electricity, gas and water. It is expected that 270 million will be operational in 2013. If we take this a step further, to intelligent houses, these will each generate 4 to 20 petabytes of data a year on the basis of 100 sensors per household. The need for Big Data applications in the utilities sector is thus increasing, and evolving deregulation is helping it along.

Healthcare

Healthcare is a broad domain that affects us all directly. With regard to clinical use of Big Data, thus for healthcare treatment, it is beneficial to be able to follow information that has been compiled in all sorts of ways over the course of time. In addition, a beginning can be made on pattern recognition, particularly the detection of events that do not occur frequently or are not perceptible when research is oriented to small populations. A good example is the way in which Google is capable, by means of Big Data analysis and in real-time, of following the way a flu epidemic is spreading. Even more impressive is the way in which the scientific Global Viral Forecasting project uses Big Data to prevent worldwide pandemics such as HIV and H1N1 infection. In such matters we must be aggressively proactive, as the absence of results has taught us that we simply cannot just sit and wait while potential catastrophes are developing all around us.

Ahead of our gene chart

A fundamental Big Data development in the field of healthcare is the ambition of the Broad Institute, an initiative of MIT and Harvard, to expand the Human Genome Project, which was eventually rounded off in 2003. Over a period of 13 years, scientists ultimately managed to chart all the 20,000 to 25,000 genes plus the 3 million basic pairs of human DNA. What the mega-project primarily proved was that genes only make up a minor part of our genome and that there are many more fundamental elements that must be identified and investigated.

The Broad Institute has been engaged with this assignment since 2003, and particularly with the issue of how cells actually process information, which not only leads to a better understanding of the genome but also has great therapeutic value. In combination with other institutes, the Broad Institute is currently researching the cell mutations that cause cancer, the molecular structure of the viruses, bacteria etc. that are responsible for infectious illnesses, and the possibilities of their use in the development of medicines.

Genome biology and the study of cell circuits belong to the most important Big Data challenges of our time. At the end of 2011, the Broad Institute had 8 petabytes of data. The institute is continually working on dozens of specialist software tools in order to be able to analyze the data in the required way. All software and data can be downloaded by everyone.

Social analytics

Warehouses use social analytics to rapidly adapt their online assortment to the customers' wishes on the basis of terabytes of search assignments, blog posts and tweets. They now do so within a few days, instead of the six weeks that it normally used to take. Modern social-analytics tools have been optimized for use by business professionals, and can cope with all kinds of data sources: publicly accessible sources, own data and that of partners.

The data flow revolution

Software for the analysis of data flows is used to uncover real-time deviations and new patterns in the data. In this way, organizations can immediately gain new insights and take quick decisions that are necessary on the basis of the latest developments. In this context, you can think of tweets that are monitored, or blog posts, video images, electrocardiograms, GPS data, various types of sensors and financial markets. Modern data-flow software makes it possible to monitor real-time complex associations in situations that are much more complicated than relational databases and traditional analytical methods could possibly cope with. Ranging from patient care to better customer service, data-flow software offers surprising new possibilities.

Preventing medical complications

In hospitals, the respiration, blood pressure and the temperature of patients are continually monitored. In order to be able to detect the often-subtle signals warning of complications, data-flow systems have to be applied. They are capable of identifying the first indicators of malfunction, well before the symptoms actually appear. In the past, 1000 measurements per second were aggregated to form patient reports every half hour or hour, but that is now considered as too crude. In this case, data-flow systems are of vital importance in order to be able to intervene proactively.

An optimum service

Another example is the service to customers. Internet and the social media have empowered the customers and made them fastidious. On average, we trust one another's opinions three times more than we trust those expressed by corporate adverts. Therefore it is essential to listen attentively to what customers and others online have to say and to the information that they are exchanging. The improvement of service currently demands close attention to comments on websites, in e-mails, in text messages and on the social media. If members of staff have to do that manually, the process occurs much too slowly and there are too many inconsistencies in the reporting and the follow-on. With advanced data-flow software for content analysis, organizations are now capable of automatically analyzing that kind of unstructured data and of categorizing it according to certain terms and clauses that occur within the text. With such a policy, the car-hire company Hertz has doubled the productivity of its customer service.

Visionary phase

The examples given with regard to Big Data are as yet rather rudimentary. This is probably an indication of the phase we are in regarding Big Data. Organizations are not yet basing their distinctive value on their capacity to deal with Big Data. This far, we have not been able to identify the real "heroes" of this era, so that the disruptive potential only glimmers through the examples. We are currently in a visionary phase, in which much experimentation is going on. During the Big Data research and in the publication of various research reports, VINT will pay particular attention to cases in different areas, from various angles and sectors.

Join the conversation

Question 6
How much privacy are you prepared to sacrifice to receive optimum service?

www.sogeti.com/vint/r1q6

8 With Big Data from Big Science to Big Business

Big Data is developing most rapidly in the world of Big Science. In 10 years, 2800 radio telescopes in the Square Kilometer Area project (SKA), the largest Big Science project ever, will generate 1 billion gigabytes of data daily. That is equal to the entire Internet on a weekday in 2012. As far back as 2008, Chris Anderson proclaimed the *Petabyte Age* in the magazine *Wired*, and Joseph Hellerstein, from UC Berkeley, announced the *Industrial Revolution of Data*. In comparison: in 2012, Google processes a total of 5 petabytes or 5000 terabytes per hour.

Big Data, Big Science and Big Bang

The terms Big Data, Big Science and Big Bang are all related to a completely different situation than the one to which we have traditionally been accustomed. For Big Bang, we can thank Fred Hoyle, the British astrophysicist, who coined the term in a radio broadcast in 1949. Atomic scientist Alvin Weinberg popularized Big Science in the *Science* magazine in 1961. And it was only relatively recently, in 2005, that Roger Magoulas of O'Reilly Media came up with the term Big Data. Its use was oriented to organizations: ranging from *Next Best Offer Analytics* directed toward the individual, to production environments and sensor data.

Big Business and Big Bucks

So, it is a good habit to call something "big" if we wish to draw attention to it. In this context we can think of *Big Brother* (1949) by George Orwell, not forgetting more profane matters such as Big Business – large (American) enterprises from the mid-nineteenth century – and Big Bucks, both of which have a direct association with Big Science and Big Data. With respect to Big Data, we are currently shifting from megabytes, gigabytes and terabytes to the vertiginous age of petabytes, exabytes and zettabytes. It's all happening extremely rapidly.

The notion that opportunities to capitalize on Big Data are simply lying there, ready to be seized, is echoing everywhere. In 2011, the McKinsey Global Institute called Big Data "the next frontier for innovation, competition, and productivity" and the Economist Intelligence Unit spoke unequivocally of "a game-changing asset." These are quotes taken from titles of two directive reports on Big Data, a topical theme that is developing vigorously, and about which the last word has certainly not been uttered. McKinsey states it very explicitly:

> *This research by no means represents the final word on big data; instead, we see it as a beginning. We fully anticipate that this is a story that will continue to evolve as technologies and techniques using big data develop and data, their uses, and their economic benefits grow (alongside associated challenges and risks).*

The Global Pulse project

As if he wished to underline the qualifying words of McKinsey, Ban Ki Moon, the Secretary-General of the United Nations, presented the so-called "Global Pulse project" at the end of 2011, geared to keeping up to date with a number of developments all over the world via large online datasets – *New Data* in Global Pulse terminology. The project is being run as a co-operative endeavor with various commercial and academic partners, with the ultimate aim of being able to intervene earlier and better in crisis situations if that should be necessary. There are five main projects:

1. A Global Snapshot of Well-being through Mobile Phones
2. Real-Time E-Pricing of Bread
3. Tracking the Food Crisis via Online News
4. Unemployment through the Lens of Social Media
5. Twitter and Perceptions of Crisis-Related Stress

Data Science rules!

Despite such indicative initiatives, the Big Data concept is most closely related to what we call Big Science. There, the Volume, Variety and Velocity aspects, in combination with state-of-the-art hardware and software, are most obviously present, although some people may contest Relevance and Value, certainly in times of crisis. Moreover, the CERN particle accelerator and hyper-modern telescopes are somewhat larger than what we have to deal with businesswise, and they are of a completely different order in terms of data techniques. So, how does Big Data bring us from *Big*

Science to *Big Business*? The heart of the answer is *Data Science*, the art of transforming existing data to new insights by means of which an organization can or will take action.

Without mentioning the currently much-discussed concept of Data Science, Chirag Metha, the former Technology, Design & Innovation Strategist for the SAP Office of the CEO, emphasized above all the importance of the tools and the corresponding collaboration, as Big Data is certainly not only for experts. In Metha's opinion, it is important to involve as many people as possible in the data chain:

> *Without self-service tools, most people will likely be cut off from the data chain even if they have access to data they want to analyze.* ***I cannot overemphasize how important the tools are in the Big Data value chain.*** *They make it an inclusive system where more people can participate in data discovery, exploration, and analysis.* ***Unusual insights rarely come from experts; they invariably come from people who were always fascinated by data but analyzing data was never part of their day-to-day job.*** *Big Data is about enabling these people to participate – all information accessible to all people.*

9 Big Data as new Data Science era

Right from the outset, a key characteristic of Big Science was the fact that the isolated scientist, working in his ivory tower, had become a thing of the past. But it did not remain a distinctive feature of Big Science, as co-operation soon became the norm across the whole of society. Modern science without well-coordinated collaboration has become inconceivable. The report entitled *Big Science > Big Data > Big Collaboration: Cancer Research in a Virtual Frontier*, dating from October 2011, emphasizes that from a Big Data perspective. In this report, Big Science is put into the same category as Big Data and Big Collaboration. In the report itself, the three "Bigs" mentioned in the title are supplemented by *Big Technology* or *Big Compute*:

> *Big Science generates dimensions of data points and high-resolution images to be deciphered and decoded. In cancer research, Big Data often require on-demand Big Compute across settings using a private cloud, a public cloud or mix of the two.*

It is exactly this that changes for organizations when they decide to work with Big Data. If existing technologies and working methods in an organization are not able to cope with Big Data, a new approach will be needed. This means: investing in hardware, in people, in skills, in processes, in management and in governance. According to Gartner, Big Data is primarily literally the Volume component at the basis of what is referred to as *Extreme Information Management*. An integral part of that is *Data*

Science, the “science” that inevitably enters the organization along with Big Data, Fast Data, Total Data and Dynamic Data. Chirag Metha gives the following profile sketch of a data scientist:

> ***The role of a data scientist is not to replace any existing BI people but to complement them.*** *You could expect the data scientists to have the following skills:*
>
> - *Deep understanding of data and data sources to explore and discover the patterns at which data is being generated.*
> - *Theoretical as well practical (tool) level understanding of advanced statistical algorithms and machine learning.*
> - *Strategically connected with the business at all the levels to understand broader as well deeper business challenges and being able to translate them into designing experiments with data.*
> - *Design and instrument the environment and applications to generate and gather new data and establish an enterprise-wide data strategy since one of the promises of Big Data is to leave no data behind and not to have any silos.*

Big Data: a new microscope

With his *Principles of Scientific Management*, dating from more than a century ago, Frederick Taylor put the “scientization” of organizations on the agenda; in his particular case this was scientific management. This was important but it was essentially an issue of continuous improvement. With Big Data, the enthusiasts see a fundamental change, somewhat similar to the advent of the microscope. This is currently a favored analogy: we are on the brink of a new era, comparable with the beginning of modern science around 400 years ago. Owing to the digital “microscope”, which is currently being invented for Big Data, as it were, we will soon be able to analyze and predict events much more scientifically and accurately in all fields, according to MIT professor Erik Brynjolfsson. Eventually we will be able to zoom in and out rapidly thanks to advanced hardware and software, with the ultimate aim of discovering structures and connections that enable us to obtain spectacularly better insight and solutions, and make better decisions: *Data Driven Decisions* en *Predictive Analysis.*

Join the conversation

Question 7
Can Big Data help you predict the future better?

www.sogeti.com/vint/r1q7

10 Closing the knowledge gap is essential

As a topical business theme, with sky-high economic and societal promise, Big Data is currently the subject of much interest and is gathering momentum. This will remain the case, at least in the near future, and accordingly there is a need for a clear picture. In that context, as the McKinsey Global Institute has calculated, 140,000 to 190,000 data experts (data scientists) will have to join organizations in the USA alone, and the number of business people who can deal with such data will have to increase by 1.5 million. First of all, a certain knowledge level is required in order be able to handle Big Data responsibly. Unfortunately there is a structural lack of knowledge in organizations across the entire spectrum. According to an IBM study dating from 2011, organizations are most willing to introduce structural improvements, as indicated by the percentages shown adjacently. A few years ago, the excuse could still be applied that the development of Big Data was only possible for scientific people and a select number of organizations. For all other parties it was simply too difficult and too expensive. That is no longer the case. Pioneers such as Walmart, Tesco and Google have demonstrated that data can be the source of steady competitive advantage. According to IBM, no fewer than 83% of the CIOs currently nurture visionary plans to significantly improve the competitive position of their organization by means of new Business Intelligence & Analytics on the basis of Big Data.

1 in 3	Business leaders make decisions based on information they don't trust, or don't have
56%	Say they feel overwhelmed by the amount of data their company manages
60%	Say they need to do a better job capturing and understanding information rapidly
83%	Cited "BI & Analytics" as part of their visionary plans to enhance competitiveness

The Economist Intelligence Unit underlines this, but also subdivides Big Data conduct in large organizations into the following maturity quartet:

- **Data wasters**
  Of the data wasters, 30 per cent give no priority to the gathering of data. The 70 per cent from this category who do give priority to data-gathering use the data much too sparingly. Such organizations are below-average achievers. *We find them in every economic sector.*
- **Data collectors**
  These organizations recognize the importance of data, but do not have the resources to capitalize on them. They can only store them. They have immersed

themselves completely in data. *We find them primarily in healthcare and professional services.*

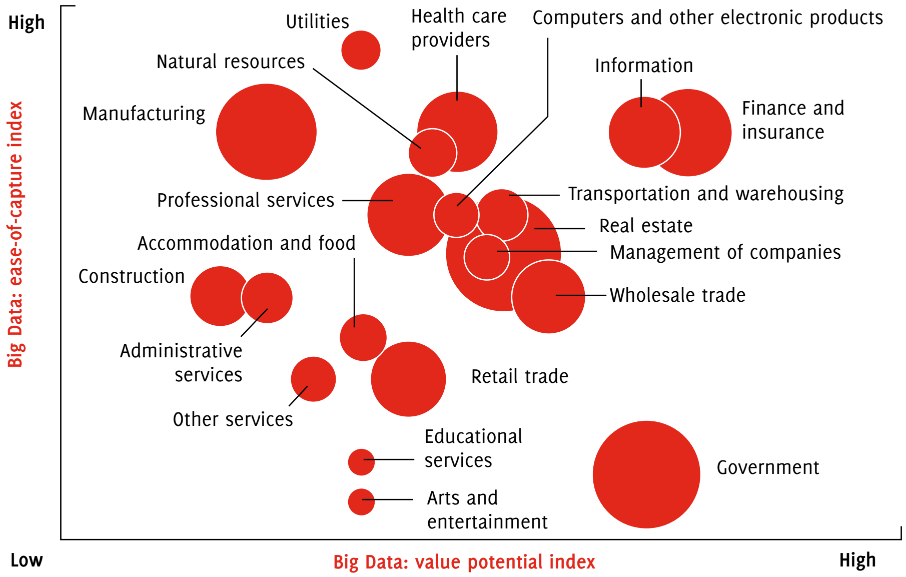
- **Aspiring data managers**
  This is the largest group. People are fully aware of the importance of Big Data for the future of the organization. They use data for strategic decision-making and make solid investments in that area. But they have never reached the upmost level in terms of achievement. *We find them mainly in the communications branch and in retail services.*
- **Strategic data managers**
  This is the most advanced group of Big Data users. These organizations first of all identify specific metrics and data that are related to their strategic targets. *We find them primarily in the manufacturing industry, in financial services and in the technology sector.*

Thus, organizations should not merely collect all kinds of data, but should also develop the wish and competence to work with as much data as possible. In conjunction with business professionals, data scientists must help interpret all the data and generate insights that are genuinely beneficial to the organization. This may concern specific issues or exploratory data research. The intention is to transform an organization from an intuitive decision-making instance into a data-intensive one, shifting from the *heroic manager* who takes decisions simply hoping for the best and knowing that there is too little data available, toward the more *scientific manager* who first seeks data and insight.

11 Big Data in hard cash

Precisely why arrears in Data Science should be made good has been quantified by McKinsey as follows. According to the office, trillions of dollars and Euros can be generated in value worldwide on the basis of Big Data. For example, 300 billion dollars in American healthcare, 250 billion Euros in European government, more than 100 billion dollar in the American telecom business and up to 700 billion for their customers, can be earned on an annual basis. By capitalizing on Big Data, the American retail trade could increase net yield from turnover by more than 60 per cent, and the manufacturing industry would eventually only need to give out half of the current expenditure on production development and assembly, while working capital could decline by 7 per cent.

These are examples from the following overview picture of American economic sectors (see the next page). The great sectoral differences between the ease with which Big Data can be obtained, set against the value that can be expected from using Big Data, are obvious. The McKinsey Center for Business Technology published the chart early 2012 in the reader *Perspectives on Digital Business*, on the basis of information from the report entitled *Big Data: The Next Frontier for Innovation, Competition, and Productivity* by the McKinsey Global Institute, May 2011.



To determine the ease of obtaining data ("ease of capture") on the vertical axis, the researchers have investigated four factors: the analytic talent available, the IT intensity, the data-driven mindset, and the availability of data in a sector. The potential value (horizontal axis) is a function of the following five factors: the amount of data present, the variation in business-economic performance, contact with clients and suppliers, transaction intensity, and the competitive turbulence in a sector. The size of the circles in the figure indicates the relative contribution of a sector to the Gross Domestic Product.

Big Data has great potential particularly in areas that involve many people, such as the utilities and healthcare. This is mostly so due to the relative ease with which Big Data can be obtained, as the figure above shows. In that context, utilities take the title. In terms of the combination of Big Data ease-of-capture, client relevance, financial profit and contribution to the economy, the information-processing industries, including financial service-providers, occupy top position.

Summary and Social Business Analytics

Big Data in the year 2012 is comparable to what the worldwide web was in the early nineties. An enormous acceleration has taken place, everything is being connected to everything else, and the corresponding visions are being formulated. Many people expect that the current data focus will turn the world upside down: in terms of economics, society, innovation and social interaction.

Organizations are currently faced with the major challenge of having to imagine the concrete possibilities of Big Data. How could Big Data generate a revolution in your professional field? Or what would change if you truly succeeded in knowing everything you wanted to know? Could you cope with that? Would you like that and, if so, in which way? And can you allow yourself to wait for further developments in the realm of Big Data, or perhaps avoid participating altogether?

The core of Big Data is that we are dealing with one data spectrum, one continuum. Organizations will explore this continuum step by step, because we do not wish to ignore new possibilities to make better decisions. To help define the urgency of transformation within your organization, we presented and explained the following issues in Section 3:

A. Your Big Data profile: what does that look like?
B. Ten Big Data management challenges: what are your issues?
C. Five requirements for your Big Data project: are you ready?

The interaction covering this and related matters takes place on our website, but may certainly also occur face-to-face as far as we are concerned. We shall share new research insights with you on a weekly basis, via Blog posts, e-mail and Twitter alerts. The accompanying video material featuring leading experts is intended as inspiration to think through and discuss the entire Big Data theme from various angles.

Not all answers can be given immediately, not by a long shot Indeed: even more questions will arise. The Big Data theme is a mission with many questions at the beginning of, and certainly during, the journey. For this reason we will be only too pleased to exchange thoughts with you: online at www.sogeti.com/vint/bigdata/questions and, of course, in personal conversations.

By actively participating in the discussion you will help yourself and us to sharpen ideas relating to Big Data, in order to arrive at lucid and responsible decisions based on progressive insight. In this way, we can jointly determine the concrete substantiation of the coming three research reports after this kick-off on the topic of Big Data.

In many organizations, the focus currently lies on the challenge to chart relevant customer behavior and its consequences as richly as possible, and to steer them in desired directions. This is the core of *Social Business Analytics*, the main theme of the second Big Data research report of a total of four published by VINT.

Literature and illustrations

Anderson, C. (2008): "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete"

Appro Supercomputer Solutions (2012): "From Sensors to Supercomputers (Part 1)"

Appro Supercomputer Solutions (2012): "From Sensors to Supercomputers (Part 2)"

Credit Suisse Equity Research (2012): *The Apps Revolution Manifesto. Volume 1: The Technologies*

Economist Intelligence Unit/SAS (2011): *Big Data: Harnessing a Game-changing Asset*

Frost & Sullivan (2011): *Big Science > Big Data > Big Collaboration – Cancer Research in a Virtual Frontier*

Gartner (2012): *Information Management Goes "Extreme": The Biggest Challenges for 21st Century CIOs*

Harbor Research (2012): "Smart Systems Drive New Innovation Modes"

Hortonworks (2012): "7 Key Drivers for the Big Data Market"

IBM (2011): *Big Data Success Stories*

IBM Data Management (2012): "Big Data Governance: A Framework to Assess Maturity"

IDC/SAS (2011): *Big Data analytics: Future architectures, Skills and roadmaps for the CIO*

Leadership Council for Information Advantage/EMC (2011): *Big Data: Big Opportunities to Create Business Value*

McKinsey Global Institute (2011): *Big Data: The Next Frontier for Innovation, Competition, and Productivity*

Mehta, C. (2012): "4 Big Data Myths – Part II"

MIT Sloan Management Review/IBM Institute for Business Value (2010): Analytics: The New Path to Value

Sumser, J. (2012): "Big Data: The Questions Matter Most"

The 451 Group (2010): "Total data: 'bigger' than big data"

UN Secretary-General (2011): *Global Pulse*

Wolfram, S. (2011): "Jeopardy, IBM, and Wolfram|Alpha"

World Economic Forum (2012): *Big Data, Big Impact: New Possibilities for International Development*

Yared, P. (2012): "Big Data may be hot, but 'little data' is what makes it useful"

Creating clarity with Big Data

Question 1

www.sogeti.com/vint/r1q1

Will facts now definitively defeat intuition?

Question 2

www.sogeti.com/vint/r1q2

How do you link Real-time Big Data to the operational supervision of your company?

Question 3

www.sogeti.com/vint/r1q3

What is the best approach to capture positive attention among the management?

Question 4

www.sogeti.com/vint/r1q4

For organizations, what is the most important new rule of play with regard to Big Data?

Question 5

www.sogeti.com/vint/r1q5

To what extent is Big Data a solution looking for a problem?

Question 6

www.sogeti.com/vint/r1q6

How much privacy are you prepared to sacrifice to receive optimum service?

Question 7

www.sogeti.com/vint/r1q7

Can Big Data help you predict the future better?

SOGETI

VINT | Vision ● Inspiration ● Navigation ● Trends