THE BIG DATA REVOLUTION

A SURVEY OF ITS FAR-REACHING CONSEQUENCES

ANWAR OSSEYRAN & WILLEM VERMEEND

Foreword

Fast-moving and radical developments

Over the next few decades, the world economy will see sweeping and swift changes that will affect the economic balance of power in the world between the industrialised nations in the West and the world's emerging economies such as China, India, Brazil and Russia. For a recent prognosis, see the IMF website (www.imf.org). But nations' economies, including that of the Netherlands, are being revolutionised primarily by the rapid advance of digitisation, a development that is already resulting in new ways of thinking, learning, working and doing business, with the Internet at its core. We are seeing start-ups which, armed with laptops, tablets, smartphones and websites, are successfully competing with existing traditional enterprises. In the years ahead, the turnover of many companies will increasingly be reliant on the World Wide Web, especially via the mobile Internet, in what is already referred to as the *tablet and smartphone economy*. Moreover, companies are now serving customers spread out across the world who are becoming increasingly articulate and demanding and who, thanks in part to the Internet, are exercising a growing influence on companies' products and services and have the power to make or break company reputations. In the past few years we have witnessed the demise of countless traditional companies that failed to adapt quickly enough to the new digital reality, including businesses in the music industry, the travel sector, the world of book and newspaper publishing, but also in the retail sector.

Technological developments

Countries and companies will increasingly find themselves forced to deal with new technological developments, developments that will have a defining influence on nations' entire economies as well as the profit structures of ever more enterprises. International studies have highlighted the key role to be played by new Internet applications as core innovations that will fundamentally impact the economy, particularly the use of mobile Internet and Internet applications in multiple areas such as e-business, e-government, e-health, e-education, e-towns, e-security, e-entertainment, e-energy. Furthermore, smart robot technology, the so-called Internet of Things, 3D printing (see www.3dprintwereld.com), nano-technology, cloud computing, and analysis technology for Big Data will all have a huge influence on policy and decisions in both the public and commercial spheres. It is therefore crucial that new technological developments feature high on everyone's policy agendas. The next few decades are expected to bring greater changes to the economies of many countries than the past forty years. Entrepreneurs who pro-actively seize the opportunities as they arise will be among the winners of tomorrow.

Anyone interested in what this world of tomorrow holds should read the publications of, for example, the McKinsey Global Institute, Gartner and Daniel Burrus, a renowned US innovation expert. It has already been established that a growing number of companies will have to drastically adapt their revenue models, and that failure to do so in time may well spell the end. To survive going forward, it is essential that companies make use of all available sources of internal and external digital information relevant to their profitability and financial health. International research has found that many businesses that effectively analyse and use this Big Data are able to boost their competitiveness and raise turnover and profits.

Survey

In this quick scan we survey the world of Big Data, exploring the challenges and opportunities it presents, but also the downsides. This little booklet is for anyone interested in the developments surrounding Big Data, from professionals who may already be fully familiar with the subject to lay readers without any prior knowledge and who may find it tough to stay the course. It is with these latter readers in mind that we have included appealing practical examples.

April 2014
Anwar Osseyran and Willem Vermeend

Contents

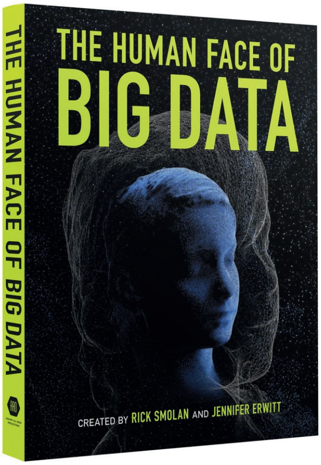
Introduction: The Promise of Big Data

The future is in the hands of those who are able to make good use of the large abundance of data around us. Big Data is not only a hype for computer scientists and data freaks but is a very hot item in many board rooms and cabinet offices. More and more businesses and governments are discovering the big potential of mining the data space. It is all around us on the internet but it is in huge amounts, scattered and unstructured, volatile and unreliable, open but also invasive. Detecting the correct trends and making wise decisions from the tangle of Big Data is far from easy and those who manage to extract and exploit the facts will own our data-driven future.

The 'trick' is to manage, aggregate, de-duplicate, clean and also curate the available data on the millions of data sources on the internet. The quality and accuracy of curated data are improving thanks to the speed of development of hardware and software tools. Decision makers and internet users are empowered by data science, Big Data, and machine learning enabling them to detect trends, make decisions and select contextually between many alternative products, services and offerings. The focus of new Big Data companies like Factual, INRIX, GNIP and Infochimps[1] is on providing global, clean and structured web data to new Big Data applications like Yelp, Foursquare, Trulia, BlockBeacon and Spindle[2], helping consumers to get useful contextual information. Quick and accurate access to internet-based data is becoming a fundamental strength for any company that desires to compete in today's marketplace.

"Essentially, all models are wrong, but some are useful", an axiomatic statement made 26 years ago by Georges E.P. Box in his book on response surface methodology with Norman R. Draper[3] reflecting the power of data and imperfection of present simulation models. Google's research director Peter Norvig went a step further at the O'Reilly Emerging Technology Conference in 2008, announcing that Big Data will make scientific modeling obsolete: "All models are wrong, and increasingly you can succeed without them"[4]. With sensors everywhere, ubiquitous internet, on-demand computing, storage clouds, and social media, we are able to capture, store and process huge amounts of scientific, business and social data predicting trends and forecasting the future well before theory and models have been able to evolve.

The promise of Big Data is growing rapidly and the recently published book, "The Human Face of Big Data," by Rick Smolan and Jennifer Erwitt[5] shows in photographs the impact of Big Data on our day-to-day lives and how Big Data is introducing new scientific, business, health, social and sustainability applications, from early warnings of earthquakes, to supporting business decision and policy making, improving medical drugs and protocols, tracking animal life, assisting police in preventing crimes and improving power consumption in buildings. Big Data will enable us to measure our world in real time and develop a 'planetary nervous system' using sensors, real-time data, cloud computing, data analytics and visualization tools.


THE HUMAN FACE OF
BIG DATA
CREATED BY RICK SMOLAN AND JENNIFER ERWITT
THE HUMAN FACE OF BIG DATA

But will processing massive amounts of data with applied mathematics, supersede the three steps of scientific method [of] theory, experiment and modeling? We do not believe that Big Data will be replacing those; we think it is complementary. Big Data will help us gain a better understanding of data-intensive phenomena especially in scientific disciplines where data is abundant, diverse, volatile and unstructured. In fact, data analysis and data mining form the fourth paradigm[6,7] next to theory, experiment and computational modeling. Vast volumes of scientific data streamed real time by sensors and instruments (like the Square Kilometer Array[8]) or captured in experiments (like the Large Hadron Collider[9]), along with data generated by simulations of computational models, will be curated, analyzed and will require a special communication and publication infrastructure. In a certain sense, this fourth paradigm offers us an integrating framework allowing theory, experiments and computer-based modeling to interact and reinforce each other and to look both backward and forward. Data and software must therefore be integral parts of the scientific record, enabling scientists to reproduce experiments, conduct new analysis or proceed where others have left off. This will require systematic record management and adequate data stewardship.

2 Defining Big Data

A good summary of what 'Big Data' encompasses is given by Mayer-Schönberger and Cukier, in their book 'Big Data: A Revolution That Will Transform How We Live, Work, and Think'. By analyzing large quantities of data, it becomes possible to discover patterns and interactions that were previously invisible. In this way, new solutions can be found for difficult problems, as well as new opportunities be discovered. An excellent example of the use of Big Data is predicting seasonal influenza (flu) by Google[10]. By collecting relevant search terms worldwide and analyzing and weighing them in a certain way, reasonable predictions can be made of how seasonal flu is spreading. This is something that was much more difficult to do before. Another good example of Big Data use is the election campaign by US President Barack Obama[11].

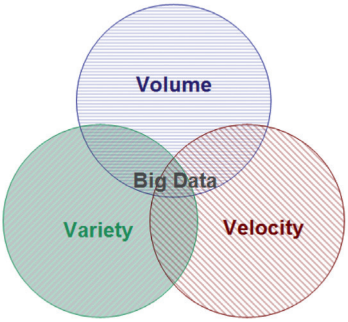
The 3 V's: Volume, Variety and Velocity

But how to define Big Data, and is Big Data synonymous with large amounts of Data? Volume is the first of three elements of Big Data, as originally defined by Doug Laney in a 2001 Meta Group research report[12]. The two other elements are *Variety* and *Velocity.*

Volume is the most obvious element to characterize Big Data, mainly relating the size of data to the processing capability. Volume is therefore a moving target as data capture will continue growing as well as the ICT capacity of storing and processing it. Dealing with volume requires scalable storage and distributed techniques for querying or aggregating data in a cost-effective way.

Variety is an indication of the large amount of various data types that are stored and that still need to be processed and analyzed while the traditional relational databases are largely unsuited to cope with this task. New data types from social networks, machine to machine communications, and mobile devices add to traditional types of structured data generated by computers through transactional processing. Examples of such new data types are pictures, audio and video files, GPS data, medical files, instrument measurements, images, RFID, log files and web documents, Blobs, RTF files and text strings. Unstructured data such as speech and social media increase the complexity of processing and categorization of data and require the deployment of new technologies like speech processing for data mining, noisy text analytics and pattern recognition.

Velocity, the third V, is a measure of the rate of change of data, indicating the temporal value of the data itself. Relational databases are generally less-suited to deal with "volatile" data. Velocity of Big Data requires fast processing of structured and unstructured (streams) of data in order to benefit from geo location data, detected hypes and trends and available real-time market and customer information. Traditional relational databases are mostly unsuited to deal with Velocity. A new approach must therefore be adopted in order to capture, store, curate, aggregate and analyze the data quickly. The greater the 3 V's, the more difficult it is to solve the technical issues, but at the same time the business and scientific opportunities would then also be bigger.



Other Big Data V's: Viscosity and Virality; Veracity and Value

Advances in social networks, mobile technologies, cloud computing and unified communications bring two additional characteristics that need to be dealt with in order to gather insight from the various data sources at our disposal: *Viscosity and Virality*. Viscosity characterizes the resistance to navigate in the dataset related for instance to variety of data sources, data flow rates or complexity of data processing required. Virality is a measure of the spread rate of data across the network. Time is an important characteristic along with rate of data proliferation.

Two other V's connected with Big Data are *Veracity and Value*. Veracity describes the quality and lineage of the data, to characterize data in doubt, conflicting data or noisy data and ultimately, data of which users are unsure how to deal with. Value is to characterize what value could come out of which data and how Big Data will enable the user to get better results from the data stored.

Other Big Data characteristics

Beside the seven V's defined above, there are other important characteristics to take into account in a Big Data case: the level of aggregation of the data stored as discarding original raw data may undermine the validity of Big Data analysis performed; availability and use of metadata (time, place, source, context, etc.) which help to significantly improve the effectiveness and usefulness of Big Data methods; and data signal-to-noise ratio as faint signals require fast and accurate methods in order to isolate the correct data effects and timely get to the right conclusions.

3 Data trends and Big Data drivers

IDC recently published a study sponsored by the storage company EMC[13], showing that the global data volume called *Digital Universe* is growing more rapidly than previously forecasted. The digital universe is doubling every two years, growing towards the 40 Zettabytes in 2020. Despite the Big Data hype in the last few years, less than 1% of the world's data was analyzed in 2012 while according to IDC 23% of the data could generate much more value if tagged and analyzed with Big Data analytics. Machine-generated data are a major contributor to this expansion growing from about 11% in 2005 to more than 40% at the end of this decade. According to the IDC, by 2020 there will be more than 200 billion devices interconnected (sensors, actuators, vehicles, cameras, smart devices, home appliances, industrial and medical devices, toys and gadgets, etc.) and will soon become the major users of internet and a major growth source of our digital universe.

The Internet of Things

This explosive growth of machine generated data is driven by the synergy offered by advances in electronics and ICT. New individual applications exploit those new technological possibilities in many areas including industry, urban development, environment, buildings, security, government, and healthcare. For example: optimization of manufacturing using sensors, cameras and actuators is not really new, but combined with internet connectivity and wireless technology it opens up the possibility for autonomous monitoring of trends and remote corrective measures. Keeping track of weather forecasts, pesticide usage, fertilizer and moisture levels in soil offers the high-tech farmer many 'precision agriculture' advantages. Smart cities are exploiting smart grids and smart street lighting to optimize sustainable energy use and minimize carbon emission. Smartphones offer a sophisticated platform for apps to record, map and share health and environmental data within urban areas. Sensors assist in keeping streets safe and clean, issue pollution or security warnings or optimize parking use. Web-enabled home outlets, smart thermostat apps, connected appliances, web-enabled lighting and-sensors help us to reduce energy costs and increase comfort and security. Smartphones, smart body devices, smart clothing and even smart pills enable us to monitor our activity level and keep online track of our breathing, body temperature, blood pressure, sleep patterns and other important health parameters. Innovative health devices at home help the world's increasingly grey populations stick to prescribed regimens, follow medical instructions, remotely monitor patients, read their biometrics and maintain the high quality of individual care services while lowering their huge costs.

The differentiating factor of Big Data

Big Data has evolved from being a huge problem at the beginning of this century to a major business opportunity a decade later. The McKinsey Global Institute characterized Big Data in their recent study[14] as being too big, too diverse and too fast to fit into the legacy database architectures in a cost-effective way. Alternative ways to store, process and mine Big Data are therefore needed. As The Economist has put it in 2010[15]: "Businesses, governments and society are only starting to tap its vast potential". The solution was mainly provided by the rapid increase of capacity of CPUs and storage devices but also by new open-source software frameworks like Hadoop[16], derived from Google technology and put to practice by Yahoo and others. Since then, other tools were developed in order to store and manage Big Data like NoSQL[17] and massively parallel processing databases like IBM's Netezza, HP's Vertica, and EMC's Greenplum, CalPont, EXASOL, Kognitio, and ParAccel [18].

The big advantage of the huge amount of data collected and stored on the web is not only its direct use, offering a snowball of business and social applications; the gold mine is in the aggregation of the data collected, the analysis of the collective meaning of the data and in the multidisciplinary interpretation of the data collected. This is where Big Data enters the picture. The storage and management of individually collected data can involve in the hundreds or even thousands of terabytes, exceeding the storage, management and affordability limits of traditional relational databases. At the same time the challenge of managing unstructured data and its high speed of creation and loss of value, requires new measures. Emerging technologies like Hadoop, are designed and built to process very large volumes of semi-structured data. NoSQL database technology deals with the scaling problem better than relational databases can. In combination with Big Data platforms allowing storage of all data in its native formats, it becomes possible to get value of huge, unstructured, volatile and in some way uncleansed data through massive parallel processing on commodity hardware and deployment of smart Big Data Analytics techniques.

4 Coping with Big Data

Google's CEO Eric Schmidt announced at Google's 2010 Atmosphere convention[19] that every two days about 5 Exabytes of information is created, equivalent to the amount of data stored between the dawn of time up until 2003. However, not all of these 5 Exabytes were stored. In fact, 2007 was the first year in which the amount of data produced exceeded the total storage capacity available. Coping with fast streams of large volumes of time-dependent largely unstructured data has become a seriously challenging task. Managing ever increasing creation rates of data requires real time analytics for data streams generated by heterogeneous sources like sensors, instruments, log files, cameras, internet traffic, blogs and tweets. The challenge is to find a way to mine with those large, fastly changing data streams with reasonable accuracy and within limited time and ICT resources constraints. On the other hand, a good prediction process will also depend on the quality of the preceding learning process of the algorithm used.

The three main constraints: accuracy, time and resources are in fact communicating vessels. More accuracy requires either more time and/or more ICT resources. Data mining requires less time with less data or faster processing but the accuracy may suffer. The strategy would then be to use distributed systems in order to cope with resource and time constraints and to exploit the learning process to smartly sample the data streams or use probabilistic techniques[20].

The NoSQL Revolution

Relational databases became a major liability in dealing with Big Data queries as scaling and customization became unaffordable. The so-called NoSQL databases offer an alternative better suited to tackle new Big Data challenges. Four categories of NoSQL databases can be distinguished[21]: Key-values stores, Column Family stores, Document databases and Graph databases. Their main advantages are their ability to scale out by adding commodity hardware, thereby increasing capacity at low cost, and the relative simplicity of changing data models when needed due to fluctuating demands.

The Key-values category is the simplest and easiest to implement and is best suited to loosely-coupled datasets, while in the other categories, the relationship between dataset items is as important as the characterization of the items themselves. Key-value stores offer speed and scalability for simple queries and simple data models, but lose performance with data that needs more context. Compared to Key-values, the Column family category offers an additional grouping of key-value data elements thus enabling it to perform more complex queries, while the Document category stores the data items as objects allowing complicated data retrieval mechanisms through object-oriented programming. Performance becomes a big issue when datasets are fundamentally interconnected and non-tabular, this is where Graph databases come into the picture. Typical examples are geospatial problems, network analysis and recommendation engines. Typical application areas are bioinformatics and financial analysis where the closest relationship between data items needs to be determined.

Big Data tools and the Open Source Community

The biggest advantage of the Big Data revolution today is that the tools used are mostly open source like the Apache-Hadoop[22], -Pig[23], -HBase[24], -Cassandra[25], -S4[26], Storm[27], Pegasus[28], Scribe[29], Cascading[30], GraphLab[31], R[32], MOA[33] and Vowpal Wabbit[34]. The MapReduce[35] technique was initiated by Google to effectively processing webpages while crawling on the web. Hadoop[36] is an open-source implementation of MapReduce and is the most used non-streaming Big Data analysis tool for data-intensive distributed applications. Hadoop offers a programming model and a software framework to distribute large amounts of data on its Hadoop Distributed Filesystem (HDFS) and to process the data in parallel on distributed clusters of compute nodes. Like its name suggests, MapReduce uses two distinct steps first map and then reduce the data to be analyzed.

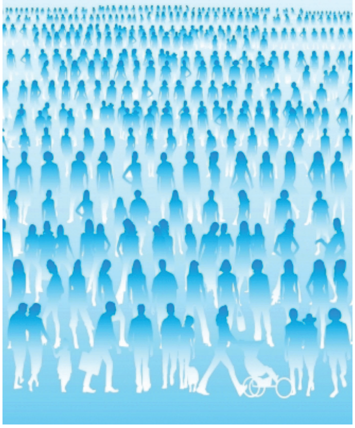
licensed under (cc) Attribution-NonCommercial-ShareAlike 2.0 Germany | Ludwig Gatzke | http://flickr.com/photos/stabilo-boss/

Large companies like Yahoo!, Microsoft, LinkedIn, Facebook, Google and Twitter contribute to the Big Data open source developments and benefit mutually from the advances made by the communities. A tight collaboration is also ongoing between the academic and industrial research communities defining research areas and exchanging results in conferences like ICDM[37], KDD[38] and ECML-PKDD[39].

Data Analytics and Social networks

Social networks provide a very interesting source of information that was not available at this scale a few years ago. Mining evolving data streams in social networks like Facebook, LinkedIn, Hyves and Twitter requires a totally different approach that copes with speed and context of the data streams and related knowledge discovery challenges. For instance, in 2013[40] Facebook had about 1.11 billion users (May 1st), YouTube had 1 billion (March 20th), Twitter 500 million (March 21st), Shazam 300 million (April 29th), LinkedIn 225 million (May 6th), iCloud 300 million (April 24th), Google's Chrome 750 million (May 15th), Gmail 425 million (February 7th), Google+ 343 million (January 26th), Sina Weibo 503 million (February 21st), Tencent's QQ 825 million and Qzone 611 million (May 15th, 2013) and Skype and Yahoo!, both had about 280 million users each (October 2012). On social networks, data arrive hence at high speed, in large volume and mostly unstructured. For instance, in October 2012 the number of tweets per day was 500 million (according to Twitter CEO Dick Costolo), up by about 25% in 4 months. And according to Twitter, you don't really need to tweet to be on Twitter as 40% of their users use Twitter as a "curated news feed of updates that reflect their passions"[41].

Sentiment analysis and opinion mining are among the applications of Big Data techniques to mine the social networks data streams and classify messages into positive or negative feelings according to the sentiment they convey. Classifiers for sentiment analysis use machine learning techniques and train the algorithm with training data labeled using for instance expressed emoticons in the data streams. Carnegie Mellon University developed NELL[42] (Never Ending Language Learner), extracting structural information using text analysis of hundreds of millions of unstructured web and continuously improving its ability in extracting instances and relations from data streams.



Tools mentioned above like Hadoop MapReduce speed up the mining of data, distributing the training process onto parallel machines and dividing the input datasets into independent subsets that are first mapped and then reduced. S4 and Storm are instrumental for processing continuous data streams. Coping with historic and real-time data at the same time requires the use of both techniques. An interesting approach, so-called Lambda architecture, was developed by Twitter's engineer Nathan Marz[43], dividing the data mining problem into three layers, so-called batch-, serving- and speed layers. The data is sent to both batch layers and speed layers. The batch layer uses tools like MapReduce to compute query functions that enable fast queries of the historic data in the serving layer. The speed layer uses fast incremental algorithms like S4 or Storm to achieve the fastest latencies possible focusing on recent data whereas the batch layer produces views on the entire data set. Combining both in the Lambda architecture offers a robust, fault tolerant, extensible and debuggable system for ad hoc queries of fast social network streams.

Big Data and Cloud Computing

The rise of Big Data would not have been possible without the on-demand offerings of Cloud Computing. Both virtualization and cloud computing made it possible to scale capacity and reduce costs of data storage and processing. Clusters of computing and storage nodes are deployed within a private or public cloud to analyze large volumes of data streams that fluctuate with time and may occasionally explode. Cloud computing offers the possibility of deploying additional virtual machines thereby ensuring that Big Data analysis is not delayed by even unusually huge datasets, while avoiding having to upgrade the database or permanently add expensive hardware.

While virtualization abstracts the underlying physical hardware, offering higher level services such as cloning a data node, high availability to a specific node, or user controlled provisioning, the clouds offer a collection of virtualized hardware with additional services such as additional resources on-demand (IaaS), various computing platforms (PaaS) or catalogs of software (SaaS)[I]. A single Hadoop image can be easily cloned, and the storage needed and computing resources expand on-demand. Public clouds differ from private clouds in that they may provide scale cost benefits but at the detriment of loss of control, privacy, security or data custody. Examples of public cloud offerings for Big Data are the Amazon Elastic MapReduce[44] and the solutions-google-compute-engine-cluster-for-hadoop[45].

Virtualization and cloud computing are not however appropriate for all Big Data use cases. Storage, processing and networking resources might cause delays or even serious disruptions if they are shared or undersized. On the other hand, Hadoop's assumptions about the underlying physical hardware do not always hold in a virtual environment. Hadoop's low-cost yet reliable Hadoop Distributed File System (HDFS) for instance is based on a three-way replication[46] on to local low-cost storage and uses physical topology awareness to optimize data blocks across racks and hosts. Optimization of run-time may therefore become obsolete when the physical topology information is not shared with the users. Another disadvantage of virtualization is its impact on latency. Virtualization does not guarantee the low latency needed for real time data streams like for example Twitter feeds, especially when milliseconds signify an important competitive edge or make the result too stale to be useful. And last but not least, public clouds will not offer the privacy and security needed for processing, for instance, sensitive financial, competitive or medical data streams. Private or community clouds may offer a better alternative here.

[I] *Iaas, Paas en Saas: Infrastructure-, Platform- en Software-as-a-Service*

The ultimate challenge: getting insight

The ultimate challenge of any Big Data approach is how to offer the user the needed insight into large volume of rapid and unstructured data streams. Visualization has always been a key tool for understanding and digesting huge datasets resulting from complex analyses, creating the possibility - through images, diagrams, graphs or animations - for humans to communicate and interpret analysis results and make better decisions.

Various visualization techniques and technologies are being developed specifically in the context of Big Data[47, 48]. Exploring emergent patterns in for instance live data streams is a big challenge to existing visualization techniques. New methods are therefore being developed to enable analysts to explore and dissect the Big Picture offered by Big Data analyses of news, comments, and social media data streams. Visualization transforms text overload into insightful and actionable information. This is why Big Data companies like Twitter and DataWatch boost their own visualization capabilities through acquisitions. Twitter acquired Big Data Visualization Startup Lucky Sort in May 2013[49] and one month later Datawatch acquired Panopticon Software, a supplier of data visualization and discovery tools[50].

The future singularity of Big Data with Quantum Computing

In the future, processing Big Data requires an ever increasing capacity of computing power. Whereas capability computing is about using the maximum computing power to solve one large problem in the shortest amount of time, capacity computing is about using commodity servers to solve many small problems in parallel. As described above, deploying capacity clusters in the clouds and using Hadoop framework is the way to go at present. But how to cope with huge Big Data problems when the required computing capacity largely exceeds that available or affordable, and this “singularity is near”[I]. Two alternatives may offer some relief on the long run.

The first variant is already being deployed to maintain the Bitcoin digital currency system. Bitcoinwatch.com estimated in May 2013 the data mining capacity of its networked computers to exceed the exaFLOPS - over 8 times the combined speed of the top 500 supercomputers at that moment[I]. The Bitcoin system was invented just four years ago to draw in computing power to solve mathematical problems of increasing complexity.

[I] *http://www.singularity.com/ The Singularity is Near door Ray Kurzweil*

On a longer run, quantum computing[II] may provide the cost-effective answer to deal with the Big Data singularity. While electronic computing is based on the deterministic bit state of a one or a zero, quantum computing uses the probabilistic quantum state of atoms as computing devices making it possible to program the atoms to represent all possible states of ones and zeros simultaneously. That means an algorithm can test its possible input combinations at once, instead of serially cycle through every possible input combination to arrive at a solution. This makes quantum computing very interesting for Big Data analytics seeking to find the best possible answer from a huge set of potential answers. Quantum computers excel at this because of their ability to parallel process many different approaches to the same problem simultaneously. This is also why Google and NASA joined forces to operate a laboratory that will feature a quantum computer to explore how quantum computing could enrich machine learning[III].

[I] *http://money.cnn.com/2013/05/23/technology/enterprise/bitcoin-supercomputers/index.html*

[II] *http://www.howstuffworks.com/quantum-computer.htm*

[III] *http://www.fedtechmagazine.com/article/2013/05/nasa-and-google-plan-tackle-big-data-quantum-computing*

5 Big Data potential in science, business and society

Nowadays it is possible to sense, collect, store and analyze massive amounts of data generating huge collections of raw data and to deploy Big Data techniques and tools to reveal much of their intrinsic value to science, business and society. Companies like Google, Facebook, Yahoo!, Twitter and Microsoft collect massive amounts of data every day and continuously add new services based on collected or pre-processed data such as satellite information, maps, tweets, images, videos and social interactions. These data repositories and services are generating new innovations and have a deep impact on science, business and society alike.

Fourth Paradigm and data-intensive Science

A historic showcase of the deep impact of data mining on science is the set of laws of planetary motion elaborated by German mathematician Johannes Kepler (in 1609), through mining the collected observation data of his colleague astronomer Tycho Brahe. With the advent of Big Data, scientific breakthroughs in this 21st Century will increasingly be powered by advances in computing and data analysis capabilities that help researchers manipulate and explore massive datasets and conduct data-instead of hypothesis-driven science. This so-called Fourth Paradigm of Science[6] is based on the full exploitation of massive volumes of measured data, uninterruptedly collected by sensors and instruments, social data registered in blogs and websites, or simulation data generated by computer models. Like in Kepler's case, those data repositories will result in many new theories and discoveries. A prerequisite is the availability of those repositories of captured, curated and analyzed datasets and a well-established communication and collaboration infrastructure within the scientific communities.

Astronomical figures

Most scientific disciplines are data-driven nowadays. Astronomy for example, with its large radio telescopes and light detectors, is one of the main reasons that the amount of data stored in the universe is doubling every year. The Large Synoptic Survey Telescope[51], operational in 2015, will generate images of three Giga pixels each. The SIMBAD astronomical database[52] provides access to basic data, cross-identifications, bibliography and measurements for more than 7 million objects outside the solar system. The Square Kilometer Array (SKA) project[8], with

67 scientific teams from 20 countries working together, is like CERN's Large Hadron Collider (LHC)[9], a global scientific collaboration project aiming at giving a better insight into our universe. The SKA radio telescope will produce an exabyte of data every day! The LHC produces about a petabyte of raw data per second, this is 86,4 exabytes per day, but stores only 41 terabytes per day, about 15 petabytes of raw data each year.

This exponential increase in data volumes also meant a paradigm shift in the way scientists deal with the data. They realized it is impossible to download and process the data locally, not only because the datasets are enormous but also because the data is diverse and also very dynamic. The big leap was to create Big Data repositories allowing the scientists to get to their data and to conduct their analysis locally without having to download and replicate it. This way, tier-1 and tier-2 sites with large data repositories, processing tools and interfaces allowed high energy physicists and astronomers to mine millions of sources of data in a short time span, and to share the results of their research within the community.

Big Data in Healthcare, Pharma and Life Sciences

Big Data techniques allow medical researchers to apply data mining techniques on the huge amounts of patient data collected through imaging modalities (scanners, MRI), genetic analysis (DNA microarrays, NGS), lab results, monitoring equipment and other medical data sources but also socio-demographic data and other public data sources. This will enable them to gain fundamental insight into the genetic and environmental influence on diseases and to develop personalized medicine in aid of higher quality care, better outcome and lower costs. For instance, optimal treatment pathways can be developed for each individual patient by analyzing and correlating patient and treatment data to identify the most efficient and effective treatment to apply per patient and per disease. Equipment for permanent monitoring of patients (wired or mobile) generates huge amounts of data that can be mined to improve the understanding of diseases, treatment protocol, improve clinical trial design and drug discovery.

Decision support systems for physicians advise them about the latest drugs and treatment protocols based on the optimal treatment pathways mentioned above but also on the most recent patient data, thus helping to prevent clinical mistakes or adverse drug reactions. Patient profiling based on segmentation and predictive modeling using Big Data analytics enables the patients to initiate preventive treatments or lifestyle changes, to minimize or avoid predicted disease risks. Finally, Big Data analytics applied to anonymized patient records and related medical procedures and protocols allow improvement of quality, performance and efficiency of the healthcare system and offer patients more transparency of the cost and quality of the service providers in this still very opaque sector.

Like many medical institutions, life science firms, insurance companies and other stakeholders, the pharmaceutical industry understood the importance of sharing data and exploiting Big Data techniques. Peer-to-peer networks and dedicated dark fibers were used in the last decade to transport data to researchers for analysis and mining. Big Data techniques offer them the ability to process the data at its location as opposed to moving the data to the user location, thereby eliminating the known privacy, security and data ownership issues. A health sector cloud for instance, relieves the burden of hosting the sensitive patient data, enables the creation of a distributed Electronic Health Record (EHR) and allows the researchers to mine the EHR using emerging Big Data analytics. The use of the EHR not only reduces time and costs of clinical trials, but also enhances the information gained as it helps design more-targeted studies on smaller populations. However, without adequate clinical interpretation, Big Data in life sciences could though be a hazard and should therefore be used carefully. Patient privacy should also be adequately protected and analytics must always be segregated from the raw data in the EHR.

Finally, a word about hereditary disorders and Big Data: the website Ancestry.com, stores data from historical records, such as census data and military and immigration records including birth and death certificates. In addition, this website offers its users the ability to store user-generated data. The result is a repository with about 40 million family trees and 4 billion people represented. The site also allows DNA collection of users making it possible to use recent advances in DNA technology to look back hundreds of years and make correlations as part of the user's family story. Machine learning techniques are also used to make suggestions to the users about their relatives. Connected to the EHR of the users, the site offers an unprecedented repository for research on hereditary disorders.

Big Data in conjunction with Energy and environment

The energy sector is just beginning to discover the potential of the huge amount of data that can be collected, stored and analyzed when the Big Data infrastructure is put in place. Utility companies in Europe have started to roll-out smart meters but most still lack the infrastructure and the analytics to store and analyze the data. When implemented, the potential is enormous as Big Data will allow customers to understand and optimize their energy use, compare theirs with that of neighbors and learn how to improve their energy efficiency. Likewise, utility companies will be able to minimize peaks reducing the size of their energy stations and the need to develop or purchase additional energy resources. But the biggest gain will be realized when smart grids become operational. Smart grids will drastically modernize the utility electricity delivery by analyzing and mining the data feeds of sensors and smart metes in the smart grid and decentralizing energy production. In conjunction with Big Data, smart grids will generate new foundations for energy consumption and production. Automation will play an important role here as production and consumption of energy will be seamlessly optimized by smart software agents negotiating between the consumer utilities and the electricity grid. The smart software agents will enable the deployment of decentralized energy sources such as sustainable energy sources (wind, solar, hydroelectric, geothermal…) or local power supplies (electric cars, local power generators…). The software agents and the application software coordinating all the decisions within the system will use the results of data analytics to continuously improve the models used for negotiations and decision making.

Beside improvement in carbon emissions and costs, another obvious advantage of deploying Big Data in the energy sector is safety. Big Data analysis allows us to learn from historical data and detect patterns of disasters, enabling the monitoring of critical signals and avoiding hazards in the future. Temporal and special data also help to tune the local production with the expected energy consumption or determine the optimal placement and orientation of solar cells or energy turbines. The location chosen to install and exploit a wind turbine for instance depends on a large number of special and temporal parameters like wind, temperature, precipitation, humidity, altitude, atmospheric pressure and surroundings and will hence have a huge impact on the efficiency of the power production as well as on the lifespan of the turbine itself. Optimizing the placement of a wind turbine requires several petabytes of historic data, that is to be analyzed using Big Data techniques.

Additionally, Big Data helps to better understand the environment by analyzing a wide range of globally available, national datasets measuring various environmental issues, ranging from carbon emissions to air and water pollution to deforestation. These measurements are used to conduct simulations of climate change and to improve predictive modeling of the planet. But they are also deployed to calculate the environmental Key Performance Indicators (KPI's). KPI's are needed in order to avoid wrong assessment of environmental policies or decisions and used for instance in setting the Environmental Performance Index ranking[I]. The challenge of having important gaps in environmental repositories is being met through the emergence of crowdsourcing and citizen science such as Dangermap[II], a Chinese site displaying pollution hazards based on government data and crowd input. Open source tools and platforms like Arduino[53] offer the crowd the ability to develop interactive sensor-actuator devices for measuring environmental data collection.

Big Data in the public sector

A wealth of data is being collected within the public sector such as tax data, social security records, socio-demographics, polling records, and other public sector data. In their study[14], McKinsey showed that with this wealth of data collected, the efficiency and productivity in the public sector can be boosted by deploying Big Data techniques. Big Data analytics can help develop evidence-based public sector measures to enhance quality, improve performance, increase productivity, boost revenues and lower costs. Acceptance and embracement of public sector policies and services also improves when policy makers are able to better understand their citizens, develop more tailored services and products resulting in less bureaucracy and waste and a better meeting of public needs. Big Data's transformative effect could also drive reforms in regulations and government policies.

Instruction
Employee Business
Section reference
Revenue Cod
Ge
or Form 8888.
Account number
Amount You Owe
75 Amount of line 73 you want applied to your
76 Amount you owe. Subtract line 72 from line 6
77 Estimated tax penalty (see page 62)
Third Party Designee
Do you want to allow another person to discuss thi
Designee's name ▶
Pho no.
Sign Here
Under penalties of perjury, I declare that I have examined this
belief, they are true, correct, and complete. Declaration of prep
Joint return? See page 17. Keep a copy for your records.
Your signature
Spouse
D
Paid Preparer's Use Only
Preparer's signature
Firm's nar yours if se address,

One of the urgent improvements that can be reached with Big Data in the public sector is the increase of transparency and increase of citizens' trust. Making the massive amount of public sector datasets available to citizens as open data[III] improves transparency and helps citizens understand which data is gathered and what the public services are doing with it. Transparency would allow citizens to monitor how public money is spent and force governments to improve efficiency and reduce costs. This would also stimulate governments to respect the privacy of its citizens and encourage public-private initiatives to develop new innovative services that better meet the public needs.

Improving the public services requires a proactive attitude from the civil servants. Data Analytics of not only structured public data, but also of unstructured social media data will enable them to detect and quickly react to the society dynamics and changing events. Voice recognition techniques can assist in detecting and better understanding national or group

[I] *http://epi.yale.edu/*

[II] *http://www.weixianditu.com/*

[III] *https://data.overheid.nl/*

sentiments and social unrest or riots could then be anticipated and possibly even prevented. On the other hand, segmentation of citizens could help develop algorithms to automate decision making and at the same time personalize the citizen experience of services offered. Segmentation allows the public services to better address the actual needs of the individual citizen segments.

Another hot issue in most countries is detection and reduction of tax and welfare fraud. Using Big Data analytics, tax officers could detect more fraud patterns and better identify suspicious transactions at even earlier stages. Analysis of regional and national public data combined with social media delivers a better insight in abnormal behavior and leads to early actions. Profiling techniques and Big Data statistical analysis help in the detection of patterns and identification of fraudulent behavior. Transparency is also important here as it strengthens the trust of citizens in the government actions as well as acting as a preventive measure, discouraging people from making the wrong choices.

Finally, a word about the sensitive subject of Big Data in conjunction with national security and people privacy. The use of Big Data tools enables government agencies to monitor the citizens and to detect social unrest, riots or suspicious criminal or terrorist behavior. Public agencies have access to huge amounts of privacy-sensitive data about citizens as well as the means to collect, process and analyze those datasets in conjunction with social media and other public (and even private[I]) data without the explicit knowledge or consent of those citizens. This enables those agencies to unmask criminal plans, intercept drug trafficking, oppugn child abuse or prevent terroristic attacks but can also be misused by government officials or intruders. This explains the social controversy around programs like PRISM, set up to secure the world against terrorist attacks. Governments and companies can gain a lot by deploying Big Data but prerequisites for acceptance by citizens and customers are transparency, and the assurance that privacy and integrity are secured and respected.

[I] *Like mail, phone calls, video calls, private chats, satellite images, CCTV footages, etc.*

Big Data and impact on e-commerce

E-commerce puts an increasing downward pricing pressure on retailers by giving consumers immediate access to powerful pricing, promotional and product information online. This price transparency is shifting huge value to consumers. Big Data on the other hand offers retailers new opportunities for capitalizing on the data e-commerce is providing. Online purchase data combined with social media leverage an immense repository of customer information.

At the same time, data storage is becoming so cheap that almost all data can be stored. More than product price alone, business intelligence has become much more important for a competitive strategy. Crunching the gathered data to understand customer behavior delivers insights that are changing marketing strategies and in-store plans. Big Data offers retailers the possibility to improve the personalization of their services despite the huge growth of their customer base. Techniques like Next Best Offer (NBO)[54] combining target segments with product categories, price optimum, preferred channel and time-to-offer, represent the convergence of real-time data processing and mobility and allow optimizing of context, location, channel and the promotion of product offerings. Segmentation of customers using Big Data analytics[55] make it possible to focus on the most valuable customers, ensuring their engagement and loyalty. Other potential Big Data levers in retail include location-based marketing, in-store behavior analysis, customer sentiment analysis, assortment and price optimization, operational performance optimization, supply chain improvements and development of new business models, all using the wealth of data gathered and the power of Big Data analytics.

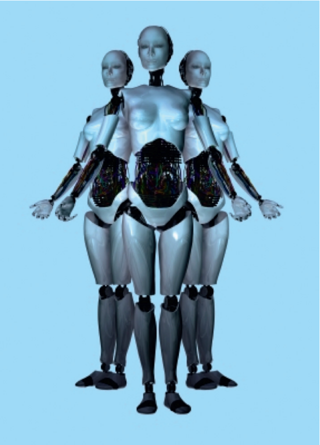
Big Data and manufacturing

Low hanging fruit benefits of Big Data in manufacturing are early detection of product defects and processes and improving supply planning. Data analytics deployed on after-sales data from sensors, service desk records and social media in real time allows triggering pro-active after-sales services and early detection of manufacturing or design issues. In fact, Big Data has an impact across the whole manufacturing life cycle. Starting with product design, by extracting insights from customer data, analytics provide the insight needed to refine product design-to-market and develop new specs for next generation models. A further step would be to involve the suppliers and customers in product innovation. Internet and proliferation of broadband made sourcing and sharing data through virtual collaboration sites feasible. This decade we entered into the prosumer era[56] where the customer is co-creating the supplier product. Examples are Lego creating customized robots out of programmable bricks, Threadless designing of t-shirts by the prosumer crowds, and Procter and Gamble, supported by Innocentive.com, using crowds of experts to solve technical challenges. These open innovation methods were so successful that the biggest challenge became how to extract the most valuable insight from the huge amount of data collected. This is where data mining, Big Data algorithms and analytics deliver the added value of product design with crowd sourcing.

Just-in-time manufacturing can best be implemented when manufacturers enhance their forecasts of customer demand and adjust their supply planning accordingly, using analytics not only to exploit gathered own data, but also data sources of retailers (specifying products, prices, sale volumes, …), dynamics of products (introduction phase, peak sale period, exit phase) and regular near real-time inventory updates (e.g. available stocks and sales per region or channel). This just-in-time on-demand approach delivers the biggest macro-economic benefit to all players involved in the supply chain: the customers are served on time with the lowest possible costs, suppliers are not running the risk of huge unused stocks and manufacturers invest just-in-time[57]. As we have seen in the retail paragraph above, the price competition in retail is going to be increasingly fierce and the only sustainable solution for manufacturers and their suppliers will be to use data to create collective intelligence and improve service quality, production logistics and minimize waste thanks to a better tuning of supply to demand and thereby better meeting the customer needs. Last but not least is the impact of Big Data on the development of flexible manufacturing and virtual digital factories through the analysis of data gathered by sensors and other means in the total production process (from raw material to recycling) to create process transparency, produce dashboards, and identify issues. Big Data analytics is expected to enhance the throughput of those virtual digital factories and to enable the necessary competitive edge of mass customization.

Two other sectors: Finance and Mobility

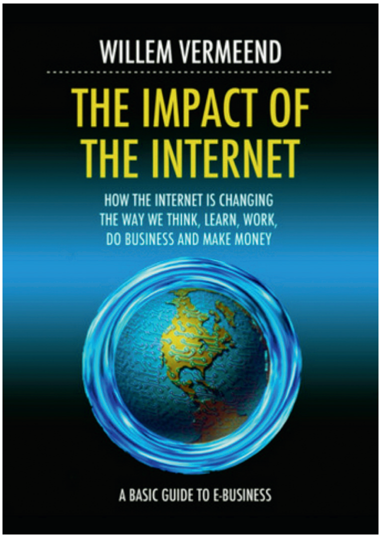
As in the other sectors, Big Data analytics has the potential to offer banks[58] and insurance companies[59] sharper business intelligence by providing better insight and discover patterns through queries, analysis, reporting, dashboards and scorecards. Scenario modeling and predictive analysis help predict behavior and improve decision making. Pattern recognition helps to detect fraud and minimize related losses while scenario modeling enables better risk management. Like in retail above, aggregation of customer information enables banks and insurance companies to determine which customers are most profitable long term, develop loyalty programs and raise the value of their services.



Big Data is also related to our ever increased mobility. Since we are living in a connected world, people's growing mobility is providing a wealth of data through internet on location and behavior. The same applies to other moving objects of interest such as goods, animals, assets and diseases. Sensors and GPS devices continuously provide us with location data combined with other contextual information and registration and correlation of this wealth of information is expected to lead to substantial added value in various sectors like transport and mobility, marketing and advertising, urban planning, disease control, environmental protection and use of sustainable energy.

What about small and medium enterprises (SMEs)?
With all the above mentioned Big Data applications, the logical Big Data users would be within large companies, which typically have their own departments for data analysis. Smaller or start-up companies are generally perceived as having neither the time nor budget to focus on Big Data as described above. Things like a shortage of qualified personnel and a lack of the appropriate software and hardware are the most cited Big Data barriers for SMEs. Still Big Data could be as useful to smaller companies as it is to larger companies. A growing number of technologies is offering SMEs access to cost-effective, sophisticated data analytics[60]. In that way, smaller businesses can also benefit from the creation, collection and analysis of Big Data. The budgets, tools and man-hours will just be more limited than with large organizations.

One of the central questions for every SME remains: if Big Data can add to the internet strategy, how can this best be organized? Generally speaking, a company will need to hire dedicated, specialized people to store and analyze the data and on that basis reach reliable strategical conclusions. This requires special skills. For larger companies this is mostly a matter of assigning responsibilities and perhaps a small expansion of already existing 'datawarehouse' departments. For smaller companies, concise external advice on how to set up Big Data in a simple and effective manner may be the best route to a good start. A number of practical examples on Big Data can be found on *www.ebusinessbook.nl*.



WILLEM VERMEEND

THE IMPACT OF THE INTERNET

HOW THE INTERNET IS CHANGING THE WAY WE THINK, LEARN, WORK, DO BUSINESS AND MAKE MONEY

A BASIC GUIDE TO E-BUSINESS

6 Finally: setting the right expectations

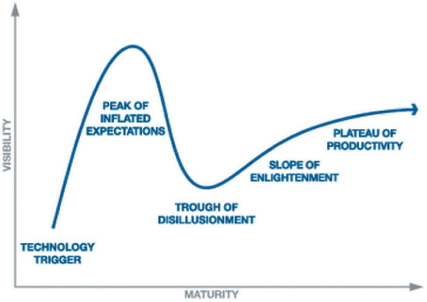
We are still at the advent of the Big Data age and two critical factors will have a decisive impact on the speed at which Big Data will deliver its promise. The first one is our limited ability to implement Big Data in the wide extent of its promising levers due to the explosive need for skilled Big Data DevOps[I]. McKinsey[14] calculated in 2011 that the USA alone will face a shortage of between 140,000 and 190,000 professionals in the next five years. A more recent study from Gartner[61] claims that 4.4 million skilled people worldwide will be needed to support Big Data by 2015. The ability to harness the Big Data potential will therefore ultimately depend on the ability of companies and societies to deliver professionals who are able to understand how to store and process huge amounts of data, develop and run data analytics, extract the needed insights and build predictive modeling. This also requires a good understanding of information technology and of business models and that is a lot of skills in one person. So to help addressing those needs, universities and training institutes need to develop crash programs to supply the right people within the window of competitiveness.

The other critical factor for Big Data is privacy and security. While a tremendous value can be unlocked by sharing all data we ever produce, and creating transparency about it, loss of privacy as well as manipulation by entities who are able to know more about individual persons than they know about themselves[62], remain a huge risk for citizens. In 2012, the European Data Protection Agencies (EDPA) issued a warning to Google about its plan to launch a new integrated platform that would allow better tracking of the users' activities using Big Data processing techniques. The French CNIL[63], acting on behalf of the EDPA, expressed "strong doubts about the lawfulness and fairness of such processing, and about its compliance with European Data Protection legislation". Another danger of Big Data is that analysis of personal data could be misused by government officials or hacked by criminals, thus strongly harming the security and interests of citizens. Existing ethics, legislations and market mechanisms will not be sufficient to guarantee citizen privacy and freedom. Revised ethics, regulations and rules to enforce privacy and security of data by law are therefore pivotal to the adoption and legitimacy of Big Data practices. A big challenge to policymakers is keeping pace with the rapid evolution of technology and its clever use. Training the policymakers and educating of the public about ethics[64], their rights and Big Data threats are prerequisites for a sustainable and legitimate proliferation of Big Data in society and businesses.

[I] *http://en.wikipedia.org/wiki/DevOps*

Big Data and the Trough of Disillusionment

Like every new technology does, Big Data has raised bold expectations and many still need to discern the hype from what is commercially viable. In terms of the Gartner Hype Cycle[I], one can argue that we are approaching the Through of Disillusionment phase, following a period of inflated expectations.



Nassim N. Taleb, professor at New York University's Polytechnic Institute and author of The Black Swan[65], even argues that the bigger the mined datasets are, the harder it becomes to sift through the noise to find the right signal and hence the bigger the chances for misinterpreting the data[66]. The promise of Big Data to work across data sets and silos delivering early clues to hard-to-predict, high-impact 'black swan' events is being questioned due to the disillusionment of existing applications and the negative impact of high noise-to-signal ratio on Big Data predictions. During the next phase of the Big Data hype cycle, we will therefore need to deal with this issue. Improving data signal-to-noise ratio and statistical methods dealing with the impact of noise on predictions should cope with the black swan syndrome. At present, many consumer profiles are partially corrupt or incomplete and many data repositories are full of noise and as a result, many of the personalized Big Data advertisements or direct marketing are ineffective or inaccurate. This in contrast with many successful Big Data application in areas mentioned above like healthcare, fraud detection, insurance and energy management. Big Data developers will need to deal with this phase of disillusionment and improve the technology to the satisfaction of the early Big data adopters and develop second- and third-generations of Big Data products.

Big Data in the Netherlands

A recent survey[67,68] on the use of Big Data in The Netherlands conducted by Keala and The METISfiles shows that 6% of Dutch enterprises with more than 50 employees has already adopted the idea, while 7% is still in the start-up phase of a Big Data project and 6% examines the possibilities for a Big Data initiative. The strongest growth of Big Data projects is expected in 2015. The size of the Dutch Big Data market was estimated at 176 million in 2012, a growth of 48% compared to 2011. This year (2013), the market is expected to grow by 52%.

SURFsara hosts one of the largest Hadoop clusters in The Netherlands. This publicly available cluster allows researchers in The Netherlands to process large data volumes by making use of the MapRe-

[I] *http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp* duce framework. In addition, SURFsara hosts a relatively small NoSQL cluster making it possible to make use of document and key-value stores for research purposes.

SURFsara not only provides Big Data infrastructure but also provides consultancy, training and support to the Dutch academic and research institutions as well as to businesses during the pre-competitive phase. The interest for Big Data techniques is growing rapidly within The Netherlands as the use of Hadoop and the demand for databases like HBase and MongoDB increases. In addition, the need for more training and consultancy is growing at a fast rate as well.

(See also Appendix)

7 Conclusion: Big Data - the road ahead of us

Today, huge repositories of structured, semi-structured and unstructured data collected across various digital platforms, social media and blogs or generated through simulation and modeling are at our disposal. These mass repositories are beyond the abilities of traditional database methods to analyze and understand effectively. Commoditization of High Performance Computing and mass storage in conjunction with cloud computing, open source software and platform interoperability made it possible to deploy data analytics techniques in order to cope with data volume, velocity and variety and to provide the insight needed to really benefit from this data deluge. The value of data at our fingertips is largely underestimated and unexploited today and in almost every sector, including science, health, e-commerce, government, energy, environment, and manufacturing, many applications need to be developed in order to deliver the promise of Big Data. Our lives will consequently be changing rapidly and a whole new way of science and business will be added to existing ones. Correlations and predictions will pave their way into data analysis next to causation, modeling and theories.

The biggest challenge does not seem to be the technology itself - as this is evolving much more rapidly than humans – but rather how to make sure we have enough skills to make effective use of the technology at our disposal and make sense out of the data collected. And before we get to that stage, we need to resolve many legal issues around intellectual property rights, data privacy and integrity, cyber security, exploitation liability and Big Data code of conduct. Like in many other technological areas, customs and ethics around Big Data possibilities and excesses take time to develop. Promises of Big Data include innovation, growth and long term sustainability. Threats include breach of privacy, property rights, data integrity or personal freedom. So provided Big Data is exploited in an open and transparent manner, delivery of the promise of Big Data is not far ahead of us.

Appendix

Big Data @SURFsara

Hardware, software and datasets

SURFsara's Hadoop cluster currently consists of 90 nodes. Each node has a dual quad-core CPU (AMD Opteron 6128) and 64GB memory with 4x 2TB hard disks. The nodes are connected with aggregated links of 2x1Gb. The total raw HDFS storage space of the cluster is 633 TB. The total CPU available for MapReduce is 720 cores. In 2014 the cluster will be extended to about 150 nodes and 1.2 PB of storage.

NoSQL Cluster

SURFsara currently hosts a cluster of seven nodes each containing 24 cores, 132 GB RAM and storage space of 12 TB. This cluster makes it possible for researchers to use several NoSQL stores like Riak, MongoDB, CouchDB and others. A few of these are already used by some users in combination with SURFsara's Grid computing services. The cluster will soon be made publicly available to the research and small business communities.

YARN

SURFsara intends to upgrade its Hadoop MapReduce software implementation to the first stable release of Apache Hadoop MapReduce 2.0 (YARN) [I] as soon as a stable version becomes available. YARN allows Hadoop to be extended by other frameworks than just MapReduce, making it possible to use other parallel computing frameworks such as MPI, Hama, HBase and graph processing frameworks like Giraph all on the same cluster and file system. YARN allows researchers to combine classic High Performance Computing techniques (MPI) with new frameworks for e.g. Graph Processing and NoSQL database technology like HBase.

HBase and NoSQL

As mentioned above, there is a growing interest in NoSQL databases like HBase and MongoDB. SURFsara is participating in an NWO program together with the Dutch national research institute for mathematics and computer science (cwi.nl) and the Royal Library (KB.nl), which will

[I] *http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html* result in setting up an HBase cluster and a few MonetDB nodes for storing and querying large amounts of web archives. Also, as part of SURF's Enlighten Your Research Contest[I], explicit interest from the group of Spinoza prize winner Piek Vossen was expressed for HBase and Big Data solutions.

Common Crawl and WikiPedia

SURFsara hosts a number of (semi-)publicly available data sets on its Hadoop cluster, such as the Common Crawl Dataset and the Dutch and English versions of Wikipedia. In addition, the datasets for the Text Retrieval Challenge (TREC) and ClueWeb are also available for researchers to use. Of course, users are able to upload their own datasets to the cluster.

Norvig Award

SURFsara and the Common Crawl foundation (*http://commoncrawl.org/)* jointly organize the Norvig Web Data Science Award[II], named after Peter Norvig, director of Research at Google Inc. and member of the advisory board of Common Crawl. The goal of the Common Crawl Foundation is to democratize access to the web by producing and maintaining an open repository of web crawl data that is universally accessible and analyzable. The web could give us a tremendous insight when we are able to understand it better. With web crawl data users can spot trends, identify patterns in politics, economics, health, popular culture and many other aspects of life. Contestants are therefore invited to make creative use of the Common Crawl open data set hosted at SURFsara's Hadoop cluster, containing six billion web pages. SURFsara offers access to its cluster, making operations on this dataset of about 25 TB very easy. This Challenge is accompanied by training and support provided by SURFsara's consultants. Subjects investigated could be for instance: how many pages in the Common Crawl data are spam, or what are the most controversial pages in Common Crawl or how wide are networks of linked pages discussing a certain event? The first Norvig Award Competition was organized in 2012 and was a great success.

User communities

SURFsara's Hadoop cluster is actively used by virtually all groups in The Netherlands which are active in information retrieval and data mining. Other research disciplines are soon expected to follow as awareness of this technology is spreading rapidly. SURFsara will proactively reach

[I] *http://www.surf.nl/eyr/*

[II] *http://norvigaward.github.io/learnmore.html* out to other research disciplines and extend its user base. In particular, in 2014 emphasis will be put on attracting more users from the Life Science and Natural Language Processing communities. Since one of the primary challenges in Big Data is having skilled personnel, SURFsara is organizing and giving Big Data courses for academia as well as for Dutch companies and through training agencies.

Knowledge Transfer to the market

SURFsara is working closely with its spin-off Vancis B.V. on setting up and operating a Hadoop cluster for commercial applications. An agreement for knowledge transfer between SURFsara, Vancis and KPMG was set up in the course of 2013. In addition, Vancis participated in the Big Data Tooling Challenge organized by SURFsara in the fall of 2013. In this challenge small businesses were invited to come up with datasets and a challenging Big Data problem. SURFsara offered its expertise and infrastructure during the course of this challenge in order to tackle a number of pre-selected problems.

Apart from Vancis, companies and organizations like Dynamic Credit, Webpower, Belastingdienst, Lucifer, Dacolt, 2Coolmonkeys and Metaphora actively participated in this challenge. In relation to this, SURFsara's Big Data activities have also sparked interest from NFI (Netherlands Forensic Institute) and the National Police.

In 2014 the Big Data Tooling challenge will probably be repeated and extended. Also the contacts with organizations like NFI and NP will be strengthened.

About the authors

Dr. Anwar Osseyran[I] is the managing director of SURFsara (the national High Performance Computing Centre) and member of board of directors of various ICT companies.

Prof. Dr. Willem Vermeend[II] is internet entrepreneur, investor and member of the supervisory board of several (internet) companies.

Anwar Osseyran and Willem Vermeend are working closely together within the context of the Breakthrough projects of the Ministry of Economic Affairs on the development of concrete Big Data applications for small and medium enterprises in the Netherlands. Prof. Vermeend is the so-called 'figurehead' of this project.

Through the Breakthrough Projects, the Dutch government aims at developing public-private partnerships in which industry and research work together on removing obstacles and deploying ICT for economic growth, better competitiveness and increased innovation. Examples of obstacles are lack of knowledge, lack of standardization or lack of technology transfer between academia and industry.

[I] *http://www.linkedin.com/in/osseyran*

[II] *http://www.linkedin.com/in/willemvermeend*

References

[1] www.factual.com; http://www.inrix.com; http://gnip.com; www.infochimps.com

[2] www.yelp.com; https://foursquare.com; www.trulia.com; www.blockbeacon.com; http://spindle.com

[3] http://www.gbv.de/dms/ilmenau/toc/025308912.PDF, blz. 424.

[4] See also Norvigs' "The unreasonable effectiveness of data", http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/ja//pubs/archive/35179.pdf

[5] http://humanfaceofbigdata.com/ see also www.bbc.co.uk/news/technology-21535739

[6] http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf

[7] http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt

[8] www.skatelescope.org/

[9] http://lhc.web.cern.ch/lhc/

[10] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3572967/

[11] http://www.technologyreview.com/featuredstory/508836/how-obama-used-big-data-to-rally-voters-part-1/

[12] http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[13] http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

[14] http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

[15] www.economist.com/node/15579717

[16] http://hadoop.apache.org/

[17] http://nosql-database.org/

[18] www-01.ibm.com/software/data/netezza; www.vertica.com; www.greenplum.com,; www.calpont.com/; www.exasol.com/; www.paraccel.com;

[19] http://www.youtube.com/eventsatgoogle#p/u/5/qBaVyCcw47M

[20] http://highlyscalable.wordpress.com/2012/05/01/probabilistic-structures-web-analytics-data-mining/

[21] http://blog.monitis.com/index.php/2011/05/22/picking-the-right-nosql-database-tool/

[22] http://hadoop.apache.org/

[23] http://www.pig.apache.org/

[24] http://hbase.apache.org/

[25] http://cassandra.apache.org/

[26] http://dl.acm.org/citation.cfm?id=1934385

[27] http://storm-project.net/

[28] http://www.kdd.org/sites/default/files/issues/14-2-2012-12/V14-02-04-Kang.pdf

[29] https://github.com/facebook/scribe/

[30] *http://www.cascading.org/*

[31] *http://select.cs.cmu.edu/code/graphlab/index.html#overview*

[32] *http://www.r-project.org/*

[33] *http://ailab.ijs.si/dunja/TuringSLAIS-2012/Papers/Bifet.pdf*

[34] *http://hunch.net/~vw/*

[35] *http://mapreduce.org/*

[36] *http://hadoop.apache.org/*

[37] *http://www.cs.uvm.edu/~icdm/*

[38] *http://www.kdd.org/conferences*

[39] *http://www.ecmlpkdd.org/*

[40] *http://expandedramblings.com/*

[41] *http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html*

[42] *http://www.cmu.edu/homepage/computing/2010/fall/nell-computer-that-learns.shtml*

[43] *http://www.manning.com/marz/*

[44] *http://aws.amazon.com/elasticmapreduce/*

[45] *https://github.com/GoogleCloudPlatform/solutions-google-compute-engine-cluster-for-hadoop*

[46] *http://wiki.apache.org/hadoop/Virtual%20Hadoop*

[47] *http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation, blz. 33-36*

[48] *http://public.dhe.ibm.com/common/ssi/ecm/en/iml14296usen/IML14296USEN.PDF, blz. 109-112*

[49] *http://techcrunch.com/2013/05/13/twitter-acquires-big-data-visualization-startup-lucky-sort-service-to-shutter-in-months-ahead/*

[50] *http://www.crn.com/news/applications-os/240156767/datawatch-boosts-big-data-visualization-capabilities-with-acquisition.htm*

[51] *http://www.lsst.org/lsst/about*

[52] *http://simbad.u-strasbg.fr/simbad/*

[53] *http://www.arduino.cc/*

[54] *http://practicalanalytics.wordpress.com/2012/01/05/analytics-case-study-schwan-foods/*

[55] *http://www.mindofmarketing.net/2007/05/customer-segmentation-why-exactly-does.html#.UhnH0hvIb_Y*

[56] *http://www.forbes.com/sites/work-in-progress/2010/07/03/the-shift-from-consumers-to-prosumers/*

[57] *http://www.controleng.com/industry-news/more-news/single-article/process-risk-assessment-uses-big-data/632b3ce8d25102b9ab558b3833cc5885.html*

[58] *http://www.banktech.com/business-intelligence/putting-big-data-to-work-for-financial-s/240153177*

[59] *http://www.businessinsurance.com/section/NEWS040105#*

[60] *http://www.forbes.com/sites/capitalonespark/2013/05/30/what-can-big-data-do-for-a-small-business/*

[61] *http://www.gartner.com/newsroom/id/2207915*

[62] *http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all*

[63] *http://www.cnil.fr/english/news-and-events/news/article/googles-new-privacy-policy-raises-deep-concerns-about-data-protection-and-the-respect-of-the-euro/*

[64] *http://it-ebooks.info/book/1984/*

[65] *http://www.riosmauricio.com/wp-content/uploads/2013/05/Taleb_The-Black-Swan.pdf*

[66] *http://www.wired.com/opinion/2013/02/big-data-means-big-errors-people/*

[67] *http://www.themetisfiles.com/wp-content/uploads/2013/09/Big-Data-Big-Decisions-Market-Dynamics-Inhoudsopgave.pdf*

[68] *http://www.themetisfiles.com/wp-content/uploads/2013/09/Big-Data-Big-Decisions-Eindgebruikersrapportage-Inhoudsopgave.pdf*