Behavioural and Societal Sciences
Brassersplein 2
2612 CT Delft
P.O. Box 5050
2600 GB Delft
The Netherlands

www.tno.nl

T +31 88 866 70 00
F +31 88 866 70 57
infodesk@tno.nl

TNO report

Thriving and surviving in a data-driven society

Date 24 September 2013

Author(s) Jop Esmeijer
Tom Bakker
Silvain de Munck

Number of pages 69 (incl. appendices)
Number of appendices 3
Sponsor Ministerie van Economische Zaken
Project name BTK 2013 – Effectie waardenetwerken
Project number 060.01752

All rights reserved.
No part of this publication may be reproduced and/or published by print, photoprint, microfilm or any other means without the previous written consent of TNO.

In case this report was drafted on instructions, the rights and obligations of contracting parties are subject to either the General Terms and Conditions for commissions to TNO, or the relevant agreement concluded between the contracting parties. Submitting the report for inspection to parties who have a direct interest is permitted.

© 2013 TNO

Contents

1 Introduction

The “new oil” is a popular metaphor to describe big data. It captures the promise of tremendous economic and societal value of a new resource when processed and combined in smart ways. Like oil, data in its raw form does not in and of itself represent much value. The path from raw data collection to actionable insights demands a value creation process in which raw data is converted to information or knowledge, which in turn may yield important economic and societal benefits.

To create value from big data, organizations need to determine what datasets need to be captured and what technologies and techniques are required for data integration, storage, analysis, and visualization. An important strategic consideration for any organization is to what extent they will maintain the data assets themselves, and when they will have to partner with other players such as technology vendors, data-owners, data-brokers, or legal experts.

As the emergence of big data analytics is likely to affect all sectors[1], multiple data platforms are being created that, combined, form a trans-sectoral (big) data ecosystem bringing together players, technologies and data. Such a data ecosystem may not only facilitate incremental innovations that are based, for instance, on the available data within a single organization, it also enables more radical transformations that require data from multiple sources, across multiple sectors, creating complex networks of public and private stakeholders.

Relatively little is known about the constellation of the emerging big data landscape. In order to assess how big data impacts innovation, entrepreneurship and core values, such as ‘right to play’ and privacy, policymakers and other decision makers need to understand how an entirely new data ecosystem is being created through the acts of existing and emerging players through technological solutions and standards, and through regulatory action and ideas.

The objective of the study is twofold.

Firstly, drawing on expert- and stakeholder interviews, desk research and workshops, an attempt is made to describe the complex big data value creation process and the associated process of data-driven innovation. Important drivers, goals and barriers related to data-driven innovation are highlighted.

Secondly, the study aims to reveal core elements of the emerging big data ecosystem: the prominent types of players, the main products and services, the key technologies, and finally its market dynamics.

[1] Manyika, J., et al (2011). Big Data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute. Available at: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Big Data Innovations
An illustrative example of a company that turned data into a valuable asset is John Deere, a manufacturer of tractors and other kinds of equipment for farmers. Some of its most recent series of tractors are enhanced with sensors that collect data from the machine, the soil and the crops it processes. The data that it collects is then analyzed and augmented with additional data about the weather and crop features, and presented on an online platform that can be used on both desktop computers and handheld devices. The results from the analyses support farmers, for instance, in deciding what to sow, where to sow and when to sow in order to yield the best results.[2] The data constellation that provides this product extends far beyond the traditional value chain of John Deere tractors, or any tractor for that matter. It now includes sensor technology, data-integration from external sources, real-time data processing and visualizations for multiple devices. It requires collaboration between many different kinds of expertise. The result is the development of innovative services that explore new territory in an unconventional way. John Deere profoundly changed the use of information for its customers.

Another example is the use of analytics by retail companies such as Walmart and Amazon to better target their customers. Walmart developed most of its big data tools in its own Walmart Labs, which it created after acquiring internet company Kosmix in 2011.[3] It combines (and crunches) its own proprietary data on customer behavior (e.g. purchasing history), combined with social media data and public data on the web, to create a so-called Social Genome of its customers that is used for marketing purposes. It has since open sourced some of the tools it developed for mobile devices, so they can be used and modified by other organizations.

1.1 Background: big data and the changing data landscape

1.1.1 Big data: what is new and what is not?

Data is nothing new, nor is a lot of data, nor is the analysis of a lot of data for strategic purposes, a practice that is often referred to as business intelligence.[4] So where is the new wine?[5]

Big data is often defined as data that cannot be handled by standard IT solutions.[6] However, this definition defines big data solely as a technological challenge, a problem that needs to be solved. It does not acknowledge nor even hint at the idea of data as a new source that can be put to use, or as a phenomenon that poses important societal issues, as will be discussed later on. It is important to recognize that the concept of big data is not just a technological development, as boyd and

[2] Big Data Startups. John Deere is revolutionizing farming with big data. Available at: http://www.bigdata-startups.com/BigData-startup/john-deere-revolutionizing-farming-big-data.
[3] http://www.walmartlabs.com/
[4] Chen, L., et al. (2012) Business Intelligence and Analytics: From Big Data to Big Impact. In: *MIS Quarterly*.Vol. 36 No. 4, pp. 1165-1188.
[5] Agrawal, D., Das, S., El abbadi, A. (2010). Big Data and Cloud Computing: New Wine or just New Bottles? Presented at: *The 36th International Conference on Very Large Data Bases*, September 13-17, Singapore.
[6] Dumbill, E. (2012). Big Data Market Survey. In: O'Reilly Rader Team. *Planning for big data*, pp. 23-34. Available at: http://oreilly.com/data/radarreports/planning-for-big-data.csp

Crawford[7], and Cukier and Mayer-Schönberger point out,[8] but also refers to the significant societal and economic transformations and impact it may bring about.

Of course technology plays a crucial role, and especially the growing affordability of technology. With the rise of cloud services and open source platforms, prices of high performance computing and data storage dropped significantly. This enables even small companies to capture, store and analyze vast amounts of data. Other important developments are the emergence of (new) infrastructures such as mobile devices, the Internet of Things - where objects from the physical world are connected to the Internet and, subsequently, data is captured with sensors, GPS, RFID and smart objects - and social networks where we share our daily social interactions, thoughts and opinions. These infrastructures are partly responsible for the so-called data explosion in recent years, which is the driver of another definition of big data in which it is described as a combination of its main characteristics; the three 'V's' of volume, velocity and variety.[9]

But these technological developments are only one side of the story. Big data and its possibilities are one dimension of a new paradigm (some say a myth[10]). Big data inspires a data-driven culture in all kinds of domains and organizations: more data is better than less data, is the common notion. According to Kaisler et al., this attitude has, for some users, become an addiction that becomes stronger as they acquire more data[11]. New York Time's David Brooks described it as follows:

> *"If you asked me to describe the rising philosophy of the day, I'd say it is data-ism. We now have the ability to gather huge amounts of data. This ability seems to carry with it certain cultural assumptions — that everything that can be measured should be measured; that data is a transparent and reliable lens that allows us to filter out emotionalism and ideology; that data will help us do remarkable things — like foretell the future."*[12]

In order to determine what is new, and what is not, it is useful to look beyond definitions and descriptions, and consider the opportunities and questions that are driven by the above-mentioned developments.

On the one hand there are known questions that are already being answered by more traditional forms of business intelligence (or gut-feeling). Questions such as "How can we better segment customers?" may be answered by collecting and

---

[7] boyd, d. & Crawford, K. (2011). Six provocations for Big Data. Presented at: *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, September 2011.*

[8] Cukier, K. & Mayer-Schönberger, V. (2013). Big Data: a Revolution That Will Transform how we Live, Work, and Think. Boston: Houghton Mifflin Harcourt Publishing Company.

[9] Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety. In: *META Group*, 6 February, Available at: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

[10] boyd, d. & Crawford, K. (2012). "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." In: *Information, Communication, & Society* 15:5, pp. 662-679

[11] Kaisler. S., et al. (2013) Big Data: Issues and Challenges moving forward. In: *IEEE Computer Society*, pp. 995-1004.

[12] Brooks, D. (2013). The Philosophy of Data. In: *New York Times*. Available at: http://www.nytimes.com/2013/02/05/opinion/brooks-the-philosophy-of-data.html?_r=0

analyzing lots of fast and varied data. Big data enables us "to see things at a large scale we cannot see at a small scale."[13]

But big data also raises new questions. Some questions frame big data as a technological challenge, for instance: "How to store and process vast quantities of real-time sensor data?" or "How to analyze unstructured texts?" But big data might also enable organizations to formulate and answer new questions that were not possible to ask before. Additionally, it enables organizations to rethink their business models on a more fundamental level, not focusing on incremental enhancements, but rather on more transformative innovations, as described above in the case of John Deere. Other sets of questions are not so much addressed by big data technology but rather triggered by it, for instance: "How to implement a data-driven mindset in my organization?" Or "What does big data mean for privacy?" These issues require organizational or legal expertise. Furthermore, as big data may bring about a restructuring of value networks or even whole sectors, organizations need to re-assess their role and partnerships in a changing landscape where new players emerge and new alliances are forged and forced. These data-driven changes in markets can pose important policy-related questions as well, for instance regarding barriers-to-entry or rapid market entrenchment.[14]

Of course, these issues do not exist in isolation. When an organization defines a goal (for instance: it wants to increase sales by three percent) it could choose to explore the possibilities of data-driven innovation and deploy big data technologies for certain questions. Subsequently, it will encounter technological, organizational or even societal challenges along the way. This variety of issues illustrates the multidisciplinary challenge of big data innovations, especially when the scope of the innovations increases and they do not only affect a single organization, but also have systemic ramifications.

1.1.2 A changing data landscape

On April first, 2013, Datasift - a company that offers social media filtering services - posted a blog on its website in which it unveiled a new "plan" to use carrier pigeons for data delivery to their customers:

> *"We're […] bringing the carrier pigeon into the 21st century and putting millions of feral pigeons to work for the Internet economy!"*[15]

April fools.

The joke illustrates how we have come a long way since Paul Reuter, more than 150 years ago, in fact did use carrier pigeons (and the telegraph) to serve companies with commercial news – information as an important asset for strategic business operations. Reuters is still around, as are similar information intermediaries such as Bloomberg and Forrester. But the landscape of data and information services and related technologies has both expanded and changed rapidly. New players have emerged, such as Google and Facebook. Newer ones

[13] Cukier, K. & Mayer-Schönberger, V. (2013). Big Data: a Revolution That Will Transform how we Live, Work, and Think. Boston: Houghton Mifflin Harcourt Publishing Company, p. 6.
[14] Brown, I. & Marsden, T. (2013). Regulating Code: Good and Better Regulation in the Information Age. Cambridge: The MIT Press.
[15] Barker, T. (2013) Introducing Datasift Pigeon Carrier Delivery. Available at: http://blog.datasift.com/2013/04/01/introducing-datasift-carrier-pigeon-delivery/

like Datamarket and Infochimp arrive on the scene every day. Governments have embraced 'open data' strategies and are contributing on a massive scale as well. The European Commission, for instance, has launched its European Union Open Data Portal, containing almost 6000 datasets.[16] Most of all, organizations are starting to leverage their own data, captured through myriad transactions, production processes and communication, using all manner of technologies for data collection, integration, storage, analysis and visualization.

A driving force behind these developments is that big data is widely regarded as a new source of growth, productivity and innovation; a new type of fuel to fire-up our economy. The concept has captured the imagination of businesses, governments and scientists alike. European Commissioner Neelie Kroes, in the same speech in which she referred to data as the "new oil", likened data to gold and encouraged everyone "… to start digging".[17] The European Commission calculated that opening up its data will provide a 70 billion Euro boost to the economy. Data and data analytics are expected to improve both business and public services to benefit society and empower citizens. The White House in 2012 presented its 200 million dollar research initiative that "promises to transform our ability to use big data for scientific discovery, environmental and biomedical research, education, and national security."[18]

Big data has sparked a management revolution according to Brynjolffson and McAffee, who discern a "fundamental transformation of the economy"[19] in which businesses that embrace these new possibilities for better decision-making are likely to outperform their competitors. Of course, these views are also heavily promoted by technology vendors such as IBM, Oracle and Microsoft. These companies herald data and especially related technologies (*their* technologies) in white papers and portfolios of big data success stories. Even at the level of individuals big data is catching on. The so-called *Quantified Self*-movement has inspired its followers to use of all kinds of tools, like the Fitbit, to track their every move and heartbeat, hoping that data analytics and feedback will improve their health and overall wellbeing. All-in-all, the potential of big data seems enormous.

Still, there are also important risks that need to be addressed even though the European Commission emphasized that big data should be seen as an opportunity and not as a problem.[20] Privacy is obviously a major concern, as are other civil liberties and consumer freedoms.[21] [22] But so are hidden biases in datasets and

[16] http://open-data.europa.eu/en/data
[17] Kroes, N. (2013). Data is the new gold. Opening Remarks, Press Conference on Open Data Strategy. Available at: http://europa.eu/rapid/press-release_SPEECH-11-872_en.htm
[18] Cohen. R. (2012) The White House is Spending Big Money on Big Data. In: *Forbes*. Available at: http://www.forbes.com/sites/reuvencohen/2012/05/13/the-white-house-is-spending-big-money-on-big-data/
[19] Brynjolfsson, B., and McAffee, A. (2012) Big Data: The management revolution. In: *Harvard Business Review*, October. Available at: http://hbr.org/product/big-data-the-management-revolution/an/R1210C-PDF-ENG
[20] European Commission (2012) Big Data at Your Service. Available at: ec.europa.eu/information_society/newsroom/cf/dae/itemdetail.cfm?item_id=8337
[21] Cukier, K. & Mayer-Schönberger, V. (2013). Big Data: a Revolution That Will Transform how we Live, Work, and Think. Boston: Houghton Mifflin Harcourt Publishing Company.
[22] Bollier, D. (2010). The Promise and Peril of Big Data. Washington DC: The Aspen Institute. Available at: http://www.aspeninstitute.org/publications/promise-peril-big-data

algorithms[23], the tendency to use data to treat symptoms by making broken systems more efficient rather than solving underlying problems[24], and a false sense of safety in numbers as big data (and spurious statistical correlations) can also lead to Big Mistakes.[25] [26] [27] [28] As Mike Loukides from O'Reilly emphasizes:

> *"Perhaps the most dangerous is the technologist who never understands the limitations of data, never understands what data isn't telling you, or never understands that if you ask the wrong questions, you'll certainly get the wrong answers."*[29]

More than ever companies and governments rely on data and analytics to make better decisions, to target customers and citizens, or to create new products and services. A widely cited McKinsey report on big data from 2011 stated that:

> *"[…] much of modern economic activity, innovation, and growth simply couldn't take place without data. Digital data is now everywhere—in every sector, in every economy, in every organization and user of digital technology."*[30]

"Software is eating the world", wrote venture capitalist and Silicon Valley veteran Marc Andreessen[31], and the world is served in big chunks of data.

1.2 Research questions

In order to better explore and interpret both the economic and societal value of big data and its risks, a deeper understanding is required of the value creation process of data and data-driven innovation, the interaction between data-driven innovation and the transcending big data ecosystem and the players, technologies and dynamics that structure this ecosystem. As Mark Graham from the Oxford Internet Institute states after describing the increasing importance of information in our society:

> *"[…] It is important to understand who produces and reproduces, who has access, and* who *and* where *are represented by information in our contemporary knowledge economy."*[32]

---

[23] Wakefield, J. (2012). Can we trust the code that increasingly runs our lives? In: *BBC*. Available at: http://www.bbc.co.uk/news/technology-19347122
[24] Morozov, E. (2013) To Save Everything Click Here: the Folly of Technological Solutionism. New York: PublicAffairs.
[25] Taleb, N. (2013). Beware the Big Errors of Big Data. In: *Wired*. Available at: http://www.wired.com/opinion/2013/02/big-data-means-big-errors-people/
[26] boyd, d. & Crawford, K. (2011). Six provocations for Big Data. Presented at: *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, September 2011.* Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431
[27] Silver, N. (2012). The signal and the noise. New York: Penguin Press.
[28] Manovich, L. (2011) Trending: The Promises and Challenges for Big Data. Available at: http://www.manovich.net/DOCS/Manovich_trending_paper.pdf
[29] Loukides, M. (2013) Big data is dead, long live big data: thoughts heading to Strata. In: *O'Reilly Radar*. Available at: http://radar.oreilly.com/2013/02/big-data-hype-and-longevity.html
[30] Manyika, J., et al (2011). Big Data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute, p.4. Available at: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation
[31] Andreessen, M. (2011) Why Software is Eating the World. In: *WSJ*. Available at: http://online.wsj.com/article/SB10001424053111903480904576512250915629460.html
[32] Graham, M. (2010) The Knowledge Based Economy and Digital Divisions of Labour. Available at: http://www.geospace.co.uk/files/The_Knowledge_Based_Economy.pdf

This report, therefore, addresses the following specific research questions:

1 What are the main **drivers** for the implementation of a data-driven strategy and big data technologies?
2 For what types of **goals** do organizations leverage data?
3 What does the **value creation process of data** look like?
4 What does the **big data ecosystem** look like?
    (a) What kinds of players and products can be discerned?
    (b) What types of organisations are dominant players?
    (c) What market dynamics can be discerned (competition, vertical and horizontal integration)?
    (d) How could this evolve in the future?
    (e) How does the big data ecosystem impact data-driven innovation?
5 What are the most important **barriers** in data-driven innovation?

2 Method

Despite the popular interest in big data and the diversity of big data initiatives, empirically-based research in this field is still rare. Drawing on actual big data practices, workshops with stakeholders and the experiences from data experts, this study aims to provide a more comprehensive view of the value creation process of data and the composition of a nascent big data ecosystem.

2.1 Data collection

Given the variety of research questions, the study decided to employ a multi-method approach. This allowed the collation and integration of results from extant studies, white papers and news media (desk research), outputs from interviews with experts and stakeholders, and, finally, insights obtained in two big data workshops.

Table 1 provides an overview of the research questions and the associated methods. In the following paragraphs, more detailed descriptions are provided.

Table 1 Research questions and associated methods

Research question	**Primary method**	**Complimentary method**
1 & 2 (drivers and goals)	*Interviews data owners* *Interviews data experts* *Desk research*	*SME workshop* *Stakeholders workshop*
3 (value creation process)	*Interviews data owners* *Interviews data experts* *Stakeholders workshop*	*SME workshop* *Desk research*
4 (big data ecosystem)	*Interviews data owners* *Interviews data experts* *Desk research*	*SME workshop* *Stakeholders workshop*
5 (barriers)	*Interviews data owners* *Interviews data experts* *Stakeholders workshop*	*SME workshop* *Desk research*

2.1.1 Interviews

In order to understand drivers and barriers of the big data innovation process, input was collected from people and organizations already dealing with big data. To generate insights from different angles, interviews were conducted with independent data experts, data owners and a data service company.

Independent data experts

Two interviews with independent data experts were conducted.

The first interview was held with Hans Wormer and Oscar Wijsman, program managers at Almere Data Capital. The goal of this program, initiated by the city of Almere, is to bring various stakeholders in the data ecosytem together and to facilitate them to exploit the value of big data. In their own words: “Almere is building an eco-system of companies, education and research facilities that will provide knowledge and services about and for big data. (...) The stakeholders can be classified into several categories: supply and demand participants; knowledge parties; small and medium businesses; patients and citizens; as well as Almere City.”[33] The program’s purpose is to position the city of Almere as *the* big data knowledge center.

SURFnet was the second organization that was interviewed. SURFnet is part of SURF, the Dutch higher education and research partnership for ICT-driven innovation. SURFnet focuses on reliable infrastructures and facilitating effective and innovative use of IT and data in research and education. Specifically, they focus on two areas: “A network infrastructure: a hybrid fixed-wireless network as the basis for all collaboration, providing efficient, unlimited data transport [and] a collaboration infrastructure: a pioneering collaboration environment that seamlessly connects systems, services, tools, and people.”[34] SURF is also partner in the eScience Center, an initiative that aims to “stimulate creative data-driven research across all scientific disciplines [and to] Develop and apply tools to enable data-intensive scientific research [and to] promote knowledge-based collaboration between cross-disciplinary researchers”. [35] The interviewees were Sylvia Kuijpers (Community manager Research) and Harold Teunissen (Department Head Middleware and Security Services).

Interviews with data owners
In order to collect detailed knowledge about the value creation process and the emerging data ecosystem, we selected two organizations increasingly confronted with data related challenges and which are likely to undergo substantial transformation as a result of deploying data.

The first organization is the Port of Rotterdam Authority (Havenbedrijf Rotterdam, shortened to HbR). The Port of Rotterdam Authority manages, maintains and develops the Rotterdam harbor. We interviewed the department Business Analysis & Intelligence. More information on the company and its data related expertise is provided in the results section.

The second organization is CIBAS, the Central Information Management Waste Services. CIBAS is a joint initiative from multiple Dutch waste management organizations which supports local governments in differentiating taxes for its citizens depending on individual household waste disposal. Management and analysis of data has a central place in this process.

Interview with the data company
In addition to the independent experts and data owners and holders, we interviewed Evert-Jan Tromp, Director Indirect and General Business Europe, the Middle East and Africa Business Intelligence at SAP. With BusinessObjects and Crystal, SAP is

[33] http://www.almeredatacapital.nl/english/what-is-almere-datacapital
[34] http://www.surfnet.nl/en/organisatie/Pages/default.aspx
[35] http://esciencecenter.nl/about-the-center/

one of the leading software companies in terms of BI and big data solutions. SAP products are used by both Port of Rotterdam Authority and CIBAS.

2.1.2 Workshops

In addition to the interviews, a workshop was organized in collaboration with the Big Data Center Almere in which a selection of both 'data-owners' from four different sectors (transport and logistics, health, safety and oversight, and energy) and data-service providers was present. The goal of this stakeholders-workshop was:

1 To explore to what extent organizations would be interested in data from other companies, and especially data from other sectors. Why are they interested, or not? Under what circumstances? In what kind of constellation of actors? And what are the most important barriers for these kind of trans-sectoral collaborations in data-driven innovation?;
2 To discuss drivers, requirements and barriers for data-driven innovation.

To explore the possibilities of cross-sectoral collaboration and to identify the dynamics, drivers and barriers in the innovation process, the participants were asked:

- to list all the internal and external datasets available to the organization. All the lists were displayed on the wall for all the participants of the workshop to see;
- to engage in a virtual and small trans-sectoral datamarket. Participants were asked to indicate – using stickers – their interests in datasets from participants from their own or other sectors;
- to form partnerships with 2-4 other participants. These groups should consist of participants of at least two different sectors and one 'data-expert', and were asked to develop a new product or service, using multiple datasets and using each other's expertise;
- to participate in an open group discussion, facilitated by the researchers, to discuss (1) the most salient issues that had arisen during the workshop and (2) to discuss drivers and barriers for establishing data-driven innovations within their own organizations.

Besides the stakeholders-workshop, two of the researchers attended and facilitated a workshop at the Big Data Value Center in Almere. In this workshops, SMEs from various sectors and with different kinds of expertise collaborated in groups to create data-driven innovations. For the researchers, this workshop both produced relevant insights for this study and it helped to fine-tune the stakeholders-workshop, held a few weeks later.

2.1.3 Desk research

The desk research strategy primarily focused on collecting a wide range of recent research, foresight, market reports and (technology) news media.

Most of the documentation collected originated from consultancy firms, public organizations and institutions and research companies. We focused on studies that contained information that was specifically collected for that study, such as the results of a survey. Also documentation of the European Data Forum 2013, which was attended by one of the researchers, was used as input.

2.2 Analytical framework

The analytical framework that is used to structure the research (e.g., the interviews, desk research and the workshops) and the results, is based on two core notions: 1) the existence of a *value creation process of data*, and 2) the emergence of a transcending *big data ecosystem*, connecting data organizations, providers, technologies, standards and ideas.

The concept of an ecological approach to describe business environments was introduced by Moore[36] to describe how companies should not be viewed as members of a single industry "[…] but as part of a business ecosystem that crosses a variety of industries." In these ecosystems, collaborative arrangements of firms combine their individual offerings to create coherent, customer-facing solutions.[37] This seems a very suitable perspective to explore the nature of the multi-faceted big data landscape in which many networks of human and non-human actors, tailored to specific data-driven innovations, create a big data ecosystem. Furthermore, these networks act in a legal and regulatory context that protect certain values but also limit certain possibilities.

The make-up of this big data ecosystem will be described along the lines of the concept of the value creation process of data, which will be described in more detail in paragraph 3.3. Such a process-centric perspective, which was also proposed in a report from the European big data technology platform NESSI[38], offers a framework to describe the different elements that need to be addressed by organizations in the process of data-driven innovation, and how this translates into a concrete network of players and technologies. Although these networks of actors may vary infinitely, the fundamental elements of the value creation process of data provide us with some guidance, as we will try to sketch the outlines of the data ecosystem. It facilitates a structured description of the various key roles (and players that perform these roles) and technologies in the big data ecosystem, its dynamics and how all this relates to data-driven innovation in terms of drivers, barriers and risks.

2.3 Outline of the study

Chapter three contains the results of the study.

In paragraph 3.1 the most important drivers behind the adoption of (big) data analytics are provided. The most common goals for using (big) data are described in paragraph 3.2.

The value creation process of data will be discussed in more detail in paragraph 3.3. We will use the value creation process of data to describe the trajectory of data-driven innovation in two organizations. We will discuss how these innovations were facilitated by a network of players, technologies, ideas, and the most important drivers and barriers during this formation.

---

[36] Moore, F. (1993). Predators and Prey: a new ecology of Competition. In: *HBR,* May-June. Available at: http://blogs.law.harvard.edu/jim/files/2010/04/Predators-and-Prey.pdf
[37] Adner, R. (2006). Match your innovation strategy to your innovation ecosystem. In: *HBR*, April. Available at: http://pds12.egloos.com/pds/200811/07/31/R0604Fp2.pdf
[38] NESSI (2012). Big Data: A New World of Opportunities. Available at: http://www.nessi-europe.com/Files/Private/NESSI_WhitePaper_BigData.pdf

In paragraph 3.4 a broad outline will be provided of the transcending big data ecosystem, which constitutes the actors from the networks that facilitate and shape data-driven innovation. Furthermore, the value creation process will be applied to assess where in this big data ecosystem the most value is added to data, what kinds of players hold key positions, how this could evolve in the future and how this relates to data-driven or data-inspired innovation.

In paragraph 3.5 the key barriers associated with deploying data-driven innovation and big data in general are provided.

Chapter 4 will discuss the most important conclusions.

3 Results

The presentation of the results follows the order of the research questions. After discussing and categorizing the drivers and goals of big data (technology) adoption, the value creation process of data is described. Using the interview data we will discuss two specific cases that illustrate how organizations fulfill this value creation process of data and what the role of data is in their innovation process. Next, the big data ecosystem will be described: its main types of players and products, and the market dynamics that shape it – now and in the near future. Based on the description of the goals, existing networks and the big data ecosystem, we identify the main barriers for companies that (aim to) innovate with big data.

3.1 Drivers: getting started

Various goals may be pursued with the exploitation of big data. Although these goals could constitute the driving forces themselves, the interviews reveal a number of additional, more generic factors that seem to drive organizations to *start* exploring big data driven innovation in the first place.

Affordable and easy-to-use technology
As mentioned in chapter one, the mere availability of raw data in combination with the improved analytical tools may set data driven innovation into motion. The emergence of social media, the integration of sensors in myriad devices and machines, and the increased accessibility of various internal and external data sources have led to a true explosion of data. In their book Information Rules from 1999 Hal Varian, currently Chief Economist at Google, and Carl Shapiro note, a bit underwhelmed, that the usable data on the internet equaled about 150.000 books, an average bookstore.[39] This was only fifteen years ago. In 2010, we produced on a daily basis the same amount of data that all of humankind produced from the first rock-paintings up to 2003.[40]

Because of this increase in data volume, many have rushed to explore the value of that data for their organization. In other cases, data has been impatiently lying around for years but has never been stored or analyzed properly simply because of the lack of affordable and suitable software, especially for organizations other than large enterprises. For example, researchers have long been limited to working with samples or limited data selections, but now the availability of affordable user-friendly analytics tools to process large volumes of data has made it possible to perform similar but more robust calculations. As one of the interviewees noted:

> *"The limitations of our calculations were always the same as the maximum number of rows allowed by Excel".*

Hype and fear of missing out
Almost two decades ago, organizations saw themselves faced with the inevitable emergence of the Internet and the need to have an online presence. About ten

[39] Varian, H. and Shapiro, C. (1999). Information Rules. Boston: Harvard Business School Press, Kindle eBook, Location 222 of 6135.
[40] Siegler, M. (2010). Eric Schmidt: Every 2 Days We Create As Much Information As We Did Up To 2003. In: *Techcrunch*, Available at: http://techcrunch.com/2010/08/04/schmidt-data

years later, they could not ignore the influence of social networks, Twitter and other 2.0 services, thus feeling forced to develop a social media strategy. Similarly, many organizations today have started to experiment with data collection and (big) data analytics, simply because of the 'fear or missing out' or 'because our competitors are doing it', a sentiment fueled by both IT-vendors and media. As Teunissen from SURFnet put it:

> *"First, there are IT companies that are pro-actively pushing their big data solutions and try to convince organizations of the need to exploit their data. Not only because it might be necessary, but because it brings in money for the software company. Second, there is an abundance of media attention for big data."*

Curiosity and belief

Notwithstanding the hype and fear of missing out, there is of course an abundance of inspiring examples of creative and valuable exploitation of data. One of Teunissen remarked:

> *"There is a lot of curiosity. People start to wonder what would happen if they would correlate their data with weather patterns or Twitter data. The phenomenon of big data hints at something that holds incredible value, but at the same time it remains hidden where that value is or how to extract it."*

Some companies have been able to generate detailed insights about customers' behavior, which in turn enabled them to optimize their services or cost models. Particularly the notion that combinations of existing internal datasets with external (sometimes open) datasets may generate valuable but unexpected insights, has led some companies to actively explore the possibilities. The Vancouver Police Department, for example, made extensive use of a wide variety of data sources and big data analytics to take a more preventive and predictive approach in their fight against crime. Ryan Prox from the Vancouver PD:

> *"With an advanced mapping and analysis tool such as CRIME, data is analyzed in mere seconds, linking non-obvious connections and uncovering relationships between individuals and criminal activities across time. The result is actionable intelligence that helps police to stay a step ahead of offenders, and optimizes the deployment of policing resources based on the situation unfolding."*[41]

A few believers who start small

Whether or not available data is actually exploited also depends, to a large extent, on the work of (or support from) a handful of enthusiastic - and often visionary or at least bold - individuals within the organization. This is especially true for the initial phase of deploying structural data collection and analytics. Given the support and commitment of more than one organizational department that is required to integrate new processes into an organization, getting started in the first place is typically tedious, slow and expensive, as will be discussed in more detail in paragraph 3.5. In this respect, interviewees noted that employees that possess multiple kinds of expertise (technical know-how, domain knowledge, commercial

[41] Prox, R. (2013). How Vancouver Tapped Big Data To Fight Crime. In: *A Smarter Blog.* Available at: http://asmarterplanet.com/blog/2013/07/how-vancouver-tapped-big-data-analytics-to-fight-crime.html

competence, access to data sources and hardware) are often important for the take-up of big data technologies and a data-driven strategy. Moreover, it is noted, that chances for a successful deployment and broader implementation of big data in an organization increase when there have already been successful experiments on a smaller scale within specific departments or even a subdivision. These examples can be used as success stories or business cases to convince higher management to exploit big data in a more substantial and integrated way in the whole organization.

3.2 Goals and benefits of big data

Using the interviews and output from the workshops as primary sources, we discuss the most common goals for organizations when deploying big data– and the benefits they perceive.

Optimizing and improving organizations
Many of the applications of big data are geared towards improvements of internal organizational processes. Management may rely on data to decide on organizational focus (and possibly changes), or use data analytics as tool to make decisions on the internal allocation of financial resources. A survey by MIT showed that using data analytics for 'financial management and budgeting', 'operations and production' and 'strategy and business development' are top priorities for companies.[42] TDWI, based on their own survey, found that professionals see 'more numerous and accurate business insights' as a main benefit of big data analytics.

The interviewee from HbR, for example, stresses the importance of using data on market developments in the landside transport sector. Given their role as one of the leading ports in terms of oil and coal transshipment, HbR's aim is to collect and analyze larger, more and more relevant datasets than before to develop future strategies. In this respect, for example, also open or accessible data about larger macro-economic statistical trends and the developments in the field of shale gas helps to develop and improve HbR's business intelligence.

Both public and private organizations are bound to numerous laws, regulations or other types of rules and agreements that require specific documentation. For example, the Dutch public transport organization Dutch Railways (NS) is required to meet certain goals (% trains on time) and has to report to the government. Waste management organizations need to report to their principals, mostly local municipalities, how their costs relate to the services they have delivered. For such accountability related issues, data can serve as an important resource that enables organizations to more easily meet the demands that are faced with. For example, NS uses insights from social networks in order to yield better insights as to the extent to which trains arrive and leave on time. This information in turn helps them to better allocate resources in order to meet the demands placed upon them by the government.

[42] Lavalle, S., et al (2010) Analytics: The New Path to Value. MIT Sloan Management Review. Available at: http://cci.uncc.edu/sites/cci.uncc.edu/files/media/pdf_files/MIT-SMR-IBM-Analytics-The-New-Path-to-Value-Fall-2010.pdf

Marketing and sales
The explosion of data has led to a multitude of data sources that enhance insight into customers' behaviors and attitudes (social media have become indispensable sources) as well as market developments. It has become common good for many organizations to personally and automatically target advertisements to (online) audiences based on individual profiles that have been assembled by integrating various data sources that contain information about (groups of) individuals. Or as put the TDWI report on big data analytics[43]: "Anything involving customers could benefit from big data analytics." The survey results in this report show the importance data-management professionals assign to analytics, with better targeted social-influencer marketing, customer base segmentation and recognition of sales and market opportunities ranking among the most important benefits.

One of the most illustrative examples in terms of leveraging the power of big data is Walmart's approach, a company with "big data in its DNA", according to Mark van Rijmenam from Smart Data Collective: "The Social Genome product allows Walmart to reach customers, or friends of customers, who have mentioned something online to inform them about that exact product and include a discount. In order to do this they combine public data from the web, social data and proprietary data such as customer purchasing data and contact information. This has resulted in a vast, constantly changing, up-to-date knowledge base with hundreds of millions of entities and relationships."[44] But in order to apply big data and analytics in such an integrated fashion, support from 'the right people' is crucial. Or as one of our interviewees states:

> *"There has to be full awareness of the potential of big data in higher management levels. There needs to be a management that values information and wishes to measure and know as much as possible."*

Also HbR uses large datasets to determine their marketing strategies for targeting potential customers. For example, to support the claims and marketing strategy that HbR's location in Rotterdam is a good (or often better) port to transport goods to the southern regions in Germany (instead of the ports of Hamburg and Bremen), information about the cargo routes in the fore- and hinterland is of crucial importance. HbR:

> *"But as of yet, we have an incomplete view of all transportation routes to conduct solid analyses. Most often, we have to rely on snapshots. We would - ideally - make use of real-time and dynamic information, visualized on maps and in graphs."*

Customization, personalization and data-products
The emergence of new means of data generation, collection and analytics has made it possible to move away from one-size-fits-all solutions that have been standard procedure in product development for a long time. For example, as the waste management case shows (see paragraph 3.3.1), by collecting data about waste disposal behavior on local and even individual level, municipalities are now

[43] Russom, P. (2011). Big Data Analytics. TDWI Best Practices Report. p. 11. Available at: http://tdwi.org/research/2011/09/best-practices-report-q4-big-data-analytics.aspx
[44] Rijmenam, M. (2013). Walmart Makes Big Data Part of its DNA. In: *Smart Data Collective.* Available at: http://smartdatacollective.com/bigdatastartups/111681/walmart-makes-big-data-part-its-social-media

able to differentiate taxes for individual households. This could not only support municipalities to adapt and differentiate their campaigning strategies - depending on the type of inhabitants in certain geographical areas, a strategy that is not yet implemented - but it also allows citizens to influence the height of taxation by adjusting their waste disposal behavior. Ayasdi, an American start-up that develops analytical tools, has created its so-called Cancer Genome Atlas. It contains a topological model of tumors and based on this typology it uses analytics to identify how different population subtypes respond to particular treatments[45], which eventually could lead to personalized and more effective care. The online music service Pandora collects and analyzes data on the preferences and music listening behavior of its users to provide personalized music experiences.

Besides the ability to tailor goods and services to individual preferences and characteristics, (analyzed) data in itself is also increasingly used as a key ingredient of new products and services. An example of such a data-product is the earlier mentioned online platform from John Deere. Another illustration of how data can become a key asset in a company is Nike. Using data collected by Nike apps and the sensors in its products, it provides a multitude of information and services related to physical activities. By expanding their traditional portfolio of sportswear to more data-centric services, Nike has been able to find a convenient way to connect with their customers. Reflecting on the efforts and development of Nike in recent years, Fortune's Scott Cendrowski notes:

> *"Getting so close to its consumers' data holds exceptional promise for one of the world's greatest marketers: It means it can follow them, build an online community for them, and forge a tighter relationship with them than ever before. It's part of a bigger, broader effort to shift the bulk of Nike's marketing efforts into the digital realm -- and it marks the biggest change in Beaverton since the creation of just do it."*[46]

Also HbR has come to the conclusion that there is much interest of third parties in all the port-related data it collects and analyzes. As a result, the initiative Port Consultancy, which delivers consultancy - to a large extent based on data analytics by HbR - to other port companies, has been set-up focusing on effective and optimal port management. Commenting on the potential value of the data that HbR holds, the interviewee comments:

> *"If [HbR] was listed on the stock market and we would have had very innovative shareholders, I am sure the exploitation of the value in all that raw data would have been stimulated much more."*

Societal goals

The commercial or financial benefits of big data are obvious. Data - when processed and analyzed - simply holds valuable information that can be exploited in various ways, as has been described above. But big data also holds many promises for positive impact on a societal level. For example, data from the energy sector (collected through smart homes and smart meters) that is shared with or made available to the public and public institutions may help to generate insight in energy

[45] www.ayasdi.com
[46] Cendrowski, S. (2012). Nike's new markting mojo. In: *Fortune*. Available at : http://management.fortune.cnn.com/2012/02/13/nike-digital-marketing/

use patterns by households and companies, which in turn can be used to improve energy-efficiency and decrease overall energy consumption[47]. Big data also holds much promise for health care. As put forward in a report from the Aspen Institute:

> *"Identifying new correlations in data can improve the ways to develop drugs, administer medical treatments and design government programs."*

This was also demonstrated in the above-mentioned example of cancer treatments based on the findings of Ayasdi and similar research.

Preferably, commercial and public benefits go hand-in-hand. As one on the interviewees states: “We use data analytics to support partners in improving and optimizing their logistical operations.” So by helping their partners to gain a more detailed picture of how their customers, in different locations and in different time periods, behave, they can also make more efficient use of the company cars. Here, environmental benefits go hand in hand with operational benefits. In the waste management case the main driver for deploying sensors in container bins and card readers for underground containers and using this data to create differentiated tariffs was to reduce the overall waste disposal. The result of personalization and financial differentiation indeed has positively impacted on environmental goals: "Since the introduction of differentiated tariffs, the amount of waste has decreased 30%."

3.3 The value creation process of data

Having identified the objectives that organizations may pursue with big data, the next phase of the study is to shed light on the steps that need to be taken in order to extract information and knowledge from raw data sources. Raw data is not yet suitable to power more efficient production processes and other kinds of innovation. Data may be stored across one or several databases, but in many cases requires refinements and transformations before they can be transformed into actionable output.

In its most basic form the value creation process of data can be represented in three main steps:

- data generation
- analysis
- output

The McKinsey report[48] describes the value chain of data as a succession of these activities, although it adds a fourth step: the aggregation of data, which follows the first step of the generation of different kinds of data prior to the analysis and the consumption of data to derive value. Still, this is a very general and abstract description. It lacks a more detailed overview of the broad variety of the types of data, infrastructure, analytical techniques and the nuances in the value chain of big data that shape the big data landscape.

---

[47] Hamilton, T. (2013). Big Data Key to Big Gains in Energy. In: *The Energy Collective*. Available at: http://theenergycollective.com/tyhamilton/187006/big-data-key-unlocking-big-gains-energy-productivity

[48] Manyika, J., et al (2011). Big Data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute. Available at: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

In order to use the value creation process of data as a framework to describe how organizations can use data to innovate, and to shed light on the emerging big data ecosystem, a more elaborate and detailed model is required. Using the expert interviews, case interviews and workshops as primary input – combined with desk research*49 50 51* – we have identified the following steps:

1 Data generation and collection (e.g., inventory of data sources and its qualities, enabling access to data sources)
2 Data preparation (e.g., filtering, cleaning, verification, adding metadata)
3 Data integration (establishing a common data representation of data from multiple sources)
4 Data storage (e.g., local databases, cloud storage, hybrid solutions)
5 Data analysis (e.g., text mining, network analysis, anomaly detection)
6 Data output (e.g., visualization)
7 Data driven action (e.g., decision making, customer segmentation)
8 Data governance & security (e.g., governance, privacy)

[49] OECD (2013) Exploring Data-Driven Innovation as a New Source of Growth: Mapping the Policy Issues Raised by "Big Data", *OECD Digital Economy Papers*, No. 222, OECD publishing. Available at: http://dx.doi.org/10.1787/5k47zw3fcp43-en
[50] Tech America Foundation (2012). Demystifying Big Data: A Practical Guide To Transforming The Business Of Government. Available at: http://www.techamericafoundation.org/bigdata
[51] Kaisler. S., et al. (2013) Big Data: Issues and Challenges moving forward. In: *IEEE Computer Society*, pp. 995-1004.

The resulting value creation process of data is represented in Figure 1.

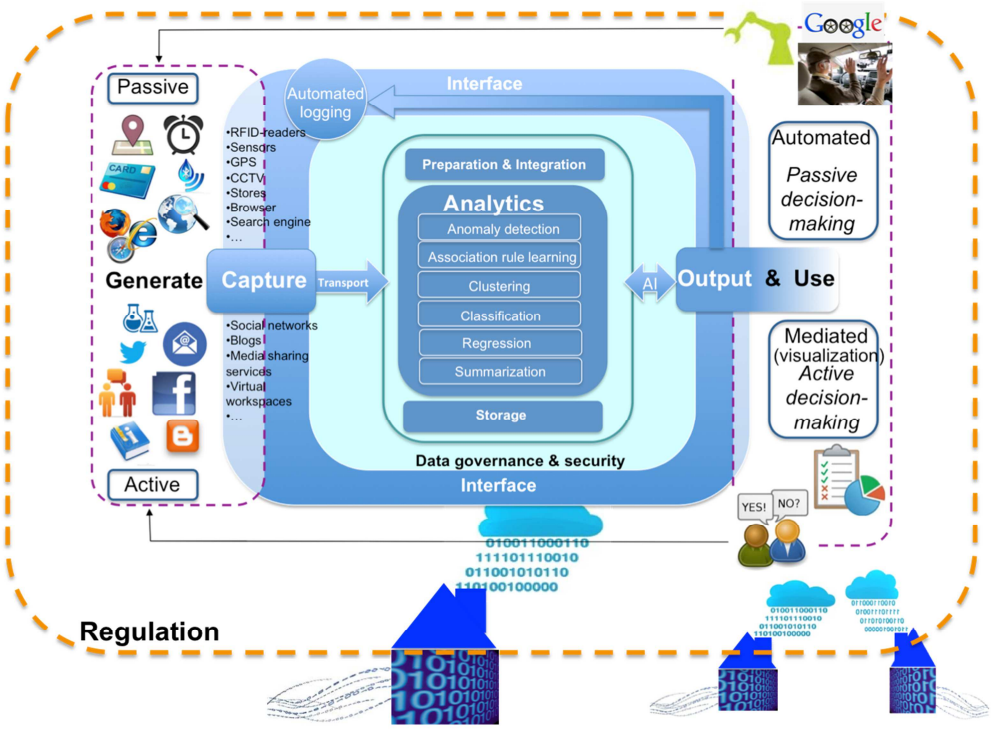

Figure 1 The value creation process of data (TNO, 2013)

As one can see, the three broad, generic steps (data generation > analytics > output) are still visible. Later in this chapter, we use the two case studies to shed light on the configuration of various steps of the value creation process. But first, we discuss three important characteristics of this model. Then, the two cases are introduced and discussed following the steps of the value creation process of data.

Intention and autonomy

First we need to discuss the concepts of intention and autonomy introduced in the above representation of the value creation process. On the left side of Figure 1 a distinction is made between data that is generated actively by humans and data that is passively generated data by sensors, cameras etc. Examples of actively generated data are, for instance, blogposts, tweets or an Excel-sheet. Examples of data that is being passively generated are browser data, GPS data, or data from CCTV cameras. In the case of passively generated data people might not even be aware of the data capturing process. And often the data might not even concern human actions but rather natural events or actions from machines.

The same distinction is made on the output-side: active and passive decision-making. The output of big data analytics is generally described as information, or 'actionable' outcomes[52], that enable humans to make better decisions and products. However, if we include the development of the Internet of Things, often the results of analytics are not directly used and interpreted by humans but by smart devices or robots. An interesting example is the autonomous vehicle that Google

[52] Howard, J., Zwemer, M. and Loukides, M. (2012). Designing Great Data Products. In: O'Reilly, *Designing Great Data Products*, p. 3. Available at: http://strata.oreilly.com/2012/03/drivetrain-approach-data-products.html

has developed.[53] The Google car does not provide its passengers with smart visualizations of the data it captures in order to enable them to decide whether or not to brake for an approaching traffic-light. The driverless car makes decisions *for* the passenger. It takes action and drives itself. Of course, the cars (or rather the set of algorithms that control it), are created by humans who have taught it how to drive, but after this initial development phase, the car takes over.

The relation between autonomous decision-making by algorithms and decision-making humans is not necessarily fixed but can be flexible and shift over time, depending on the context. As the interaction between humans and algorithms is expected to intensify in the future[54] [55] *autonomy* is an important factor that needs to be accounted for when describing the value creation process of data and the (impact of the structuring of the) big data ecosystem.

A cyclic process
A second characteristic of this model is that it represents the value creation process as a cycle, rather than a finite process from data capture to output. Once the value creation process results in data driven action, new data is generated (by humans or machines) and captured, possibly both actively and passively. Using the same example: the Google Car constantly monitors its surroundings and acts accordingly, after which the cycle begins anew. New data is generated, captured, et cetera.

Context
A third characteristic is the acknowledgement and incorporation of the context in which the value creation process occurs. In other words: data driven innovation entails a concrete fulfillment of the value creation process of data for a specific goal within an organization and a broader societal arena. For instance, a company wants to adapt its marketing efforts to reach a new target group. Subsequently, this requires an inventory of the current internal processes, the required data and an assessment of technological and organizational aspects as well as the regulatory context. The company has to determine which departments within the organization require what kind of output, what datasets are necessary (e.g., social media data from a specific target group or comments on specific topics), the analytical tools that use the right algorithm, but also infrastructural issues such as data transport, storage, computing, and so on. In many of the steps of the value creation process of data the company might need to partner with other parties – for instance social media companies such as Twitter for data, and IT vendors for the storage and analytical infrastructure or an additional partner for a specific visualization technique. In fact, these kinds of partnerships are driving the formation of the big data ecosystem, which will be described in paragraph 3.4. New practices and processes within an organization – possibly due to new partnerships – might also require organizational and cultural changes. Furthermore, if a company wants to use personal data it also has to take regulation regarding data protection into account.

[53] http://en.wikipedia.org/wiki/Google_driverless_car
[54] Brynjolfsson, E. and Mcaffee, A. (2011). Race against the machines. Lexintgon: Digital Frontier Press.
[55] Steiner, C. (2012). Automate This: How Algorithms Came to Rule the World. London: Penguin Books.

In the next two sections, we will describe the value creation process for two specific organizations in the logistic industry: CIBAS (waste management) and the Port of Rotterdam Authority.

3.3.1 Case study: CIBAS

CIBAS, the Central Information Management Waste Services, is a joint initiative from three Dutch waste management organizations: Circulus, Berkel Milieu and Twente Milieu in the East of the Netherlands, and placed under supervision of Circulus. The main objective of CIBAS is to support local governments in tax differentiation based on individual household waste disposal. To achieve this goal CIBAS manages and analyzes the data that is being collected in the fulfillment of waste management.

3.3.1.1 Activities

The main driver behind CIBAS is the introduction of Diftar: differentiated tariffs for waste management services. Diftar is one of the main instruments used by several Dutch local governments to reduce waste disposal and stimulate recycling of valuable resources. The main output of CIBAS' activities are analyses that are used as input for the software from the local government tax offices to determine the differentiated waste taxes for their citizens. Depending on the requirements from governments, these analyses are offered on a yearly or quarterly basis. CIBAS developed its own tools to export its analyses that can be processed by the software from the local government tax offices.

Besides Diftar, the activities of CIBAS support several additional objectives from the associated waste management services. First, analyses of citizens' waste disposal behavior combined with operational data enables the waste management organizations to optimize their own operations, such as the use of supplies, vehicles and manpower. CIBAS, as part of Circulus, also performs these tasks for Circulus. Second, waste management organizations can use the data to justify their operations to their contractors (the local governments).

3.3.1.2 Background

Although the city of Deventer had enhanced its waste containers with chips many years before CIBAS started and the data on waste disposal by citizens was already available to Circulus, it was not yet properly used. Within Circulus there was no single department responsible for leveraging this data, it was always an extra task that was not directly aligned with the primary processes of the various departments. Although the local government had provided the chips in the waste collectors, they did not push Circulus to actually collect and use the data.

This changed when other local governments introduced differentiated waste taxes as well. Circulus realized this posed several of the same kinds of challenges waste management organizations in other areas that it had to face itself. Consequently, they explored whether the data management and related issues could be addressed in collaboration with the other companies to make sure that knowledge and expertise was captured and nurtured in a shared service center, and costs were limited due to efficient use of resources. As a result, CIBAS was set up and placed under supervision of Circulus.

3.3.1.3 The value creation process at CIBAS

Data collection
CIBAS does not collect all the data about waste offer by itself. Instead, it uses the data that is being collected by each associated waste management organization and the local government.

These data are:
- Daily collected, via garbage containers that are tagged with a unique chip that is recognized by the garbage truck when it is emptied.
- Daily collected, via underground waste storage facilities where users need to identify themselves with a special ID-card before they can dispose their garbage.
- Daily collected, via 'mutations' in the field, for instance when citizens have one of their garbage containers replaced, or when new ones are installed.
- Daily collected, via sensors when transport is weighed at recycling centers.
- Periodically collected, via analysis of waste samples to detect patterns regarding the content of waste disposal and the various 'resources' that are being disposed.

CIBAS uses identifiers that are provided by the local tax office. These identifiers enable them to connect the data on waste disposal to individual households and, consequently, determine the individual tax rates. Currently, CIBAS does not use data from other parties, although it has experimented with external data sources, which will be discussed in more detail below (see 'ambitions'). CIBAS is also responsible for the overall integrity and quality of the data in the various waste management processes.

Preparation
The data from the associated partners is collected by the different IT companies that provide the garbage containers and underground waste storage facilities with sensors and card readers. One of the main tasks of CIBAS is to make sure that it integrates all the required data to conduct their analysis. If the data is not properly collected, which compromises the overall data quality and integrity - it has to coordinate with the executive waste management organizations that should provide this data. CIBAS sought collaboration with multiple industry organizations. This resulted in the development and adoption of an open standard to facilitate the interoperability, a process that was guided by Dutch knowledge organization TNO.

Storage
CIBAS uses the data from these organizations for the overall analyses, but does not centrally store all this data on its own server. Instead, each individual associated waste management organization stores its own data, although some use external servers. CIBAS uses an interface to perform analyses locally near the data from the other organizations which remains their property.

Analysis
For the analysis of the data CIBAS uses Microsoft Excel and Crystal Reports from SAP. Recently, CIBAS explored other options, but decided to stick with Crystal Reports. The reasons for extending the license for this software were: the ease of use to acquire quick insights, the use of this software by the associated

organizations, and the learning curve for implementing an alternative software product.

CIBAS also uses the data from Circulus to optimize their waste management operations. This analysis is partly performed manually. If there are anomalies in the data, the CIBAS-employees look at the particular days or garbage trucks in more detail to see what could be the reason for the detected deviations from planned waste disposal patterns (holidays, broken sensor readers, etc.).

3.3.1.4 Moving forward: ambitions

CIBAS has experimented with additional, external data sources such as data from CBS, the Dutch Bureau for Statistics. The goal of the experiment was to explore whether combinations of socio-economic data and data on waste disposal in different neighbourhoods, could provide valuable insights. These insights could for instance be used by the local governments to adjust and optimize their policies and communication strategies. Although these experiments did provide some interesting results, they have not yet been implemented on a structural basis.

Another experiment that provided interesting insights was an analysis of the rate of litter dispersion in different locations in the city. If these experiments were to be conducted periodically it would be possible to detect patterns that could provide insights regarding the parts of a city where litter appears faster than others, which can be used for resource management.

Another interesting data source would be consumption patterns, to better anticipate the expected waste and the different kinds of recyclable sources (paper, plastic, bio, textile, etc.).

Visualization

Just as smart meters enable consumers to better understand their energy habits and consumption patterns via visualizations, the collection of data on waste disposal could enable citizens the same kinds of services. However, the sector is not yet ready for these kinds of products. Deventer has just started to provide citizens with an online overview of their invoices.

Expansion in regions

Several waste management organizations will join the CIBAS initiative: *Avri* from Tiel and Geldermalsen and *Area* in Hoogeveen. These organizations have considered the possibilities to build the required knowledge and infrastructure themselves, but decided to join CIBAS and profit from their available know-how and experience with this new data-driven way of working.

The expansion to other areas is not primarily driven by commercial motifs. The new organizations only pay a compensation for the additional costs for CIBAS, rather than a commercial tariff. The main driver for CIBAS to expand is to improve collaboration, acquiring and sharing knowledge and negotiating better terms with local governments for waste management services, which is easier if more companies join. Therefore, CIBAS does not aspire to expand in all parts of the Netherlands, but rather in its own region.

This kind of collaboration is unique in the Netherlands. There are other companies that provide more or less the same kinds of services, but they use commercial rates (such as Atero, former Essent Mileu) and are less interested in collaboration. In other examples of collaboration the local government is the contractor, rather than the waste management organizations themselves.

Expansion in tasks
In the (near) future other tasks that concern the maintenance of public spaces, parks, street lighting etc., might be part of CIBAS as well, and could be addressed with the same kind of data-driven method. In many cases, the data is already available, but also a new way of working is required. In the end, a lot of these tasks (waste, litter, urban plantations, parks, etc.) are related and a better use of data could improve the overall cleanliness of a city. This expansion would result in a much more complex ecosystem of data-owners, collectors and IT-experts.

3.3.1.5 Moving forward: barriers

Focus and accountability
Initially the lack of focus and lack of accountability resulted in a culture within Circulus in which the data was not properly used.

Standardization
When CIBAS was initiated as a joint effort of several waste management organizations that each worked with different IT-providers and local tax offices (each using different kinds of software), the lack of standardization was a major issue. The implementation of the so called STOSAG standards, the collection of open standards for waste management, which was initiated in 2010 with support from the national government, has addressed this issue. But standardization is not only a technical issue. It also concerns a specific way of working that requires coordination between the various organizations to make CIBAS work.

Data integrity and quality
In order to deliver the analyses for the local tax offices, CIBAS has to make sure that it collects all the data. It has to guard the quality and integrity of the data to avoid the risk of ‘garbage in, garbage out’. However, the data is delivered by the individual waste management organizations and, more precisely, by the employees who actually perform the waste collection, whose primary concern in not the integrity and quality of the data, but rather (the speed of) waste collection and customer service.

Culture
The overall focus in waste management in the associated cities and waste management organizations has shifted. Before the introduction of *Diftar* there was a ‘consumer’-focused approach: waste was collected as quickly as possible and modifications in the field (e.g., applications for new collectors) were processed with the least amount of waiting time as possible. However, due to the introduction of differentiated taxes, the focus has shifted. Now, the primary goal should be to make sure that the taxes are based on the right data and reflect the actual waste disposal of individual households and, subsequently, to influence their waste disposal behaviour. This shift requires a different way of working. Although the overall goal for the different actors is the same (waste collection and reduction),

there is still a difference in attitudes, especially as the employees who collect the waste containers, and subsequently the data, are more focused on the 'customer', do not directly experience the benefits from the data collection.

Reorganizations
The ambitions regarding the use of additional, external data sources are momentarily on-hold due to reorganizations in Circulus.

3.3.2 Case study: the port of Rotterdam Authority

The Port of Rotterdam Authority (HbR) - privatized in 2004 but still owned for 70% by the city of Rotterdam - manages, maintains and develops the Rotterdam harbor: it leases property to companies, it invests in both nautical and landside infrastructure to expand and improve accessibility, and it promotes the port of Rotterdam for business development. Since 2008, the WA-department – which is part of the commercial division within HbR – is responsible for business analysis and intelligence.

3.3.2.1 Activities

The main goal of the collection and analysis of data at HbR is to improve its overall business intelligence. This means a better understanding of relevant market dynamics and trends that could impact (and guide) its strategy regarding both mid-term and long-term development and investments to better cater to its customers and improve business. Currently the data is analyzed and processed to create reports that are used, in addition to external market reports, to make decisions regarding development and investments. The WA-department encourages the HbR business developers to use and engage with the available data to perform their own queries.

There are other goals – which might become more relevant in the future as the data-related activities develop – such as acquiring more data and intelligence, which can be used to improve HbR's marketing activities and consultancy to other ports. As will be discussed below ('ambitions') more real-time data on logistical flows from competing ports such as Hamburg and Bremen, in combination with logistical landside data from trains and trucks, could be used for these kinds of marketing purposes. Additionally, more detailed information on cargo et cetera could support ships and landside logistical partners with their logistical planning. With these kinds of data, HbR aspires to act as an overall logistical planner for both sea and landside flows. Such integrated logistical services, that combine and streamline both nautical and landside flows, become increasingly important to stay competitive.

3.3.2.2 Background

The WA-department started in 2008 as part of the commercial division of HbR to provide them with strategic business intelligence for mid- and long-term development and investments for the port of Rotterdam. The WA-department started with limited funds, which resulted in sub-optimal tools to perform their tasks (see 'barriers'). However, there is now more awareness that data and data analytics can be valuable. The willingness to spend money on more advanced tools is growing.

3.3.2.3 The value creation process at the Port of Rotterdam

Data collection
HbR collects (but does not own) many different kinds of data from different kinds of ships, containers and goods. These data are collected via a 'statement of harbor dues', which is declared – manually – via a web interface by the shipbrokers.

External data sources are CBS (the Dutch Central Bureau for Statistics) and Destatis (the German Central Bureau for Statistics). HbR subscribes to a service from Seabury, which has global data on trade and logistical flows between different countries and continents. In addition to these datasets, HbR works with Panteia (former Neia) that provides analyses on landside processes.

Furthermore, HbR collaborates with the Technical University of Delft to create a Wiki (hosted at the Technical University of Delft) to collect and store information on competing ports as well as their landside information, data on traffic nodes and terminals. This Wiki will contain both hard data (reports), and 'soft data' (e.g. conversations).

Storage
HbR stores all its data on local servers.

Preparation
The data from the shipbrokers that is being used by the WA-department is collected by the HbR department that handles the harbor dues. This data is tailored to their specific billing and processing requirements. For the billing process the exact contents of cargo are not important, only the overall value is relevant. Consequently, when the tariffs for different kinds of goods of a single cargo are the same, this leads to many errors: instead of describing all the different kinds of goods in detail, shipbrokers often put the goods with the same tariff together under a single denominator. This means that the contents declared on the statements are not the same as the actual contents on the ship. While these details might not be important for the billing process, but they are very important for the analysis of the WA-department.

Analysis
The WA-department of HbR uses various software tools for its data analysis: a lot of the time they use Excel, although the original version of Excel is no longer sufficient to handle the increased amount of data. In addition to Excel, they use the tool 'Business Objects' from SAP and the prognosis tool 'Business, Planning and Control'.

3.3.2.4 Moving forward: ambitions
HbR wants to improve decision-making for development and investments, the connection to landside logistics and its marketing. Therefore, it needs to collect and analyze all kinds of new data.

HbR would like to collect the so-called 'manifest-information' from ships, which shipping companies are obliged to share with customs and CBS. The manifest information contains much more detailed data (contents of cargo, previous and planned destinations, related shipping companies and freight rates, etc.).

One of the ambitions of HbR is to aggregate the ‘manifest-information’ and the real-time data location data from ships to improve both the infrastructure of the port of Rotterdam and the connection to landside logistical partners. It may also enable them to expand their expertise beyond the port itself and take up a role as an overall planner for different kinds of logistical flows (e.g., sea, rivers, trucks, trains). This kind of streamlining can benefit the shipping companies as well.

The ambitions of HbR require the company to collaborate closely with Portbase, who is responsible for the collection of the ‘manifest-information’ for customs and CBS, but also with the Port of Amsterdam and landside logistical partners. By collecting and analyzing these different kinds of logistical data, HbR can convince (potential) clients that its harbor is, for instance, the most efficient route to the South of Germany, rather than the port of Hamburg or Bremen. However, in order to make such statements, it needs more information on the logistical flows in those ports and their landside infrastructure as well.

3.3.2.5 Moving forward: barriers

Data Quality
As mentioned above (see ‘preparation’) the quality of the data that is collected via the harbor dues contains a lot of mistakes. These mistakes require manual, tedious workarounds.

Access to data
The WA-department has ambitions to use more and real-time datasets. In some cases they knows where to find these datasets, but do not have access to them. In other cases it does not even know where it may find the data it wants to use.

Privacy
The tracking system IES generates location data that could enable HbR to track and trace ships for logistical planning purposes. However, this can be very privacy-sensitive as many people actually live on the ships, especially in inland navigation. Using the data from the IES system means that HbR would not only track ships and the contents of their cargo, but also the people who live on them. HbR has its own legal department, but for specific privacy or data-ownership issues it consults external legal experts.

Sensitive data
Shipping companies are reluctant to provide HbR with their more detailed information, which is competition sensitive. The shipping companies are afraid that their information will end up in the hands of their competitors as well.

Trust and collaboration
So far it proved to be difficult to convince the shipping companies of the added value of the services that HbR and Portbase want to develop with their data. As mentioned above, they are afraid that sensitive information will be leaked or shared with competitors, even though Portbase, as a neutral broker, will aggregate and anonymize the data.

The shipping companies also feel that they run quite some risk when they open up their sensitive data to external parties. The question what the companies can expect in return (reports, financial compensation, a better functioning port) has not yet been explicitly on the table. As a first step, they will have to be convinced of the overall benefits when HbR is better able to improve the infrastructure of the Rotterdam port, and can offer better planning services and alignment with landside logistics.

Development of software tools
Some of the software tools that the WA-department uses for its analyses are developed by third parties. The development process of these tools proved to be very difficult and time-consuming. The lack of domain specific knowledge ((port) logistics) was a major bottleneck. HbR is still not really happy with the tools. However, as HbR does not have the knowledge and expertise to build its own software, they have to rely on others.

Finance
The current state of the IT-systems are also the result of the financial restraint from HbR. It takes time before management realizes that investments are necessary and profitable in the long-term.

Negotiations with shipping companies regarding their more detailed shipping information have been postponed for now. These companies initially demanded reduced harbour dues in return.

3.4 The big data ecosystem

Data-driven innovation comes down to the concrete fulfilment of the value creation process of data in order to reach a specific goal; to tackle a problem to which data analytics could provide (part of) the solution.

For each specific goal, an organization (or a consortia of organizations) needs to configure a value creation process. It is likely that for many of these steps, organizations will have to involve third parties, because they lack experience, technological resources and/or talent to deal with the multidisciplinary aspects of big data on their own. The resulting network of actors is in most cases specifically tailored towards the goal that is being pursued.

The aggregate collection of big data players, services and products, technologies and regulations that shape the data-driven constellations create a transcending big data ecosystem.

3.4.1 Players
Describing the big data ecosystem is like taking a snapshot of a moving target. It is a very dynamic field, as new technologies and practices are constantly being developed. The developments are largely driven by traditional IT-players in infrastructure and business analytics (e.g., IBM, Oracle and Microsoft), and a vast array of start-ups.

There are various depictions of the big data ecosystem that reflect the main outlines of the value creation process of data to structure the different types of players. Two

of such representations can be found in Figure 2 from Matt Turck[56] and Figure 3 from Forbes' Dave Feinleib[57].

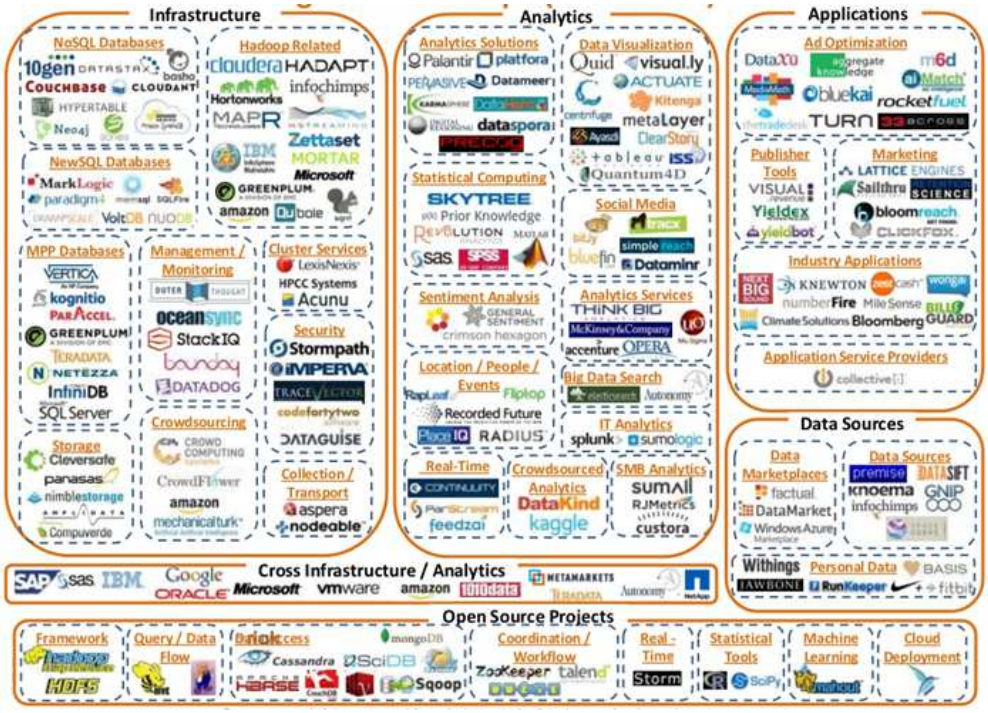

Figure 2 The big data landscape (Matt Tuck, 2012)

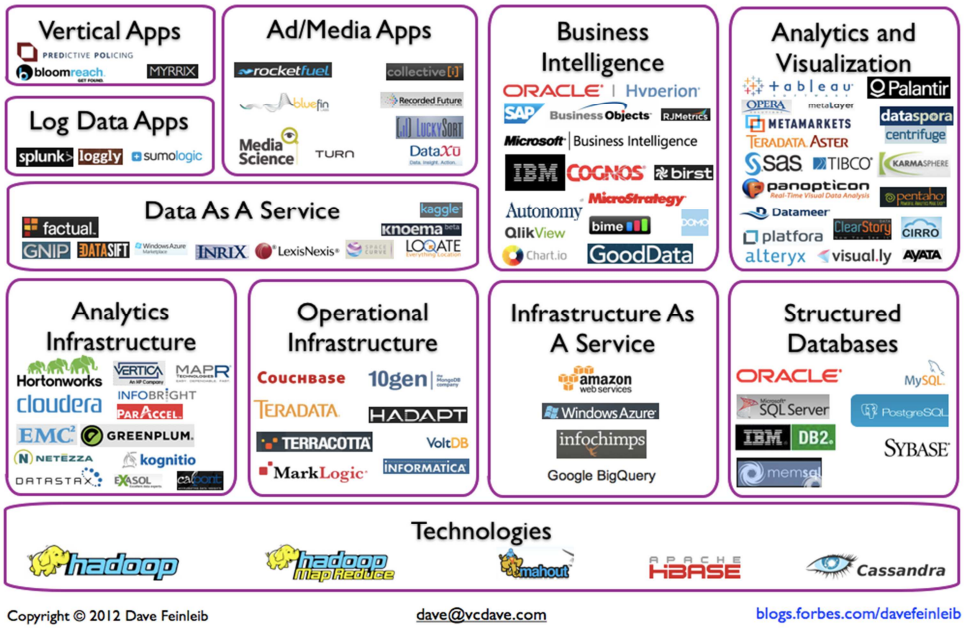

Figure 3 The big data landscape (Dave Feinleib, 2012)

[56] Turck, M. (2012). A chart of the big data ecosystem, take 2. Available at: http://mattturck.com/2012/10/15/a-chart-of-the-big-data-ecosystem-take-2/
[57] Feinleib, D. (2012). The Big Data Landscape. In: *Forbes*. Available at: http://www.forbes.com/sites/davefeinleib/2012/06/19/the-big-data-landscape/

Both figures discern the same basic elements, which are clusters of more detailed typologies of services, products and technologies:

- Underlying core technologies, such as Hadoop
- Technological infrastructure, e.g., storage and computing
- Analytical techniques and tools, e.g., sentiment analysis
- Applications that are tailored to a specific domain
- Data sources

In addition to the above mentioned main elements of the big data ecosystem, or rather in a more detailed assessment, these two descriptions each highlight different aspects, such as data transport, types of storage solutions and database technologies, data integration, data management and monitoring, security, consultancy services, and different kinds of analytical techniques.

They provide an extensive and useful overview of the most relevant big data players. However, there are also some important gaps that need to be addressed in order to properly describe the big data ecosystem and how it is being structured.

First, these two representations of the big data landscape are strongly technology-oriented. They describe big data technologies, the various types of big data products and services, and the companies that provide them. However, the expert interviews, case studies and workshop output strongly suggest that data-driven innovation is not only a technological challenge. As our assessment of the value creation processes of data and existing barriers shows, most big data innovations also consist of organizational challenges: working processes, attitudes, change management and HR policy. However, the services or products that support these organizational challenges are not mentioned in either of the two big data landscapes. Also, the descriptions do not include legal consultants. Especially for organizations that deal with personal data, legal expertise is very important.

Second, due to their focus on providers of big data technologies, products and services, the landscapes above paint a one-sided picture. Stakeholders that are not primarily concerned with providing big data products and services are not included: the organizations that consume the big data products and services, regulators and policymakers, and individual end-users and citizens. These players generally have a different position in the big data ecosystem than technology vendors and service providers – although not in principal as will be discussed below – but they are important actors that create the demand and preconditions that shape the big data ecosystem.

By solely focusing on the technological aspects, players and services, the descriptions of the big data landscape above do not account for the interactions and relationships of a substantial share of other actors and relationships within the big data ecosystem. A more comprehensive approach warrants a more reliable view on possible networks of organizations, technologies and products that solve big data problems or seize new opportunities, and also important regulations (such as the EU data protection framework) and policies that could stimulate or hamper data driven innovation and the development of the big data ecosystem as a whole.

Below four types of players will be described in more detail: 1) providers of data infrastructure, 2) providers of analytics, applications and visualizations, 3) data sources and 4) policymakers.

3.4.1.1 Providers of data infrastructure

The market for big data infrastructure contains providers of both databases and related technologies and services (management, security, transport, storage), and platforms based on the Hadoop distribution system, which has almost become the standard technology to deal with more complex, unstructured big data problems.[58]

Until a few years ago, the nascent Hadoop space was dominated by a few products and their providers, such as the open-source Apache Hadoop distribution, the independent Hadoop distribution provider Cloudera and Amazon's Elastic Map Reduce.[59] But this space has rapidly become densely populated. According to GigaOm's Derrick Harris, the infrastructure market is already near its point of saturation:

> *"The market for horizontally focused products is filling up fast with both startups and large vendors […] Yes, there's still room for startups to get in here, but the door looks to be closing fast. It's not just Hadoop, either; other techniques, from traditional data warehouses to, arguably, predictive analytics, all are nearing the saturation point in terms of vendors selling the core technologies."*[60]

On the one hand new independent Hadoop distribution players emerged, such as Hadapt, HortonWorks (a Yahoo spin-off) and MapR. On the other hand, traditional infrastructure vendors that offered servers, storage and database technologies, moved into this space as well. IBM, EMC, Cisco, Oracle, HP and VMware all have adopted the Hadoop distribution technology in order to tackle big data challenges for their customers – or they partner with the independent Hadoop distribution providers. They align their Hadoop products with the rest of their database and analytical offerings for business intelligence.[61] As cloud-based services increasingly become viable alternatives for (parts of the) infrastructure, players such as Amazon, Google and Microsoft – in addition to the likes of IBM, EMC and HP – are important players as well.

Although Hadoop has proved to be very popular – especially for big, unstructured data challenges – classical (relational, column based) database technologies are still important, as are next generation massive parallel processing database technologies and their related analytical tools. Companies that provide these analytical platforms that combine databases and analytical tools are Vertica (owned by HP) Asterdata (owned by Terradata), SAP (with Hana), ParAccel, Attivo and Datastax to name a few. However, these products are often used in combination with Hadoop. Principal analyst Edd Dumbill from O'Reilly notes: "Practical big data implementations don't in general fall neatly into either structured or unstructured

---

[58] Dumbill, E. (2012). Microsoft's Plan for Big Data. In: O'Reilly, *Planning for Big Data*, p. 35. Available at: http://oreilly.com/data/radarreports/planning-for-big-data.csp

[59] Harris, D. (2011). As Big Data takes off, Hadoop wars begin. In: *GigaOm*. Available at: http://gigaom.com/2011/03/25/as-big-data-takes-off-the-hadoop-wars-begin/

[60] Harris, D. (2011). Why Big Data Startups Should Take a Narrow View. In: *GigaOm*. Available at: http://gigaom.com/2011/03/28/why-big-data-startups-should-take-a-narrow-view/

[61] Dumbill, E. (2012). Big Data Market Survey. In: O'Reilly Rader Team. *Planning for big data*, pp. 23-34. Available at: http://oreilly.com/data/radarreports/planning-for-big-data.csp

data categories. You will invariably find Hadoop working as part of a system with a relational or MPP database."[62]

3.4.1.2 Providers of analytics, applications and visualizations

Whereas the field of big data infrastructure seems to leave limited room for new entrants, as it is dominated by traditional vendors and a few independent Hadoop distribution providers, there seems to be an explosion of start-ups that focus on data analytics, applications and visualizations. As Gigaom's Ed Harris notes:

> *"The great thing about big data is that there's still plenty of room for new blood, especially for companies that want to leave infrastructure in the rearview mirror."*[63]

The products and services of these start-ups and smaller companies sit on top of the more foundational layers of database technologies and Hadoop solutions. Cloudera's CEO Mike Olson[64], discussing the future of Cloudera and its products, noted that he sees large potential for specialized companies in this layered construction:

> *"I believe there's an enormous opportunity for smart companies, and even open-source projects, to build a new generation of data analysis tools on top of that platform."*

These new companies focus on specific analytical or visualization techniques, target specific industries or even specialized tasks within an industry. According to Teunissen and Kuijpers from SURFnet there are pragmatic reasons for this that are inherent to the nature of start-ups. The extent to which a company is able to expand horizontally or vertically is related to its financial resources. Generally these are limited for starting companies. A more narrowed focus requires fewer investments. Similarly, Tromp from SAP notes:

> *"The market of BI and big data software is dominated by larger player because of two reasons: (1) the investments to develop solid software to meet all the complex customer demands are simply too high, and (2) smaller software providers are a risk for customers because they lack extensive track records and their future is less secure."*

However, because of their focused approach, smaller companies can offer value and ease of use that generic tools and techniques lack. All experts emphasized during the interviews the limitations of generic off-the-shelf tools. According to Clive Longbottom, founder of analyst house Quocirca, many IT-suppliers have a tendency to sell one-size-fits all offerings, whereas these new start-ups try to cater to very specific data-needs.[65] As independent data-consultant Paul Miller notes:

---

[62] Dumbill, E. (2012). Big data market survey: Hadoop solutions. In: *O'Reilly Strata*. Available at: http://strata.oreilly.com/2012/01/big-data-ecosystem.html

[63] Harris, D. (2012). Five low-profile start-ups that could change the face of big data. In: *GigaOm*. Available at: http://gigaom.com/2012/01/28/5-low-profile-startups-that-could-change-the-face-of-big-data/

[64] Harris, D (2011). As Big Data Takes of, the Hadoop wars begin. In: *GigaOm*. Available at: http://gigaom.com/2011/03/25/as-big-data-takes-off-the-hadoop-wars-begin/

[65] Heath, N. (2012). Slow start for big data in Europe. In: *TechRepublic*. Available at: http://www.techrepublic.com/blog/european-technology/slow-start-for-big-data-in-europe/

> *"At the 'softer' end of the market, specifically, there has been an explosion of new startups rushing to offer tools that make it easier to create visualizations and dashboards to deliver some value from the data whilst hiding its complexity."*[66]

This trend contradicts, according to Miller, another trend in business software, which is consolidation:

> *"[…] with fewer and fewer software tools trying to do more and more. These bloated behemoths become ever-harder to use, ever-harder to support, and developers find themselves increasingly constrained by dependencies and legacy as they try to innovate or encourage the next upgrade cycle."*

3.4.1.3 Data sources

The stacks of technological infrastructure, analytics platforms, applications and visualization tools are all tailored to process data, transforming it into valuable information and insights or otherwise actionable output. Which brings us to the data sources that provide the "new oil".

Most organizations initially apply analytics to their own internal datasets, possibly combining several databases from various departments and processes. But the value of big data also lies in the combination of both internal and external data.[67] A white paper of the European Technology Platform NESSI stresses the importance of improving the awareness of integrating existing private data with external data to enhance existing products and services.[68] As O'Reilly's Ed Dumbill notes:

> *"Mixing external data, such as geographical or social, with your own, can generate revealing insights. […] Your own data can become that much more potent when mixed with other datasets."*[69]

Noting that "critical information often resides outside companies", Biesdorf, Court and Willmott from McKinsey discuss the impact of the integration of external data sources:

> *"Making this information a useful and long-lived asset will often require a large investment in new data capabilities. Plans may highlight a need for the massive reorganization of data architectures over time: sifting through tangled repositories (separating transactions from analytical reports), creating unambiguous golden-source data, and implementing data-governance standards that systematically maintain accuracy."*[70]

In addition to using data from external sources to create value, it could also be valuable to open up proprietary datasets to others. As Rufus Pollock stated at the

---

[66] Miller, P. (2013) Visualization, the key that unlocks data's value? Available at: http://cloudofdata.com/2013/04/visualisation-the-key-that-unlocks-datas-value/
[67] Redman, T. (2008). Data driven. Boston: Harvard Business School Publishing. p. 25
[68] NESSI (2012). Big Data: A New World of Opportunities. Available at: http://www.nessi-europe.com/Files/Private/NESSI_WhitePaper_BigData.pdf
[69] Dumbill, E. (2012). Microsoft's Plan for Big Data. In: O'Reilly, *Planning for Big Data*, p. 38. Available at: http://oreilly.com/data/radarreports/planning-for-big-data.csp
[70] Biesdorf, S., Court, D. and Wilmott, P. (2013). Big data: What's your plan? In: *McKinsey&Company Insights & Publications*. Available at: http://www.mckinsey.com/insights/business_technology/big_data_whats_your_plan

OECD's Technology Foresight Forum in April 2012: "The best thing to do with your data will be thought of by someone else", affirming one of the drivers of the Open Data movement. Some organizations offer their data for free via their website or specific online portals, especially governments and NGO's. The European Commission recently launched its Open Data Portal and on his first day in office President Obama of the United States announced his strategy for 'open government'.[71]

Other organizations sell their data. Well-known examples are social networks such Facebook and Google, whose vast collections of personal data are valuable sources for advertisers. In some cases social media companies work with third parties that commercially exploit the social data. Examples are Gnip and Datasift. These companies have access to the so-called Twitter Firehose and other social media data, which they prepare and manage to make it more accessible and useful to their customers by adding all kinds of filters that fit their specific needs.

In addition to these data sources, data is also provided through online marketplaces that host data from publishers and offer it to interested parties.[72] The most established data markets are Infochimp, Datamarket, Factual and Microsoft's Azure, although there are several more.[73] Some data marketplaces try to offer all the data they can, such as Infochimp. Others focus on specific kinds of data, such as Factual, which originally started with location data and is now branching out to a few new specific verticals. Another way of specialization is to choose a specific target group, such as Figshare, which is a data platform for researchers.

According to Dumbill, these kinds of data marketplaces are useful in three ways[74]:

1 They provide a point of discoverability and comparison for data, along with indicators of quality and scope.
2 They handle the cleaning and formatting of the data, so it is ready for use.
3 They provide an economic model for broad access to data that would otherwise prove difficult to either publish or consume.

These three propositions illustrate how data marketplaces can facilitate the process of finding the right kind of data and fulfill even some additional steps in the value creation process, such as data preparation, to ease further data integration. Some marketplaces allow their customers to explore data; to mix them together with their own or other available datasets to create new value. Although most marketplaces are focused on developers as their main users, as Dumbill notes, some data marketplaces try to target less IT-savy users as well. Microsoft's Azure, for instance, has aligned its datasets not only with its other big data products, but also with its business tools such as Excel. This makes it easier for smaller organizations (even individual users) to download and combine different (internal and external) datasets.

[71] Veenstra, van A. en Broek, van den T. (2012). Opening Moves – Drivers, Enablers and Barriers in Open Data in a semi-public Organization. In: M.A. Wimmer, M. Janssen, and H.J. Scholl (Eds.): *EGOV 2013*, LNCS 8074, pp. 50–61, 2013.
[72] Dumbill, E. (2012) Data marketplaces. In: O'Reilly, Planning for Big Data, pp. 47-52. Available at: http://oreilly.com/data/radarreports/planning-for-big-data.csp
[73] Big Data startups. Available at: http://www.bigdata-startups.com/public-data/
[74] Dumbill, E. (2012) Data marketplaces. In: O'Reilly, Planning for Big Data, pp. 47-52. Available at: http://oreilly.com/data/radarreports/planning-for-big-data.csp

Furthermore, as Dumbill notes, data marketplaces enable a new economic model for data use and sharing. They do not only offer open data, they also allow publishers to sell or trade their data; they match supply and demand. As Factual's CEO Elbaz explained at the Strata 2011 conference:

> *"Another dimension that is relevant to Factual's current model: data as a currency. Some of our most interesting partnerships are based on an open exchange of information. Partners access our data and also contribute back streams of edits and other bulk data into our ecosystem. We highly value the contributions our partners make."*[75]

Even individual end-users can become data providers. A number of start-ups, such as Personal, are currently offering so called "data lockers" where people can gather, store and manage their personal data. These services allow people to take control over their personal data and "re-use it to their own benefit" [76]. Another interesting development could be the rise of personal data marketplaces. Start-ups like Handshake and Enliken offer platforms where users can sell their personal data to interested parties.[77]

The potential of re-using data for new, unintended purposes has opened op the possibility for organizations (and even individuals) that are not primarily in the business of big data, to be part of the big data ecosystem as a provider. In this case they are not solely consumers of big data products; they also contribute data that can be valuable for other organizations for very different purposes.

Governments, for instance, are important data providers. Although they are often consumers of big data (services), for instance to monitor specific policy goals, they can also act as data providers in other value networks where their data is used, and possibly combined and transformed, to create new services. The Dutch energy network service provider Alliander recently organized a workshop with partners and stakeholders to explore the possibilities of opening up their data. Nike has developed an API to enable trusted partners to develop new services on top of their Nike+ platform, which collects exercise and health data from millions of users via the Nike+ app and its wearable technology FuelBand. As the data locker products illustrate, individuals can also contribute to the big data ecosystem. In addition to these lockers, many alternative and more open initiatives of user participation abide as well. The social traffic app Waze – acquired by Google in June 2013 - collects and aggregates data generated by its users to create real-time traffic information. In the Netherlands over 10,000 iPhone owners joined the collaborative research project iSPEX to measure aerosols via their mobile phones and an additional sensor.[78]

---

[75] Watters, A. (2011). An iTunes model for data. In: *O'Reilly Strata*. Available at: http://strata.oreilly.com/2011/04/itunes-for-data.html

[76] The Economist (2012). Know Thyself. In: *The Economist*. Available at: http://www.economist.com/news/business/21568438-data-lockers-promise-help-people-profit-their-personal-information-know-thyself

[77] Lomas, N. (2013). Handshake Is A Personal Data Marketplace Where Users Get Paid To Sell Their Own Data. In: *Techcrunch*. Available at: http://techcrunch.com/2013/09/02/handshake/

[78] http://ispex.nl

However, the re-use of personal data for unintended purposes is limited due to restrictions as laid out in data protection regulations. This does not only concern the re-use and combination of datasets. Even storing personal data for a long period on internal databases for its original intended purposes may be difficult.

3.4.1.4 Policy makers

Policymakers are important players in the big data ecosystem. The Open Data initiatives of the European Commission and the White House illustrate how they contribute to innovation in the field of big data. An additional driver in Europe is the Commission's ambition to create a Digital Single Market, which also includes data and a big data ecosystem.

Furthermore, laws and regulations provide important principles by which organizations must abide. Major reforms of the EU legal framework on the protection of personal data will have a profound impact on the workings of the big data ecosystem. Other regulations – which may differ between different countries, can have a similar impact.[79]

3.4.2 Market dynamics: competition, collaboration and integration

The next session will discuss market dynamics that structure the big data ecosystem. It will describe the relation between competition and collaboration in the field of (big) data, and how this translates in horizontal and vertical movements by the various players.

Competition, collaboration and integration

In one of her speeches on data as a new valuable source, European Commissioner for Digital Agenda Neelie Kroes emphasized the importance of a competitive data ecosystem.

> *"[…] there's a great, competitive market out there to get the maximum value from "big data". That competition is a good thing. It's helping our digital society. Helping innovation, with new and exciting services available for people every day. And it's good for our economy, giving us a much needed boost at a time of crisis."*[80]

But not only competition, also collaboration is key to leveraging the potential of the multidisciplinary field of big data. This multidisciplinary quality of its challenges and opportunities is not only confined to the complex stack of infrastructural elements, analytical techniques and visualization tools. It also includes organizational and HR-expertise and extensive domain knowledge from the specific vertical where big data is applied.

Teunissen and Kuijpers from SURFnet note that collaboration in the big data arena is becoming increasingly important. According to them, trying to create a one-stop shop for big data is not feasible, even for major actors such as IBM or Google. New opportunities simply require collaboration, especially in solutions for more

[79] European Commission (2012). Commission proposes a comprehensive reform of the data protection rules. Available at: http://ec.europa.eu/justice/newsroom/data-protection/news/120125_en.htm

[80] Kroes, N. (2013). The Economic and social benefits of big data. Webcam conference on big data / Brussels. Available at: http://europa.eu/rapid/press-release_SPEECH-13-450_en.htm

specialized companies. In that respect, the workings of the big data ecosystem fit the general description of innovation ecosystems[81] in which the collaboration between individual companies allows them to create value that no single company can deliver on its own. This combination of competition and collaboration in the big data ecosystem was expressed by Ben Woo from Forbes as follows:

> *"What impressed me most about the Big Data industry is something that I will shamelessly coin collab-petition. With the exception of distribution vendors, every other vendor in the Big Data space has a need to collaborate with multiple partners, many of whom are also competitors. I guess this is the beauty of a multi-vendor open-source market. Everyone has to get along (in some way) with everyone else."*[82]

The need for collaboration corresponds with the statement of Evert-Jan Tromp from SAP regarding open standards and interoperability of their products:

> *"All large companies maintain open standards and actively collaborate to enable their systems to work together. Having closed systems is rare these days, as most customers are running a patchwork of software from different vendors and thus demand that our systems can be integrated or connected with the software packages they are already running."*

It is difficult to properly assess what companies are currently dominating the big data ecosystem, as exact numbers of market share are hard to come by and would be difficult to interpret as the ecosystem comprises many different kinds of intertwined services. However, if collaboration is an important necessity in the big data ecosystem, the number of partnerships could serve as a potential indicator for market dominance. Figure 4 (also available as a list in Annex 3) provides an overview of more than 50 companies with the highest number of partnerships with other organizations in the Hadoop ecosystem.[83]

They contain many of the players from the two big data landscapes from paragraph 3.4.1. Especially the independent Hadoop distribution providers Cloudera, Hortonworks and MapR are well connected. But also the more traditional IT-vendors such as IBM, EMC, HP, Microsoft, Oracle, SAP, VMware, Cisco and Intel are well connected. Amazon and Dell, that did not make the list last year[84], have improved their network considerably in 2013. A notable company that is not included in the list is Google.

[81] Adner, R. (2006). Match your innovation strategy to your innovation ecosystem. In: *HBR*, April. Available at: http://pds12.egloos.com/pds/200811/07/31/R0604Fp2.pdf
[82] Woo, B. (2013). A Mind Blowing Big Data Experience: Notes from Strata. In: *Forbes*. Available at: http://www.forbes.com/sites/bwoo/2013/02/27/a-mind-blowing-big-data-experience-notes-from-strata-2013/
[83] Datameer (2013). Hadoop Ecosystem: Who has the most connections. Available at: http://datameer2.datameer.com/blog/wp-content/uploads/2013/01/hadoop_ecosystem_full2.png
[84] Taylor, R. (2012). The Hadoop Ecosystem, Visualized in Datameer. Available at: http://www.datameer.com/blog/uncategorized/the-hadoop-ecosystem-visualized-in-datameer.html

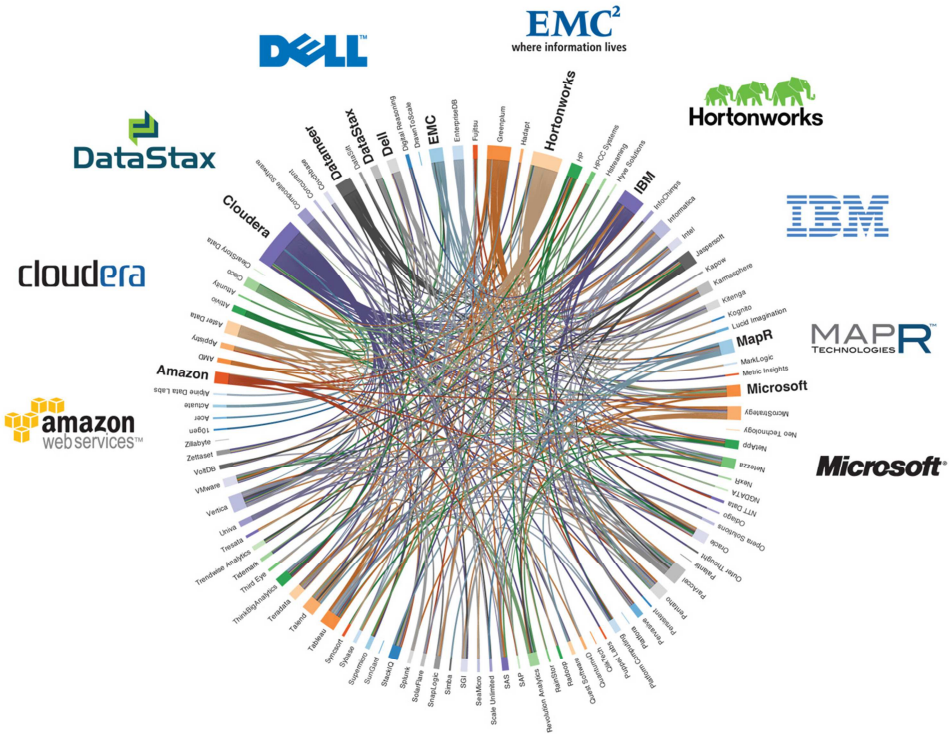

Figure 4 Partnerships in the Hadoop Ecosystem (Datameer, 2013)

As noted before, many traditional IT-vendors are active in different domains – or steps of the value creation process - within the big data ecosystem. Matt Turck, creator of the big data landscape in Figure 2, noted that there are many organizations that could fall in more than one category. In his representation of the big data landscape there is a special category for organizations that explicitly operate across both infrastructure and analytics. This category hosts some of the biggest players such Microsoft, Google, Amazon, Oracle, SAP, SAS and VMware.

As the big data ecosystem evolves, many new companies emerge. Subsequently, larger companies try to strengthen their position. They will not only develop new products and forge partnerships, but they will also acquire promising startups to improve and augment their propositions with analytics platforms, visualizations and applications.[85]

Infochimps CEO Nick Ducoff provides an explanation for this dynamic between the specialized nature of many big data start-ups and the more generic platforms they build on:

> *"If you are best at the presentation layer, you don't want to spend your time futzing around with databases […] What we're seeing is startups focusing on pieces of the stack. Over time the big cloud providers will buy these companies to integrate into their stacks."*[86]

[85] ESG (2012). Boiling the Ocean of Control Points in the Hadoop Big Data Market. Available at: http://www.esg-global.com/blogs/boiling-the-ocean-of-control-points-in-the-hadoop-big-data-market/

[86] Watters, A. (2011). Scraping, cleaning and selling big data. In: *O'Reilly Strata.* Available at: http://strata.oreilly.com/2011/05/data-scraping-infochimps.html

According to a report form Orrick on emerging big data companies, based on deals and investments (with a focus on the US), big data financing activity has increased significantly since 2008 (see Figure 5).[87] Recent years also have seen the take-off of the first IPO's of big data companies. The number of mergers and acquisitions has increased rapidly, from 1 in 2009 to 17 in 2012. According to the Orrick report IBM was the most active acquirer of big data companies in 2012, followed by Oracle that also bought a few companies and numerous others who have bought one data company.

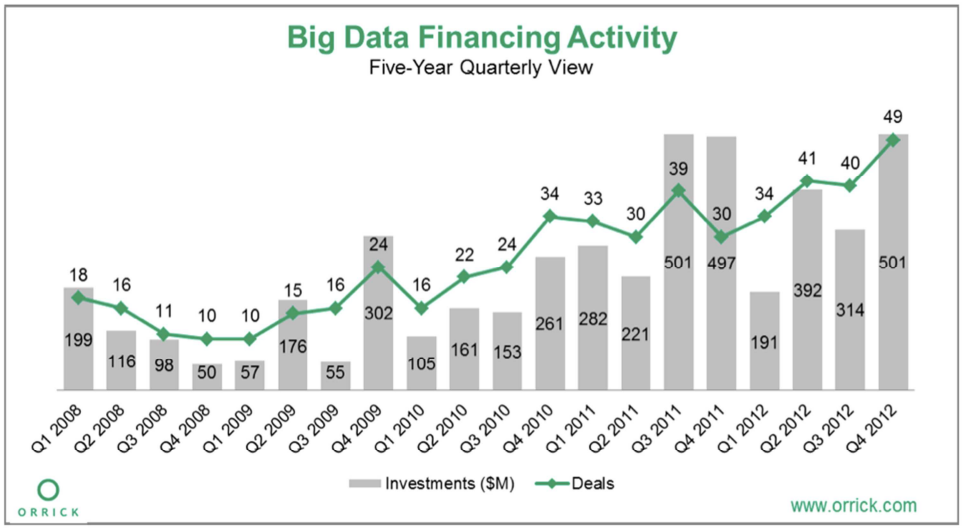

Figure 5 Big data financing activities (Orrick, 2012)

In terms of integration, the development of active roles for non-traditional data companies in the big data ecosystem is also interesting. This is most obvious in the role of data providers. As described above, many governments, companies (e.g., Nike, Alliander) and even individuals can play an active role in addition to their role as consumer of big data services.

Some companies are even expanding their role in the data ecosystem and take on unexpected roles. One of the most telling examples is Amazon. Its role as a big data powerhouse is clearly illustrated by a statement from VMware COO Carl Eschenbach in response to Amazon's rise, at the expense of VMware, in the big data field:

> *"[…] I look at VMware and the brand reputation we have in the enterprise, and I find it really hard to believe that we cannot collectively beat a company that sells books."*[88]

Companies, building on their own experience and expertise, might become data service providers for others. John Deere transformed itself, to some extent, from a manufacturer of tractors to a highly advanced business intelligence service for

[87] Orrick (2012). The Big Data Report. Available at: http://www.cbinsights.com/big-data-report-orrick
[88] Assay, M. (2013). VMWare: If Amazon wins, We all lose. In: *Readwrite*. Available at: http://readwrite.com/2013/03/01/vmware-if-amazon-wins-we-all-lose#awesm=~ohpgw6RGoncktJ

farmers. Retailer Walmart develops its own big data tools in its Walmart Labs and recently released some of its mobile tools as open source, for others to use. The case of CIBAS (paragraph 3.3.1) shows how they have evolved to a data knowledge center for other waste management services that want to leverage data. They are considering to extent their activities in other public spaces in the future as well. These examples illustrate how even organizations for whom data originally was not part of their primary process, can become data players for different steps of the value creation process of data.

3.4.3 Trends and future developments

In this section we will assess how the big data ecosystem could evolve in the future by describing the relation between value creation, competitive differentiation and innovation.

Big data and competitive differentiation

MIT surveyed over 3000 executive managers about their thoughts regarding the main challenges for their business for the next couple of years. 'Innovation to achieve competitive differentiation' came out on top, cited by six out of ten respondents, followed by 'growing revenues' (five out of ten) and 'saving costs' (4,5 out of ten).[89]

These challenges are in line with the main qualities ascribed to big data and the drivers for organizations to implement (big) data solutions. A review from Chen[90] of scientific literature on business intelligence and analytics showed that "competitive advantage" was one of the top four most discussed topics. According to the OECD "Data are a core asset that can create a significant competitive advantage and drive innovation, sustainable growth and development."[91] Both the McKinsey report and the MIT report consider big data an important instrument for organizations to differentiate themselves from others. Brynjolfsson and McAffee arrived at similar conclusions after studying 330 companies in the US:

> *"[…] the more companies characterized themselves as data-driven, the better they performed on objective measures of financial and operational results. […] Companies in the top third of their industry in the use of data-driven decision making were, on average, 5% more productive and 6% more profitable than their competitors."*[92]

A company that has adopted data analytics seems to have an advantage over competitors that lack such a strategy. And generally, in terms of competitive differentiation from the non-data-users, the exact nature of the applied data strategy is of second importance: initially, generic business intelligence platforms might

---

[89] Lavalle, S., et al (2010) Analytics: The New Path to Value. MIT Sloan Management Review. Available at: http://cci.uncc.edu/sites/cci.uncc.edu/files/media/pdf_files/MIT-SMR-IBM-Analytics-The-New-Path-to-Value-Fall-2010.pdf

[90] Chen, L., et al. (2012) Business Intelligence and Analytics: From Big Data to Big Impact. In: *MIS Quarterly*.Vol. 36 No. 4, pp. 1165-1188.

[91] OECD (2013) Exploring Data-Driven Innovation as a New Source of Growth: Mapping the Policy Issues Raised by "Big Data", *OECD Digital Economy Papers*, No. 222, OECD publishing. Available at: http://dx.doi.org/10.1787/5k47zw3fcp43-en

[92] Brynjolfsson, B., and McAffee, A. (2012) Big Data: The management revolution. In: *Harvard Business Review*, October. Available at: http://hbr.org/product/big-data-the-management-revolution/an/R1210C-PDF-ENG

suffice. Once in place, these first steps in the field of business intelligence can be expanded with big data technologies as the organization gains experience.

From infrastructure to analytics and visualizations
As mentioned in chapter one, it has become much cheaper and easier for organizations to leverage data by using big data technologies. As Paul Miller notes:

> *"There is, of course, still a need for people with specialist knowledge, hard-won skills, and painfully gained experience. But you no longer (if you ever really did) need to install a Hadoop cluster with your bare hands and juggle complex statistical formulae in your head to benefit from the growing prevalence of data in all aspects of business."*[93]

However, when more and more organizations adopt big data technologies – it becomes increasingly difficult to differentiate oneself. As business intelligence becomes more advanced and more focused on real-time insights rather than historical and periodical information, the demands from the users of big data technologies change as well. The value creation process of data must not only provide analytics, they need to leverage actionable output that is easy to understand and act on[94]. The NESSI report states:

> *"The technological improvements in infrastructure, database, telecoms, and so forth are nothing without applications that can take advantage of them."*[95]

As Figure 6 illustrates, it is expected for the next couple of years that most of the value of big data will be added by advanced analytical techniques, predictive analytics, simulations and scenario development, and advanced data visualizations.[96] These are the most important growth areas for the near future.

[93] Miller, P. (2013) Visualization, the key that unlocks data's value? Available at: http://cloudofdata.com/2013/04/visualisation-the-key-that-unlocks-datas-value/
[94] Dumbill, E. (2011). Five data predictions for 2012. In: *O'Reilly Strata*. Available at: http://strata.oreilly.com/2011/12/5-big-data-predictions-2012.html
[95] NESSI (2012). Big Data: A New World of Opportunities, p. 17. Available at: http://www.nessi-europe.com/Files/Private/NESSI_WhitePaper_BigData.pdf
[96] Russom, P. (2011). Big Data Analytics. TDWI Best Practices Report. P. 24. Available at: http://tdwi.org/research/2011/09/best-practices-report-q4-big-data-analytics.aspx

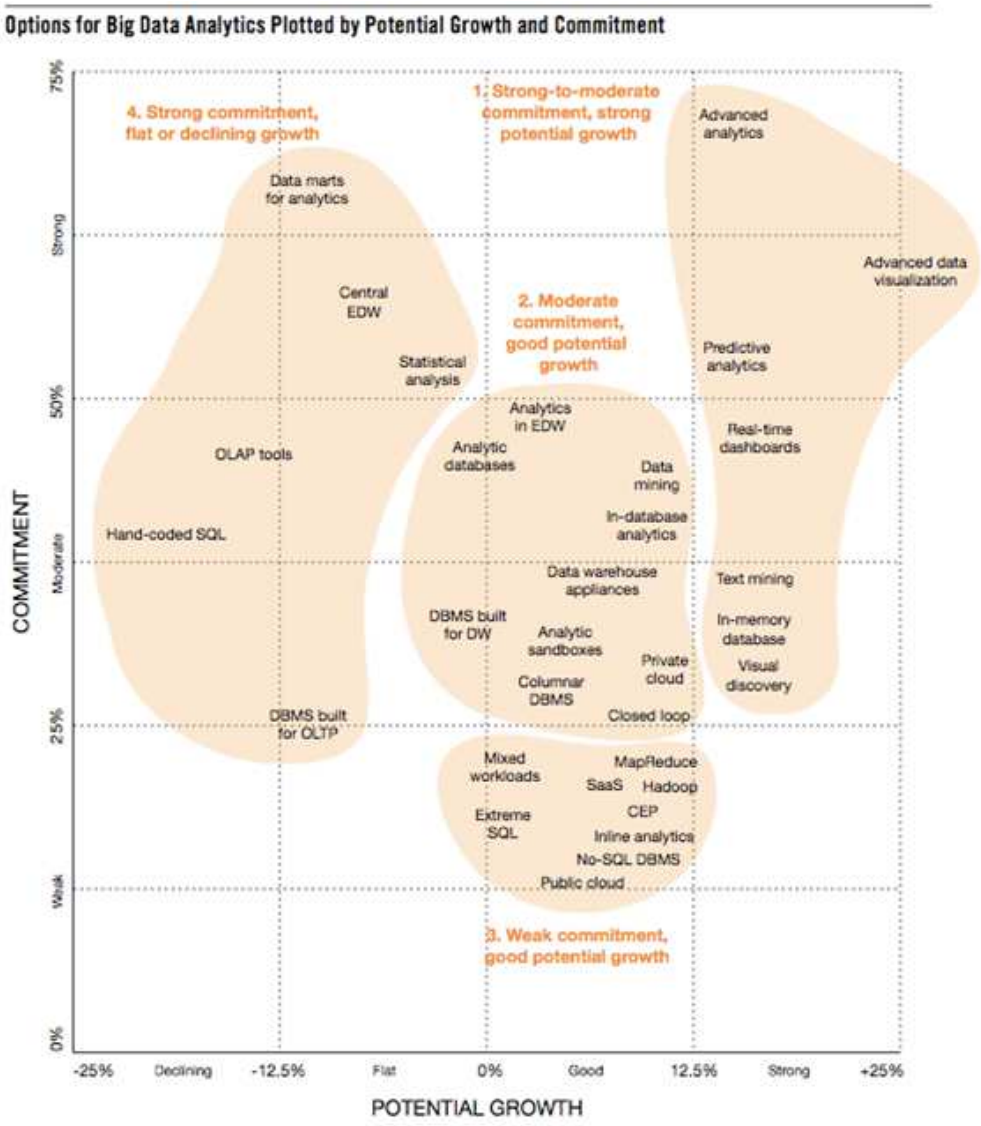

Figure 6 Potential and growth of big data technologies (TDWI, 2011)

The evolution in value creation with data seems to be reflected in the figures from the earlier mentioned Orrick report on financial activities regarding big data. In the last five years, both in terms of deals and especially investments, the focus has shifted from big data infrastructure to big data analytics and applications. Whereas in 2008 infrastructure accounted for 46% of big data investments, this decreased to 31% in 2012. These numbers also illustrate that the analytics, visualization and application layer, the 'last mile of big data'[97] is where most of the value of data is generated and where true differentiating quality resides as the commoditization of big data technologies continues.

Specialized solutions and expertise

In order to establish competitive differentiation, it is becoming increasingly important to not only generate the best actionable output, but also to present it in such a way

[97] ESG (2012). Boiling the Ocean of Control Points in the Hadoop Big Data Market. Available at: http://www.esg-global.com/blogs/boiling-the-ocean-of-control-points-in-the-hadoop-big-data-market/

that it is aligned with the business process that it strives to support. Generic analytical techniques can be important building bricks, but as the threshold for competitive, differentiating big data technologies increases, it needs to be optimized for the context in which they will be used. As Teunissen and Kuijpers from SURFnet stated during the interview:

> *"The more a data product or service shifts to the eventual output (from storage to analytics to visualization), the less can be delivered by generic and sector-independent tools. You need more and more knowledge about the domain in which the output will be used, and account for its specific context."*

Wijsman and Wormer from Almere Data Capital confirm that in order to gather truly new insights that are sector-specific, the combination of generic data-skills with domain-expertise are crucial. At the European Data Forum 2013 Siemens director Big Data Initiatives Gerhard Kress emphasized the importance of research on vertical algorithms. In an analysis of the big data market ESG, an IT market research and advisory firm, noted how big data service companies try to obtain dominant positions in certain vertical industries: "[…] where whomever has 'the most data scientists with a vertical bent' may win."[98]

However, this is not only a matter of vertical specialization. Another route to differentiating value creation is the democratization of big data technologies that enable non-experts in the field of big data, but experts in their own field, to explore and experiment with data to look for new combinations and possibilities, and tweak the available tools to fit their own goals and working environment.[99] This kind of engagement with the analytics through visualizations was also considered an important development during the workshop.

Incremental versus radical innovation

As discussed in paragraph 3.2, companies use big data technologies to various ends, ranging from cost saving through financial monitoring to revenue growth through new marketing strategies and product development. As the MIT report points out, these goals strongly depend on the maturity of an organization regarding its ability to deploy big data and related technologies. As companies gain big data experience, the balance between cost saving and revenue growth will shift. Deploying big data for marketing and sales becomes more important, as does, to a lesser extent, product research and strategy development purposes. Such an organizational evolution in terms of big data maturity was confirmed during the workshop by both data service providers and data service consumers: the first priority for organizations is to improve primary business processes. More radical innovations that upend current practices, or create new ones, require more experience, commitment and a more solid belief in the potential of leveraging data. Still, the use of big data for incremental changes can be a helpful precursor for more radical innovations.

Although some companies have indeed shifted their big data priorities from cost efficiency to revenue growth to realize innovation for competitive differentiation, this

---

[98] Idem

[99] Brave, S. (2012). We don't need more data scientists – just make big data easier to use. In: *GigaOm*. Available at: http://gigaom.com/2012/12/22/we-dont-need-more-data-scientists-just-simpler-ways-to-use-big-data/

has not yet resulted in a grand scale proliferation of more radical innovations in which (networks of) organizations rethink products or even whole systems.[100]

Two interesting examples of such kinds of innovation are sports brand Nike and the Dutch IJkdijk initiative, a new dike monitoring system. Both have redesigned some of their products as 'data products'. Nike introduced the online Nike+ platform, the Nike+ sensor that can be clipped on running shoes, an app that tracks runs and more recently the FuelBand, a wristband that tracks activities and calories burned during the day. Although its core value proposition – supporting people to be physically active and healthy – has not changed, Nike is now more and more providing this proposition by using data that enables users to set their goals, track their progress and include social elements. It has also created an API that allows trusted third parties to develop apps based on this data-driven platform. The IJkdijk is the result of a research program in which a dike in the north of the Netherlands was equipped with sensors. The collected data is analyzed and visualized to improve dike monitoring and water management.

The earlier mentioned John Deere tractors and platform, and the autonomous Google car are also illustrative examples of data products. The development of autonomous and smart cars is in line with a bigger transition in which the role of ICT and data in cars (and transit in general[101]) is increasing rapidly, as the shift towards collaboration between Detroit and Silicon Valley indicates.[102] Using data in education - in the form of learning analytics - holds the promise of enabling personalized, adaptive learning environments, rather than the one-size fits all system which is currently the standard.[103] [104] Such changes may a have profound impact on the way education is organized, the role of the teacher, the relation between formal and informal learning and the development and use of learning materials, rather than only slightly improving the current system.

When a specific domain is indeed subject to datafication, this could lead to a restructuring of the players involved. If – as in the case of Nike – physical exercise indeed becomes a data-driven activity it will no longer suffice to use big data to slightly improve margins on shoes or clothes that do not support this transformation. Deploying data to segment customers that no longer have an interest in products that do not allow them to track their progress is not effective and does not provide competitive differentiation. Companies need to be aware of the way data might impact the environment in which they act. With the massive adoption of smartphones and apps, Nike has to deal with new kinds of competitors as services that offer data-supported physical exercise, like Runkeeper or Runtastic, abound. In

[100] Lavalle, S., et al (2010) Analytics: The New Path to Value. MIT Sloan Management Review, p. 15. Available at: http://cci.uncc.edu/sites/cci.uncc.edu/files/media/pdf_files/MIT-SMR-IBM-Analytics-The-New-Path-to-Value-Fall-2010.pdf
[101] Madrigal, A. (2012). Driverless Cars Would Reshape Automobiles and the Transit system. In: The Atlantic. In: *The Atlantic*. Available at: http://www.theatlantic.com/technology/archive/2012/09/driverless-cars-would-reshape-automobiles-and-the-transit-system/262953/
[102] Muller, J. (2013). Silicon Valley vs. Detroit: The Battle For The Car Of The Future. In: *Forbes*. Available at: http://www.forbes.com/sites/joannmuller/2013/05/08/silicon-valley-vs-detroit-the-battle-for-the-car-of-the-future/
[103] Siemens, G. (2010) What are Learning Analytics? Available at: http://www.elearnspace.org/blog/2010/08/25/what-are-learning-analytics
[104] NMC (2012) Horizon Report: 2012 Higher Education Edition. Available at: http://www.educause.edu/library/resources/2012-horizon-report

a new market of wearable technology it might have to compete with Google (Google Glass) and Apple (the highly anticipated iWatch[105]) that are also its partners for the use and distribution of its apps on the iPhone and Android phones.

During the workshop, one of the data service providers stated that in each field (possibly new) disruptive companies that initiate these kinds of changes towards 'datafication', could force more traditional ones to adapt as well[106]. The companies that shape these changes will have the best opportunities to profit from it. And the ones that are most able to adapt their business to such changing environments will win.

However, creating more radical data-driven innovations is very hard for an organization. According to Maxwell Wessel from the Harvard Business School think-tank Forum for Growth and Innovation, this is especially true for more mature organizations (not in the big data-sense) as opposed to start-ups that try to bring a new solution to the market, a phenomenon that was also described by Clayton Christensen in 'The innovators Dilemma'[107].

> *"Once a business figures out how to solve its customers' problems, organizational structures and processes emerge to guide the company towards efficient operation. Seasoned managers steer their employees from pursuing the art of discovery and towards engaging in the science of delivery. Employees are taught to seek efficiencies, leverage existing assets and distribution channels, and listen to (and appease) their best customers."*[108]

Furthermore, large organizations can have many different departments and many different IT systems, which can make the implementation of more radical innovation very challenging. On the other hand, as was discussed during one of the interviews and the workshop, especially larger companies are able to direct (IT) resources to experiment with new possibilities with data. SMEs are more focused on their daily activities rather than innovation with big data technologies, which – although they are becoming cheaper – still require an investment that is relatively large for smaller companies. The idea of more innovative and agile SMEs might be more applicable to new players than existing SMEs.

As the results from MIT and TDWI illustrate, data is currently predominantly used for incremental improvements for existing products and services. Especially for organizations that do not have their roots in data nor have gained experience with big data and data driven discovery and decision-making, more transformative innovations are hard to realize. But even more experienced organizations often direct their big data prowess primarily at internal processes. These kinds of innovations mostly affect the company itself rather than create value for its customers.

---

[105] Bilton, N. (2013). Disruptions: Where Apple and Dick Tracey May Converge. In: *The New York Times*. Available at: http://bits.blogs.nytimes.com/2013/02/10/disruptions-apple-is-said-to-be-developing-a-curved-glass-smart-watch

[106] Redman, T. (2012). Integrate Data into Products, or Get Left Behind. In: *HBR*. Available at: http://blogs.hbr.org/2012/06/integrate-data-in-products-or-get/

[107] Christensen, C. (1997). The Innovator's Dilemma. New York: HarperCollins Publishers Inc.

[108] Wessel, M. (2012). Why Big Companies Can't Innovate. In: *HBR*. Available at: http://blogs.hbr.org/2012/09/why-big-companies-cant-innovate/

Data marketplaces and data-driven innovation
Rethinking products and systems from a data-perspective, rather than incremental innovation, also puts the role of data markets in another light. In the same MIT report, top priorities regarding data were integration, followed by consistency and trustworthiness. Access to data was mentioned last in a list of ten, selected by fewer than 20% of the respondents. Integration of the available data, however, topped the list and was mentioned by almost 45% as one of their top three priorities.

Although it is not clear why access to data is not a top priority, one explanation could be that most organizations innovate with the data they already have, hence the priority of data integration. During the workshop this issue was discussed as well and both the data service consumers as the big data experts agreed that for most organizations this is difficult enough considering the most prominent goals mentioned earlier. For the kinds of innovations that organizations create they often use their own data and do not require new or external datasets that are more difficult to acquire. During the workshop most data-owners also focused primarily on the data they already captured and collected, rather than identifying new and potential internal and external data sources.

But when organizations strive for more radical innovations, new data sources might be crucial and data access becomes a bigger problem. In his presentation during the European Data Forum 2013 Knut Sebastian Tungland, Chief Engineer IT at the Norwegian energy company Statoil, explained how they were considering the use of sounds and behavior from sea animals (fish, clams, seals) to detect problems like gas- or oilspills or explosions. This means they will have to be able to capture this data and learn how to make sense of it by working with marine biologists.

Although some companies have successfully created or integrated multiple data sources (e.g., Nike, John Deere), generally, as Alistair Croll notes on the idea of combining various external datasets:

> *"[…] it's hard to find data, be confident of its quality, and purchase it conveniently. That's where data marketplaces meet a need."*[109]

Furthermore, the barriers are not only finding the right data, but also finding organizations willing to share or sell their data. During the workshop most of the participants stated they were not willing (or allowed) to share their datasets with others. The few organizations that were willing to share data, only wanted to do this with specific datasets, in specific partnerships under very specific conditions, which were yet to be examined and determined. The reasons for not willing to share data were regulations (e.g., privacy) and the fact that they deemed their information to be competition sensitive. This friction in data sharing between data producers and consumers, according to the NESSI whitepaper, hinders the development of new, innovative data-driven business models[110].

[109] Dumbill, E. (2012). Microsoft's Plan for Big Data. In: O'Reilly, *Planning for Big Data*, p. 38. Available at: http://oreilly.com/data/radarreports/planning-for-big-data.csp
[110] NESSI (2012). Big Data: A New World of Opportunities, p. 17. Available at: http://www.nessi-europe.com/Files/Private/NESSI_WhitePaper_BigData.pdf

As companies mature in terms of their ability to leverage big data, the need for competitive differentiation will drive more radical innovation beyond mere efficiencies. In such a scenario, the role for data markets could evolve from simply offering a rich variety of datasets, to a more active role in which they offer both context to the data (using metadata and identifiers) and the tools to explore datasets and new combinations – as Infochimp already does.[111] This kind of integration of both datasets and analytics platforms is also part of Microsoft's big data strategy. Microsoft takes big data integration even a step further with a portfolio that also includes cloud infrastructure, business applications and tools for visualization. It even released an operating system for houses[112], which makes it easier to monitor and automate them.

It could be a preview of things to come in the big data ecosystem as the increasing size of datasets makes transport more and more a bottleneck. If data users want to have access to new, big datasets to experiment with the possibilities of new combinations, the close proximity of these datasets to each other and analytical tools and applications might become a necessity. As Ed Dumbill observes:

> *"Data that is too big to process conventionally is also too big to transport anywhere. IT is undergoing an inversion of priorities: it's the program that needs to move, not the data. Because of this, we're seeing the increasing integration between cloud computing facilities and data markets*."[113]

A few centralized data services, from players like Microsoft, who integrate both data markets and analytical tools – or that allow others to build analytical tools on top of it – might become the norm.

On the other hand, leveraging data is not only a matter of big data in the sense of big volumes. Rufus Pollock advocated the use of 'small data'[114], in the same spirit as his earlier mentioned statement on the value of open data. According to Pollock the explorations and experiments that many individuals can perform on their own desktop computer or laptop could add more value than a single organization can create by itself with a big data infrastructure. According to the NESSI paper:

> *"Many data providers, who have stockpiled vast amounts of interesting data, struggle with the problem of finding ideas for creating novel services using their data, identifying what makes their data relevant for potential consumers, and deploying solutions for rapid integration of data for loosely defined services."*[115]

Providing easy access to smaller, representative samples of big datasets, which can be more easily transported, could lead to more innovation. Data market places

---

[111] Dumbill, E. (2012) Data marketplaces. In: O'Reilly, Planning for Big Data, pp. 47-52.
[112] Simonite, T. (2013). Microsoft Has an Operating System for Your House. In: *Technology Review.* Available at: http://www.technologyreview.com/news/517221/microsoft-has-an-operating-system-for-your-house/
[113] Dumbill, E. (2012). What is big data? In: O'Reilly, planning for big data, p. 14. Available at: http://oreilly.com/data/radarreports/planning-for-big-data.csp
[114] Pollock, R. (2013). Forget big data, small data is the real revolution. In: *The Guardian*. Available at: http://www.theguardian.com/news/datablog/2013/apr/25/forget-big-data-small-data-revolution
[115] NESSI (2012). Big Data: A New World of Opportunities, p. 19. Available at: http://www.nessi-europe.com/Files/Private/NESSI_WhitePaper_BigData.pdf

could play a role in this kind of interaction between data owners and innovative users.

Still, data-driven innovations that combine datasets from different organizations, let alone sectors, remain scarce. Even when valuable data is available, only organizations that have reached a high big data maturity and are thus willing to explore new possibilities will look for them.

During the second workshop the participants were asked to make a list of their datasets with a brief general description of its contents. Combined, these datasets and a list of additional 'open datasets' constituted a hypothetical trans-sectoral data marketplace. Next, both data consumers and data experts were asked to look for interesting combinations that could possibly add value to their business processes or products, and to explore these ideas further in small groups. Table 2 provides an overview of the number of combinations between the different sectors. The data-owners can be found in the top row, with the same data-owners as interested parties in the left-hand column.

Table 2 Potential use of datasets from different sectors, by different sectors in the workshop

Sector (Potential user/ owner)	**Safety and oversight**	**Energy**	**Mobility and transport**	**Health**
Safety and oversight	1	2	5	2
Energy	1	5	3	-
Mobility and transport	3	2	3	2
Health	2	1	-	1

The datasets from the energy and mobility sector were the most popular ones. Both for the National Police Services Agency and the Tax and Customs administration hypothetical use cases were found for investigation and monitoring purposes (e.g., spikes in energy consumption that could indicate the cultivation of cannabis, or patterns in traffic flows in transport and postal deliveries).

Organizations in both the energy and mobility sector were also interested in data from their own sector. On the one hand, especially for the companies in the energy sector, this kind of data would be valuable for benchmarking and marketing. In the mobility sector there was much interest to combine the data from different companies to better streamline supplementary services. Only one combination that was explored during the workshop was deemed interesting enough for the involved parties to consider for further development. This was a combination of energy and mobility data that could create new services for their customers in terms of climate control in the home.

The data marketplace exercise and the plenary discussion afterwards confirmed that creativity and the willingness to explore new combinations are crucial and rare qualities in organizations. Additionally, even if new combinations can be found, the daily routines in organizations allow for little room for experimentation. The lack of willingness to share data is an important barrier as well, as will be discussed in more detail in paragraph 3.5. Furthermore, in some cases companies are in a good position to capture data that is extremely useful for others, but they do not have the knowledge or incentive to do so. Again, John Deere is an interesting example. The data John Deere gathers and analyzes is valuable for many users, but it was not

available until the tractors were equipped with sensors and became data-capturing devices. It required vision and determination to create this new kind of infrastructure.

Based on the observation that it is hard to initiate and implement data-driven innovation, data marketplaces might take up a more pro-active role in the big data ecosystem with the purpose to stimulate innovation by exploring relevant combinations themselves and delivering information and initial analyses to those organizations that need them, when they need them (see Figure 7).[116]

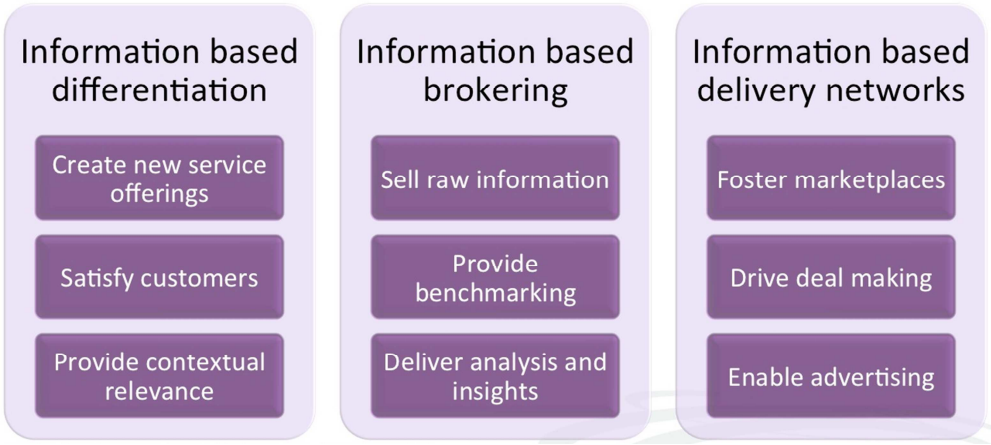

Figure 7 New roles for data marketplaces and data-intermediaries (Wang, 2011)

In this scenario, these data players actively broker deals for organizations that capture information that could be valuable to others, or data that is not yet being used or even captured. Wijsman and Wormer also expect that this kind of role for data marketplaces, in which they stimulate collaboration and new combinations will become more important. In the Netherlands, the Big Data Value Center in Almere aspires to foster data-driven innovation by stimulating and facilitating (cross-sectoral) combinations between different private and public stakeholders. In one of its workshops it challenged representatives from over twenty SMEs in the field of ICT to come up with new products and services based on a wide variety of available datasets. During the various brainstorms the ideas for intermediating data-platforms (between data-owners on the one hand and app developers and data-users on the other) was most prevalent, rather than specific data-products. In this light, the Aspen report noted:

> *"New types of data-intermediaries are also likely to arise to help people make sense of an otherwise-bewildering flood of information. Indeed, data-intermediaries and interpreters could represent a burgeoning segment of the information technology sector in the years ahead."*[117]

These new kinds of brokers could actively engage all kinds of organizations as active players in the big data ecosystem as data-providers, driving innovation and also taking a central role in all kinds of new data constellations.

[116] R. Wang (20120. What a Big Data Businessmodel looks like. In: *HBR*. Available at: http://blogs.hbr.org/2012/12/what-a-big-data-business-model/
[117] Bollier, D. (2010). The Promise and Peril of Big Data, p. 40. Washington DC: The Aspen Institute. Available at: http://www.aspeninstitute.org/publications/promise-peril-big-data

3.5 Barriers to innovation

The road to data-driven innovation is not paved. Despite the promise of economic and societal benefits, the multidisciplinary and breadth of the data ecosystem and the complexities involved in the value creation process present a number of barriers that seem to hinder the optimal and innovative exploitation of available and potential data sources.

3.5.1 Getting started and awareness

Combining the right data sources and applying the appropriate analytical tools to extract the right information and conclusions requires an important initial quality in an organization: the awareness of and knowledge about the value of big data, and knowing how to get started in the first place. In that sense, big data – as a generic concept – lacks a clear killer application that convinces aspiring big data users to structurally exploit their data.

In many cases, organizations are not fully aware of the range of available datasets – even in their own databases - and the potential value they might hold. While datasets may already exist in some departmental IT systems, they have not been identified as holding added value when combined and analyzed with other data sources. In some cases data is not yet captured, because an organization lacks the right infrastructure or processes. But the lack of awareness especially applies to the potential use of external datasets. Many organizations are reluctant to explore the options of using open datasets or acquiring or exchanging datasets with other parties.

Some organizations know they have a lot of data but do not know how to leverage it – as will be discussed in more detail below. Commenting on the lack of familiarity with the possibilities of data analytics in the scientific field, Teunissen and Kuijpers from SURFnet state:

> *"There are many researchers with very large and complex datasets. It might very well be that they are simply not aware that there are supercomputers that can handle their data in an affordable way."*

According to Wormer, awareness of the possibilities of big data is not enough for an organization to set innovations into motion. Instead of considering the use of big data as a straightforward input-analyze-output, "there should be a stronger focus on creativity. Adding value requires an intuitive creative process." Or, as put by Stefaan Verhulst in the Aspen Institute report: "The real challenge is to understand what kind of data points you need in order to form a theory or make decisions."[118]

3.5.2 Technological issues

As discussed in chapter one, an often-used definition of big data describes the three V's of big data: volume, variety and velocity. Although this definition takes a rather limited view on the broader big data phenomenon, it refers to the most important elements of many of the most common technological issues. For most of these challenges traditional database and analytical technologies no longer suffice.

[118] Idem, p. 14.

Although the volume of big data in terms of storage may not be the biggest concern for most organizations, data transport of multiple or large and dynamic datasets is a major technological issue that merits much attention. In this respect, Hans Wormer from Almere Data Capital states:

> *"There need to be very solid plans regarding the preparation and configuration of big data before one can start to apply analytics. It does not concern one dataset that needs to be transported over a simple highway. It is rather oil tankers of data that are shipped."*

In cases where the transport of datasets is an important issue, there is an emerging trend to keep data stored in central (cloud) platforms. Analytics are then 'brought' to the data, instead of the other way around. This trend is particularly relevant in domains - such as health care - where there are true explosions of data, combined with the demand for fast and real-time analyses.

As for the variety of data, according to the MIT Sloan Management report the main priority of organisations is the integration of datasets.[119] For one of the interviewed organizations, the lack of standardization was the most prominent and urgent technological issue. In situations where various organizations opt to collaborate while using one central system, problems regarding incompatibility are bound to arise. Although in some cases conversion is possible, in most cases it requires substantial changes in software systems to produce output that can be used by other parties. An interesting example is the CIBAS case. Because its partners produce data in different kinds of formats, it had to create a custom-made solution to integrate them. However, as CIBAS illustrates, standardization is also a matter of non-technical procedures: "there also needs to be some kind of central concepts and understanding of working with data. If too many deviations or custom-made and ad-hoc solutions need to be used, all envisioned synergetic profits get lost."

In addition to the three V's, a lack of big data talent, especially in combination with relevant domain knowledge, is a big challenge for organizations[120]. So-called 'data scientists' are rare and the combination of data skills and vertical expertise will become crucial for competitive differentiation. In many cases, data and professional fields are so specific that current technological tools are not easily integrated, or user-friendly enough to work with. Teunissen and Kuijpers from SURF described a common scenario that features a scientist that does not have the required skills to explore certain combinations of datasets:

> *"For instance: 'I am a biologist and I know everything about bacteria and I have a nice dataset. But I do not know how to combine this with datasets on the weather.' That is the biggest problem that scientists face, especially in the field of life sciences. They would like to explore and examine these kinds of combinations, but get stuck because the data and required knowledge is too complex for them."*

---

[119] Lavalle, S., et al (2010) Analytics: The New Path to Value. MIT Sloan Management Review. p. 14 Available at: http://cci.uncc.edu/sites/cci.uncc.edu/files/media/pdf_files/MIT-SMR-IBM-Analytics-The-New-Path-to-Value-Fall-2010.pdf

[120] Rooney, B. (2012). Big Data's Big Problem: Little Talent. In: WSJ. Available at: http://online.wsj.com/article/SB10001424052702304723304577365700368073674.html

According to Teunissen organizations and their employees often do not have the right tools to process and use the data. An additional problem with the increasing importance of vertical expertise in data is that in many cases the technology offers solutions that are too generic in nature and do not meet the demands from specific domains. The Rotterdam Port Authority experienced this issue first-hand in the development and implementation of the tools it uses to handle its data:

> *"It is not that we need everything to be customized, but the data that we need to process is so specific […] it's just different from production plants or the service industry. They have different business models and different information needs and requirements for analytics. You cannot just take something from the shelf."*

In the following comment from Rotterdam Port Authority, the difficult consideration for any organization - balancing dependency on third parties for expertise and in-house development and acquiring of expertise - is illustrated:

> *"I have often thought: if only we could have built these tools ourselves. Now we are often too much dependent on external parties. And you can put all kinds of Service Level Agreements in place, but when a software developer has never seen a ship form the inside it becomes difficult to create good software."*

The interviewee explained how this became especially clear when they had to test new iterations that just were not good enough and did not properly address the domain specific requirements. As described earlier, for a successful implementation of big data, combining data science with domain specific knowledge, and being able to make analytics relevant and usable for the people who have to work with it is key.

The ability of organizations to harness the potential of data depends on its human resources and their data capabilities. Therein lies a major challenge for the next couple of years. According to McKinsey the shortage of data scientists and a data-savvy workforce will increase by 2018 to somewhere between 140.000 and 190.000 people with deep analytical skills and an additional 1,5 million managers and analysts in the US alone.[121] This will pose a problem for many organizations. According to Teunissen from SURFnet this also applies to science where there is a lack of expertise that combines general technical knowledge with domain specific knowledge such as biology or geology. Not only are there not enough data scientists, there is also a shortage of educators to teach future data experts.

Although education and training will be important, another option is the democratization of big data tools to fill the gap between the required tasks and the capabilities within organizations. The latter solution will enable professionals with domain expertise to deploy big data technologies without extensive data knowledge. However, there are still important limits as to what extent humans and human decision-making can be taken over by automated tools and algorithms. The democratization of tools may also increase a lack of understanding of the steps from input to output, creating a chasm between data experts and vertical professionals.

[121] Manyika, J., et al (2011). Big Data: the next frontier for innovation, competition, and productivity. McKinsey Global Institute, p.10. Available at: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

The democratization of big data technologies is – to a large extent – also a matter of visualization. As Wijsman from Almere Data Capital noted:

> *"Presentation of data is a specific skill. It is too important to dispose as just a small part of data analytics. How do you present 100 terabytes in an accessible and user-friendly fashion? This is a crucial part of big data and very difficult."*

3.5.3 Organizational and management issues

Big data challenges are as much organizational as technological. We will discuss some of the most pressing barriers below.

Stubbornness, a-symmetry and fear

A major hurdle in big data-driven innovation is scaling up from small data experiments to a more structural implementation. C-level support is required to bridge this gap. This means that the CEO and other Chiefs have to be convinced of the added value of data analytics. They need a business case and have to realize that an initial investment is required even though it is not yet certain what results these investments will yield.

But even when a data-driven strategy has proven itself (most likely on a small, departmental scale), and the boardroom has been won over and the technology is in place, this does not automatically mean that a broader implementation and expansion of data-related activities is a mere formality. There also has to be a clear data-strategy, as Wormer from Almere Data Capital emphasize. For example, in the city of Deventer, before the introduction of differentiated taxes for waste disposal, there was awareness in the waste management organization Circulus, that the data was available and that it could be valuable if used effectively. However, there was no organizational focus. Nobody was explicitly responsible for the use of all this data. Only when the processing of the data was explicitly assigned to the new department CIBAS did this change.

Furthermore, fear for organizational changes that are the result of a data-driven strategy can have a negative impact. During the workshop one of the participants noted that employees might fear losing their job to an algorithm that processes a bulk of data. Big data technologies that enable (semi-) automated decision-making could also have a negative impact on employability.

A related issue is the integration of data-driven strategies in daily practices. In some cases there is an a-symmetry in the required effort for data collection and the benefits that the analytics yield. The CIBAS waste management case illustrates how the benefits of a rigid data-collection process that is vital for differential tax tariffs, are not experienced by the employees who are responsible for the actual data-collection. Although the data-collection enables the municipalities to achieve their waste reduction goals and allows the waste management organizations to better allocate their resources, it does not have a real positive impact for the actual waste collectors who perform the data collection. In some cases it even has a negative impact for them. As described in paragraph 3.3.1, it changes the relation with their customers. Before the introduction of CIBAS the general notion was: "If there are customers with problems, we need to fix them as quickly as possible." Now they have to make sure that all adaptations are registered and processed before customers can be helped.

Furthermore, especially in larger organizations, bureaucracy and organizational structure impede cross-departmental use of data and further exploration of valuable data combinations. This notion was supported in the workshop discussion when the difference between smaller and larger companies was discussed.

Collaboration across companies and sectors
Even within a single organization the collaboration between different departments can be an obstacle to leverage available data. This problem is amplified when the data value network expands and third parties are included.

The collaboration across companies has to overcome even bigger differences in IT-systems, standards, processes, culture and priorities. As one of the participants of the workshop noted:

> *"Some organizations don't even have their own IT and databases properly assessed, organized and used, let alone that they are able to explore new data-driven partnerships."*

This barrier is not limited to the use of data from other parties. As mentioned earlier, organizations may not be aware of the potential value of their own data for others. And even when they do, they might not be willing to share their data with others because of regulations or because they deem the data to be too competition sensitive.

Accessibility and conflicts of interest
Especially for more transformative innovations that require complex constellations of partners, accessibility and conflicts of interest can be serious obstacles. The Port Rotterdam Authority case illustrates how the lack of willingness of the shipping companies to share the data with them - even when the most sensitive data is removed - impedes innovation. Again, this is also a matter of perceived asymmetry because the shipping companies are not convinced that sharing their valuable data could be beneficial to them as well.

Costs and financial resources
The importance of managerial support also becomes apparent in the allocation of financial resources. Although big data technologies have become much cheaper over the years, they still require an initial investment, especially when the costs also involve acquiring talent and external datasets.

These large upfront costs can be an important barrier, in particular for smaller companies. According to Wijsman larger companies have more leeway to allocate both money and resources to experiment with the possibilities of big data separately from their primary processes. Wijsman took one bank as an example:

> *"They appointed a number of people, both IT and domain experts, to experiment with the potential of a number of datasets. They started to explore the words that were used as additional information for money transfers. They discovered that this kind of data was not only interesting for themselves, but possibly – aggregated and anonymized – for other parties as well."*

If the experiment is successful, organizations can try to scale it up. These experiments on a smaller scale are important to make a business case that will convince management to make the initial investments. These investments will be even bigger when organizations insist on developing their own tools and acquiring the necessary big data talent themselves, although cheaper alternatives might be available.

3.5.4 Institutional and regulatory issues

While technical and organizational challenges are relatively concrete and - theoretically - possible to overcome, there are also a number of institutional and regulatory issues that may seriously hinder easy implementation of big data innovations.

Privacy and data protection

Privacy is an important barrier for organizations, especially for those that deal with consumers and personal data. In some cases organizations are explicitly aware of the limitations that data protection laws impose on their data-activities. The regulations determine how specific data can and cannot be used. Privacy is one of the main factors that the participants of the workshop mentioned for not sharing their data with others – not only in a regulatory sense but also because they feared a backlash from a PR perspective.

Trust and privacy also play an important role in the formation of value networks. In some cases the data is too sensitive and organizations want it to be stored on a local server or at least in the same country to make sure that it falls under the same jurisdiction. This is especially the case for medical and financial data. Teunissen mentioned during the interview the Utrecht Medical Center where data is in principle not even allowed to leave the campus. Exceptions can be made only under very strict conditions in terms of encryption and anonymization.

However, according to Wijsman and Wormer, although privacy regulation definitely has a big impact on the way organizations can use data, they feel that some organizations use it as an excuse to stay away from data and data analytics altogether, masking other technical or organizational issues, such as fear for organizational change.

3.5.5 Data quality

"Garbage in, garbage out" is a typical statement when quality of data is concerned. Although this statement is of course also true for smaller datasets, the issue seems particularly pressing when big datasets are concerned. Various interviewees mentioned the problems that arise in the value creation process when data is not recorded, stored or labeled in a proper way. When compromised datasets increase in volume and, in a later stage, enter the process of transport and analytics to be used for important decisions, it becomes increasingly difficult to determine initial flaws when they were originally collected. In that sense, data quality is a combination of technological, organizational and regulatory elements.

Discussing data quality, Teunissen and Kuijpers note:

> *"It is falsely assumed that acquiring technological tools is the solution to create value from big data. But most organizations simply do not know how to handle all their data. We see an increasingly important role for data stewardship."*

Good data quality and thus good data stewardship is not only important within organizations, but becomes even more critical when datasets are exchanged and combined with data from different stakeholders. Organizations have to rely on or trust other parties. In some cases, some stakeholders may not have the same interests when quality is concerned, and supplying incomplete or invalid data may actually be beneficial.

CIBAS is an example of an organization that fulfills an intermediary role in this respect. One of the main activities of CIBAS is the management and improvement of data quality: controlling the integrity and reliability of data that is supplied by all the stakeholders that are involved in the logistical process. The importance of CIBAS' role can for example be illustrated by the fact that the employees of the companies that supply the data, are mainly driven (and rewarded) by the speed of container collection. CIBAS: "For our partners, it is far less interesting to control whether or not all containers have been rightly registered by the vehicles. They have, and still largely are, judged by speed, not by data handling. This process is now slowly changing, as is the awareness that correct registrations eventually also benefit them."

Data quality becomes an especially important issue when data is used to make delicate decisions, for example when people are involved. An interesting example, pointed out by Katie Crawford from MIT, was the use of social media data in the aftermath of hurricane Sandy in New York. Although this was a very smart and innovative use to determine where and what kind of help was needed, it also had an important flaw: the data did not represent the whole picture:[122]

> *“The greatest number of tweets about Sandy came from Manhattan. This makes sense given the city's high level of smartphone ownership and Twitter use, but it creates the illusion that Manhattan was the hub of the disaster. Very few messages originated from more severely affected locations, such as Breezy Point, Coney Island and Rockaway. As extended power blackouts drained batteries and limited cellular access, even fewer tweets came from the worst hit areas. In fact, there was much more going on outside the privileged, urban experience of Sandy that Twitter data failed to convey, especially in aggregate.”*

As this example illustrates, there are substantial negative consequences when data analytics are applied based on incomplete or non-representative datasets. In other cases the original data collection tools – designed with a specific goal in mind - hamper the use for other goals. As the Port Authority of Rotterdam explained, the data from shipping companies that is collected by the financial departments, which focuses on the aggregate value of the content of cargo, does not require the same rigor in terms of detailed information of what that content actually is. This makes it less useful for the department that is responsible for business intelligence.

[122] Crawford, K. (2013). The Hidden Biases in Big Data. In: *HBR*. Available at: http://blogs.hbr.org/cs/2013/04/the_hidden_biases_in_big_data.html

4 Conclusions

The objective of the study was twofold. The first main objective was to describe the value creation process of big data and the associated process of data-driven innovation, including the most important drivers, goals and barriers. The second main objective of the study was to identify the core elements of the emerging big data ecosystem: the prominent types of players, the main products and services, the key technologies, and current and expected market dynamics. For the analyses, we draw on the results of a variety of expert- and stakeholder interviews, workshops and desk research.

The main findings and directions for further exploration are described along the following lines:

1 Differentiation and transformative data-driven innovation
2 The big data ecosystem and value at the fringes
3 Barriers, risks and potential losers

4.1 Differentiation and transformative data-driven innovations

Data-driven innovation comes in many shapes, ranging from incremental changes to more radical transformations.

Based on our exploration of the current make-up of the big data ecosystem, we see that the most common application of big data is an extension of more traditional business intelligence, primarily directed at improving internal organizational practices such as more efficient resource-allocation and production processes or more effective targeting of (potential) customers. These innovations can be lucrative, but the value they generate may very well be restricted to the organization and its shareholders, especially when the impact of increased efficiency does not spill over to other stakeholders or other sectors.

On the other end of the spectrum of big data-driven innovation, data occupies a truly prominent or even central role. In such cases, data constitutes a key ingredient of the product or service that is being produced. This “datafication” of products changes their nature and the way they are consumed. Illustrative examples are the new line of John Deere tractors, which are enhanced with sensors, and Nike’s Nike+ online platform and wearable FuelBand technology. These more transformative data-driven innovations enable new products and offer new value to consumers. They could even lead to systemic transformations (e.g., in agriculture and food, and lifestyle and health) that upend a whole ecosystem, as it forces other organizations to respond.

Despite their enormous potential, transformative data-driven innovations are very difficult to realize and the societal and economic benefits are difficult to assess beforehand. This notion was supported by every stakeholder and interviewee in this study. As the balance shifts from cost saving to new products and (trans-sectoral) systemic changes, the constellations that are required to establish these innovations become increasingly complex. Such innovations demand a higher level of big data maturity from organizations (in terms of technology and organizational

capabilities) and – as the number of stakeholders increases – they have to take different or maybe even conflicting interests into account.

As big data technologies become a commodity, more transformative innovations become increasingly important in terms of competitive differentiation. There is general consensus among all experts and stakeholders that the organizations that are able to use data to redefine and shape markets, sectors or ecosystems, are the ones most likely to take pole position and profit from it.

4.2 The big data ecosystem: value at the fringes

This study confirmed our notion that the big data ecosystem constitutes all parties that are involved in data-driven innovations and that impact the value creation process of data (see Figure 8), which is a simplified version of the value creation process of data which is discussed in more detail in paragraph 3.3). This means that it not only includes the data service providers and their technologies and products, but also the consumers of these products who drive demand and often provide domain knowledge, and also organizational experts, legal experts, data providers, policy makers, end-users and citizens.

The results of this study suggest that as the generic big data technologies are becoming a commodity, the competitive differentiation that big data offers will be realized in the 'first mile' of big data, on the left side of Figure 8, and the 'last mile' of big data[123], on the right side (both are marked green).

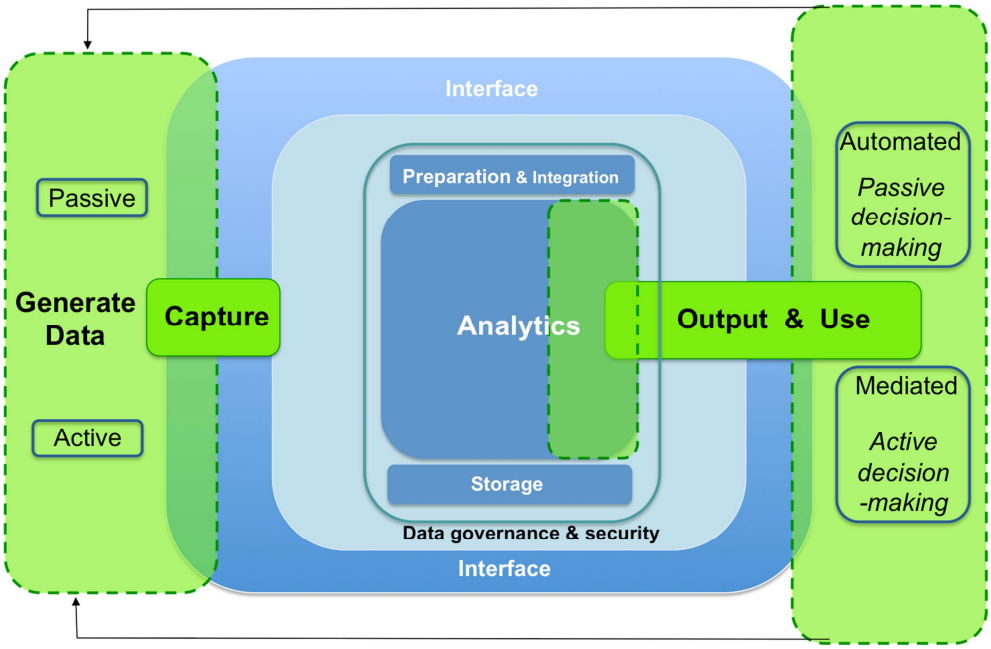

Figure 8 Value at the fringes (marked green) in the value creation process of data (TNO, 2013)

[123] ESG (2012). Boiling the Ocean of Control Points in the Hadoop Big Data Market. Available at: http://www.esg-global.com/blogs/boiling-the-ocean-of-control-points-in-the-hadoop-big-data-market/

The first mile of big data
When the overall level of big data maturity increases, organizations are able to shift their attention from incremental changes – that are generally based on already available internal datasets – to more radical innovations that often require datasets that are yet non-existent or reside in external databases. John Deere, for instance, includes external weather-data and it has also built a new data-collecting infrastructure (its tractors). Subsequently, access to the most relevant datasets becomes a crucial factor in innovation. This could lead to a more important role for data-marketplaces that actively look for new combinations between datasets and organizations. Such an intermediary role for data marketplaces might also entail that many organizations (and even individuals) that have no specific data-expertise, are no longer solely consumers, but also become key players in the big data ecosystem. They can share or sell their data to others.

Despite the potential for radical and transformative innovations, a vibrant, truly trans-sectoral data marketplace that facilitates the exchange or sale of datasets from organizations and individuals across sectors does not yet seem to be taking off on a grand scale.

The last mile of big data
The ‘last mile of big data’ constitutes the domain specific and context-specific analytics, applications and visualizations. These final steps in the value creation process of data enable organizations to create the best actionable output that is easily integrated in their business processes by humans, or machines.
Whereas the field of big data infrastructure providers is reaching some sort of equilibrium, dominated by a limited collection of (U.S. based) Hadoop distribution providers and traditional infrastructure vendors, the 'last mile' of big data is brimming with new start-ups, who are building on top of the infrastructure from the first category. Specialization is key in the last mile: it has to make big data domain-specific and usable in a certain context. This requires extensive vertical knowledge and the integration of idiosyncrasies of sectors and organizations.

Due to the diversity of data-driven innovations and the many disciplines required, collaboration in the big data landscape is essential. However, this does not withhold the biggest players such as IBM and Oracle from buying themselves into specific verticals by acquiring the most promising start-ups with a specific product or domain-expertise. This allows them to diversify their generic big data technologies and to control as many steps in the value creation process of data as possible.

4.3 Barriers, risks and potential losers

The most common definition of big data uses the so-called *three V's* to describe some of its characteristics: the Volume, Velocity and Variety. These characteristics put big data in a technological perspective, a problem that cannot be addressed by traditional tools and techniques. However, the nature of the challenge of big data is as much organizational as technological, especially because big data also constitutes a paradigm shift of datafication that leaves no sector untouched. The impact of these barriers should not be underestimated. The results of the workshops and interviews revealed that despite the abundance of raw data, many organizations are not yet able to leverage the potential of this data and data analytics.

We have identified a number of non-technological barriers that hinder the start and development of data-driven innovation.

- Lack of vision, creative talent and awareness regarding the value of data and the possibilities of data analytics.
- Lack of dynamic capabilities to respond to cultural and structural organizational changes.
- Institutional and regulatory barriers (i.e., European Data Protection framework)
- Data stewardship and data quality (i.e., integrity and representativity of the data)
- Governance, especially when data constellations expand and public values are at stake, and multiple stakeholders with potentially conflicting interests are involved.

On a more abstract level, the non-technological barriers come down to trust, or rather a lack of trust that hampers innovation: trust in the return on investments, trust in the quality and integrity of the datasets, trust in the algorithmic black boxes that process the data, trust in the benefits of the collaboration with machines and trust regarding the use of personal data and privacy, both in and outside an organization.

The issue of trust also reflects some of the most important risks that big data imposes. Privacy is an obvious major concern. Other risks arise from corrupted or incomplete datasets that do not represent what they are expected to represent. Furthermore, biased or wrong assumptions can be coded into black boxes that will provide us with skewed or even incorrect output. As these black boxes become more complex and act autonomously without human intervention, mistakes become harder to correct, as was the case in the financial flash-crash of 2010. When data-driven innovations concern systemic changes, risks reverberate throughout the constellations that expand as organizations build on top of each other's data, infrastructure and analytical tools. The question of responsibility and accountability will become increasingly important and complex in these kinds of partnerships.

The notion that the restructuring that follows the datafication of a certain domain will have winners also means that there will be losers. These losers might come from changes in production processes in case manual or knowledge-intensive tasks are outsourced to machines or otherwise made redundant. It is yet unclear what (and how many) current jobs will disappear, if new data-intensive jobs are able to make up for that loss, and whether the people who performed those jobs are able to upgrade their capabilities.

Other losers of the datafication are the organizations that are not able to adapt to a changing landscape. As data and the various steps in the value creation process become more and more important, the organizations that manage the data and fulfill these steps become more important as well. An interesting case is the development of the automotive sector in the US. Where Detroit was once the unchallenged epicenter of the American automotive industry it is now losing ground to Silicon Valley as data and data-driven innovation is defining the future of cars. It is yet unclear to what extent these kinds of transformations offer chances for new entrants or whether it is more likely that the incumbents are able to keep their dominant position. Either way, the influence of data service providers from the transcending big data ecosystem will increase as well, as the Detroit example illustrates.

Another issue, related to the restructuring of sectors due to datafication, is the war for talent and the control over data. It is likely that the scarcity of data scientists will make it hard for governments and non-commercial organizations to attract this kind of talent. Furthermore, whereas governments used to be the most prolific data-collectors, now – as data is considered to be a valuable asset to businesses as well – more and more datasets reside in databases from private companies outside the public domain. This means that especially in the case of more systemic changes where big data can have the most value for consumers/citizens and the society as a whole, governments will become dependent on these private parties for both data talent and data.

4.4 Concluding remarks

The data explosion and the emergence of affordable tools that process large, fast and varied datasets go hand-in-hand with a spreading paradigm shift that dictates the structural deployment of data. Still, for many organizations leveraging big data is still a major challenge for which they must overcome many barriers that go beyond technological hurdles.

But as big data matures – this study found - transformative data-driven innovations will become increasingly important in terms of competitive differentiation. Inevitably, this will lead to the restructuring of markets or even whole sectors. In order to profit from these changes, dynamic capabilities of both individuals and organizations are essential to thrive (or even survive) in the emerging data-driven society.

Considering the potential economic and societal impact of big data, policymakers should not only think about ways to facilitate data-driven innovation. It is also important to monitor to what extent its restructuring effects impact societal values such as privacy, entrepreneurship and right to play.

A Interviews

The duration of the interviews varied between one and two hours. Most interviews were recorded and were conducted by two of the authors. Interviewees were sent the outline of the interview beforehand. An interview protocol was created that covered all relevant aspects for answering the research questions of this study.

Data company interview Big Data

1 Please provide a short description of your company and your role within the organization.
2 Which step(s) from the value creation process is your organization involved in? [assist by showing and discussing the Value Creation Process / Ecosystem figure]
3 Which types of products or services do you deliver within this process?
4 What are the key technologies and tools you are using?
5 Are you using open standards? If so, under which conditions are they being used or can they be used?
6 Do you collaborate with other parties/stakeholders to provide the service/product you are delivering? What is their complementary value?
7 Are there other companies that offer the same kinds of product(s)/service(s) [name company]? If so, what are the most important competitors?
8 How does [company] differentiate itself from competing products/services?
9 What kinds of customers does [company] cater to with [product]? For what kinds of purposes do customers use the product/service? What are their drivers?
10 What are the most important barriers for customers of a product/service like [product] have to overcome (financially, organizational, technological, legal, …)
11 Looking at the value chain of data, roughly the following steps can be distinguished [show steps]. For which of these steps does [company] offer products/services?
12 Does [company] collaborate with other companies to offer Big Data solutions? If so, under what circumstances? With what (kinds of) companies?
13 What are the most important ambitions from [company] regarding Big Data services in general and [product]-like products/services specifically?

Case Interview Big Data

1 Please provide a short description of your company and your role within the organization.
2 To what extent are you systematically using data analytics? For which purpose/goal/end?
3 What were the most important diverse to start using data analytics?
4 Please provide a description of the process that led to the decision the apply and implement data analytics within your organization.
5 Which step(s) from the value creation process is your organization involved in? [assist by showing and discussing the Value Creation Process / Ecosystem figure]
6 Which types of products or services do you deliver within this process?
7 What are the key technologies and tools you are using?
8 What have been the most important barriers before you could systematically use data analytics?

9 What are the most important barriers you are faced with when using data analytics?
10 Could you give us some specific examples of big data analytics that you are using? In which departments of part of of your organizations are they using? What is the output of the analysis and how is this being used?
11 What is or has been the actual result of this big data application (e.g., efificiency, new products, ...)
12 Do you also produce data products or services (e.g., analysis of own data, consultancy) as part of your portfolio?
13 What types of data are you using for these analyses (e.g., historical/real-time, structured/unstructured, human/social/sensor)
14 What is the source of these data? Generated in-house of retrieved from third parties?
15 If third parties are the source for data, how is this data payed for? (e.g., licenced?)
16 Which steps in the value creation process do you fulfil yourself and for which steps do you rely on or collaborate with third parties (for transport, storage, management, analysis, visualization, security, etc.)
17 Regarding parties you collaborate with: which parties? And why these parties? Did you consider other parties as well?
18 What are your short-term and long-term ambitions for your organization and how does data analytics play are role?

Expert interviews

1 We call the process of transforming raw data to useful information and insights, the value creation process of data. [show figure]. To which extent do you agree with the proposed visualization, do you think it represents reality? Which steps or elements are you missing? Other suggestions?
2 What are the main reasons for organizations to start using data (products, services, analytics)? (e.g., personal hobbies/interests, competition, inspirational examples, external pressure, ...)
3 What are the main barriers for organizations as regards the application of big data (analytics) within their organization? (e.g., money, privacy, unclear or long-term benefits, ..)
4 What are the main reasons organizations are not able to *successfully* use big data? (lack of integration, alignment with strategy, ad-hoc, no support within organization; lack of support key decision makers)
5 What are the main barriers that producers of data products and data services (storage, analytics, etc) are facing when developing their products and bringing these to the market? (e.g., costs, legal, privacy, consumer demand, ..). Difference between smaller and larger organizations?
6 What are the types of products and services that are offered within the big data ecosystem? [use VCP/ecosystem figure)
7 Looking at the various products in the value creation process, what are the main companies at the moment? (names; regions/national/international; technological/non-technological). What are the most dominant players?
8 What are the means of payment that are common? (buy/rent, contracts, licences, trading, advertisement)
9 Do you see significant differences between the value of different types of data? (e.g, sensor data, analytics data, social media data, ...)

10 What are the main factors or characteristics that define the value of data? (completeness, availability, frequency, scarcity, ...)
11 Some questions about vertical integration in the ecosystem, i.e., organizations that fulfil more than 1 function/role in the ecosystem (like Apple does).
    11.1 Is this a common trend in the current ecosystem?
    11.2 Is it increasing/decreasing, what do you expect will be the trend?
    11.3 Could you give some examples of (type of) organizations or sectors where vertical integration is happening?
12 Some questions about horizontal integration (same service, multiple sectors)
    12.1 Is this a common trend in the current ecosystem?
    12.2 Is it increasing/decreasing, what do you expect will be the trend?
    12.3 Could you give some examples of (type of) products (e.g, storage, analytics, legal, cloud, ...) where horizontal integration is happening?
13 In the data ecosystem organizations can both be suppliers and buyer/consumer of data-related products (e.g., a company that buys storage and analytics technology, but also sells its own data about its production process or customer information).
    13.1 Is this a common trend in the current ecosystem?
    13.2 Is it increasing/decreasing, what do you expect will be the trend?
    13.3 What types of companies are doing this, and which types of products do they buy/supply?

B List of workshop participants

List of workshop participants with data-owners and data-experts.

Name	**Organisation**
Mr. Joris Binkhorst	Delta Lloyd
Mrs. Anne Verhage	KLPD
Mr. Ben in 't Veld	Belastingdienst
Mr. Mark Abspoel	Qurrent
Mr. Edwin Postma	Eneco
Mr. Thomas de Groen	Eneco
Mr. Bram Munnik	9292
Mrs. Frederika Wella Donker	TU Delft
Mr. Machiel Jansen	SURFsara
Mrs. Chiara Spaltro	SURFsara
Mr. Tom Demeyer	Waag Society
Mr. Mark van der Net	OScity / Cloud Collective
Mr. Reind van Olst	2coolmonkeys
Mr. Koen Havlik	Algoritmica
Mr. Tim Salimans	Algoritmica
Mr. Joost de Wit	TNO
Mr. Alwin Sixma	Big Data Value Center (Almere)

C List of companies in the Hadoop ecosystem

List of companies with the most partnerships in the Hadoop ecosystem.