MASTER THESIS

EFFECTS OF BIG DATA ANALYTICS ON ORGANIZATIONS’ VALUE CREATION

NIELS MOUTHAAAN
Student No. 6377289, MSc

University of Amsterdam
Faculty of Science and
Faculty of Economics and Business

Information Studies: Business Information Systems

UNIVERSITEIT VAN AMSTERDAM

Final version: August 22th, 2012

Supervisor: Arjan Vreeken
Signature: ..............................................................

Second examiner: .....................................................
Signature: ..............................................................
Executive summary

This thesis examines the effects of big data analytics on organizations’ value creation. Identified as the biggest IT trend for 2012, big data gets global attention and can be best described using the three Vs: volume, variety and velocity. It provides a comprehensive definition of the evolution of data. The former, volume, is the main cause of the “big data” phenomenon and describes the explosive growth of data: in 2007, more data was created than storage capacity could handle. The variety aspect indicates the diversity of contemporary data. An increasing 80% of data organizations own, can be classified as unstructured data: for example data packed in emails, social media and multimedia such as video and audio. Traditional data analytics tend to analyze structured data only: the remaining 20%. Hence, a potential source of valuable insights is ignored. Finally, also velocity plays an important role: more data is being generated in a shorter timeframe and analysis tasks should be able to keep up as the business value of an action decreases, the longer it takes from the occurrence of the event to taking action.

As data is becoming more extreme and will continue to do so, new data-warehousing and database technologies have been introduced to address this problem. Cloud computing plays an important role as it gives organizations the ability to analyze evolutionized data often in a more economically way as it offers computing resources on demand. The introduction of new NoSQL databases is a consequence of the increasing amount of unstructured data as it is capable of storing data having an unknown or dynamic structure. Also, it delivers the ability to scale more easily and efficiently in a horizontal way, by adding relatively cheap commodity nodes. The same applies to Hadoop, often identified as a synonym for big data due to its capabilities to store and handle huge amounts of data in an economically way by effectively making use of parallel and distributed computing.

Big data analytics can be seen as both the successor and enhancer of traditional data analytics, yet in current literature, its definition is still missing concreteness. In this thesis a more specific definition is provided, arguing that big data analytics focusses on analysis tasks requiring the need of distributed computing power or an in–memory database, or at least one of the used data sources uses data which can be identified as unstructured or multi-structured data, or no fixed or predefined connection between datasets exist.

In order to define big data analytics in terms of value creation, different theories on value creation are used to describe how value can be created when an organization is successfully performing its value activities. According to the value chain analysis, value is created when value activities can be performed at lower cost or result in more sales. To achieve this, according to Schumpeterian’s innovation theory and resource–based view, resources can be combined to create new products, services, production methods, etc. Also partnering through a strategic network, excluding other organizations from utilizing the same advantages, can create situations in which the buyer is willing to spend more of its own resources. Finally, by improving transaction efficiency, costs are reduced resulting in added value as the value activity can be performed at lower cost, either for the seller, buyer or both. By linking these theories with respect to the topic of this thesis, aggregated hypotheses are defined. Each hypothesis indicates how an organization might create value when succeeding in doing big data analytics. The first hypothesis argues that value can be created when big data analytics improves transaction efficiency leading to a reduction of costs or an increase in sales. The second hypothesis argues that by utilizing
an exclusive set of resources through a strategic network, making a specific big data analytics case possible, the amount buyers are willing to spend, increases. The third hypothesis, argues that big data analytics creates, supports the creation or improves products, services, markets or channels, leading to significant advantages for customers.

In this study, two cases have been studied: X FACTOR Heartbeat, an application showing live statistics about the participants of the music show X FACTOR and bol.com, an Internet-based retailer. Analyzing Twitter’s multi-structured streaming feed is what made the former case a big data analytics case and indicates that big data analytics is not only about volume. The sentiment of each tweet was calculated using list of words. In this case, big data analytics acted as a value creation driver as it offered significant advantages for the customer and was a result of the combination of different innovations seen as resources by both the Schumpeterian’s innovation theory and the resource-based view. The latter case represents a case in which the amount of data determines the label of being a big data analytics case as distributed computational power is required. By analyzing clickstream data of all visitors of their website, a list with search suggestion is created every night. These search suggestions are showed when a visitor starts typing in a search query and helps the visitor finding products. This case study showed that big data analytics acts as a value creation driver as it improves transaction efficiency of the transaction between bol.com and its customers by improving the search activity within the shop. It also confirms that big data analytics supports the creation of new or improved services, with significant advantages for customers.

Concluding, this study supports the first and third hypothesis, by arguing that big data analytics might create value in two ways: by improving transaction efficiency and by supporting innovation, leading to new or improved products and services.

For organizations, this thesis can act as an eye-opener as it shows the potential of big data analytics in relation to traditional data analytics. It also introduces different techniques which can be used to analyze different data sources. More importantly, it helps organizations defining expectations on how big data analytics will affect their organization, especially their value chain. Also, this thesis can be helpful for further research as it helps solving the “subjectiveness”-issue of big data analytics.

Finally, this study also shows that big data analytics is indeed a hype created by both potential users and suppliers and that many organizations are still experimenting with its implications as it is a new and relatively unexplored topic, both in scientific and organizational fields.

Additional research might focus on specific markets or types of big data analytics such as realtime analytics, which receives an increasing amount of interest. In addition, its impact on competitive advantage and the creation or reinvention of new business models, might be studied too, as value creation and competitive advantage are closely related. In addition, since this study fails to support the second hypothesis, the impact of big data analytics on competitive advantage can be considered doubtful and interesting to study.

Keywords: big data, data analytics, distributed analysis, value creation
Acknowledgements

This thesis has been written in order to finish my master “Information Studies” track “Business Information Systems” at the University of Amsterdam. I have always been interested in technical matters and my goal is to make innovations, especially within the IT field, more tangible by translating these innovation to business opportunities. I believe the phenomenon “big data” is such an innovation that in fact is not just one innovation but rather, a sum of innovations that even may create a new era in IT (Bloem, Doorn, Duivestein, Manen, & Ommeren, 2012).

During the last six months I have spoken with many people about the subject of this thesis. Naming this people would create an endless list of acknowledgments. In special, I would like to thank the members of Almere Data Capital, especially Hans Wormer, for letting me participate in the formation of a new data-mining master study and Brent Meulenberg, for helping me to connect with other people affiliated with the subject of this thesis. Also thanks to members of the Hogeschool Rotterdam for giving me the opportunity to give a presentation about my study and its progress. Special thanks to the organizations of Big Data Forum 2012 and Big Data seminar for their inspiring conferences that have been of great value to me.

Special thanks to all interviewees. Without them, getting sufficient data was just impossible: Harrie Vollaard (Rabobank), Marcel Kuil (Rabobank), Niels Basjes (bol.com), Jan Willem Mathijsen (bol.com), Frank Lippes (Triple IT), Gijs Zonneveld (Lost Boys Mobile), Ruud Burgel (Red Data)

I would like to thank Arjan Vreeken for his help during the last six months, but most of all for his inspiring and sometimes out-of-the-box lectures the last two years.

Marco, Michael and Matthijs, without you this study would have felt a lot longer but most of all, less interesting. I will never forget the great moments we have experienced: from having lunch in Vanessa’s class, visiting Budapest and experiencing Amsterdam’s nightlife in a total different way. Thanks guys!

Last but not least I would like to thank my girlfriend Shirley for keeping me motivated when I needed to. Also she did great work reading and checking this thesis.
# Table of contents

1. Definitions & abbreviations .......................................................... 6
   1.1. From data to wisdom ......................................................... 6
   1.2. Other definitions .......................................................... 7
   1.3. Abbreviations .............................................................. 8

2. Introduction & motivation ......................................................... 9

3. Research approach ................................................................. 10
   3.1. Research goals & relevance ............................................. 10
   3.2. Research questions ...................................................... 11
   3.3. Research method ........................................................ 12
   3.4. Reliability & validity ................................................... 14

4. Big data analytics ................................................................. 16
   4.1. Introduction ............................................................... 16
   4.2. The evolution of data ................................................... 17
   4.3. A new generation of data-warehousing and analytics .......... 24
   4.4. Summary .................................................................... 29

5. A potential value creation driver .............................................. 31
   5.1. Value creation theories ................................................ 31
   5.2. Linking value creation theories ..................................... 33
   5.3. Value creation hypotheses ............................................ 35
   5.4. Potential value creation driver ..................................... 36

6. A true value creation driver .................................................. 41
   6.1. Case I: X FACTOR Heartbeat ....................................... 41
   6.2. Case II: bol.com’s search suggestions ......................... 45

7. Conclusions & beyond ............................................................ 50
   7.1. Research conclusions ................................................... 50
   7.2. Achieved goals and relevance ...................................... 51
   7.3. Additional observation notes ....................................... 52
   7.4. Limitations ................................................................. 53
   7.5. Further research ........................................................ 54

8. Bibliography ........................................................................ 56
1. Definitions & abbreviations

This chapter starts with an important subject as this study studies the impact of data analysis: the relation between, data, information, knowledge and beyond. Furthermore, other definitions and abbreviations are explained, acting as an useful reference when reading this thesis.

1.1. From data to wisdom

Ackoff, a systems theorist and professor of organizational change, classified the human mind into the following five categories: data, information, knowledge, understanding and wisdom (Bellinger, Castro, & Mills, 2004). These categories are very useful in determining the different states of the interpretation of data. Each category has a certain amount of benefit for humanity as illustrated in figure 1. Data refers to raw symbols: it simply exists and does not have a meaning without interpretation of a human. Databases and files hold data, which is hard to see by human beings since “seeing something” always invokes some form of interpretation which makes it information: data that has been given meaning by relational connections acquired in the past. Information is not always useful, which brings us to the next category: knowledge. Knowledge is the appropriate collection of information which becomes knowledge when a human starts remembering this particular “piece” of information (Bellinger et al., 2004). For example, memorizing “2 x 2 = 4” gives you the ability to answer the question “what is 2 x 2?” in the future but does not give you the ability to answer the question “what is 323 x 234?”.

Creating new knowledge from the previously acquired knowledge is called “understanding”: the fourth category, which is a special case since in each transition between the categories, some form of understanding, a cognitive and analytical process, is present: it is not separate state of its own. Most useful for humanity is wisdom: the only category that deals with the future because it incorporates creativity, vision and design (Bellinger et al., 2004). Bellinger et al. (2004) argues that computers cannot create wisdom since a soul is needed which computers do not possess. This statement is very interesting for big data analytics since in the media, big data is described as a technology which, in the future, enables organizations to get answers without having questions (Leeuwen, n.d.; Ras, 2012, p. 17). Yet, this requires some form of creativity, vision and design.
1.2. Other definitions

Below an overview of the most-used definitions are listed which can be used as a quick reference when reading this thesis.

**Analytics**
Information resulting from the systematic analysis of data or statistics\(^1\).

**Big data**
Data that is complex in terms of volume, variety, velocity and/or its relation to other data, which makes it hard to handle using traditional database management or tools.

**Big data analytics**
Refers to analysis techniques operated on data sets classified as “big data”.

**Cloud**
The presence of IT services such as computing power and storage as a service accessible via a network such as the Internet.

**Data governance**
A technique to manage data within an organization efficiently and effectively (“Je moet met geweld de board in!,” 2012)

**Hadoop**
An open-source analytics toolset that supports running data-intensive applications on different nodes.

**MapReduce**
A model, mostly known as a part of Hadoop, used to distribute the processing of a large dataset across different nodes by using map and reduce jobs.

**Multi-structured**
Since data is often somehow structured, the term unstructured is misleading in these cases. Therefore multi-structured is a better term, referring to content not having a fixed structure. Terms like semi-structured and “grey data” are also referring to this.

**Node**
A node refers to a (virtual) terminal (or computer machine) in a computer network.

---

1 Retrieved from the online version of Oxford dictionary on June, 7th 2012 from http://oxforddictionaries.com/definition/analytics?q=analytics
1.3. Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACID</td>
<td>Atomicity, consistency, isolation and durability: a set of properties guaranteeing basic functionalities of most databases.</td>
</tr>
<tr>
<td>API</td>
<td>Application Programming Interface: a programmed specification to enable easy communication with other software components.</td>
</tr>
<tr>
<td>BASE</td>
<td>Basic availability, soft–state and eventually consistency: a successor and looser version of ACID making horizontal scaling more feasible.</td>
</tr>
<tr>
<td>BI</td>
<td>Business Intelligence: analyzing and combining data in order to create knowledge which helps the organization to create and exploit opportunities.</td>
</tr>
<tr>
<td>CRM</td>
<td>Customer Relationship Management: managing organization's interactions with customers, clients and sales prospects.</td>
</tr>
<tr>
<td>EDW</td>
<td>Enterprise Data Warehouse: a centralized database used for reporting and analysis.</td>
</tr>
<tr>
<td>HDFS</td>
<td>Hadoop Distributed File System: part of the Hadoop toolset making distributed storage of data possible.</td>
</tr>
<tr>
<td>IoT</td>
<td>Internet of Things: the phenomenon of connecting devices to a global network such as the Internet resulting in a interconnected digitally world.</td>
</tr>
<tr>
<td>IMDB</td>
<td>In–memory database: a database management system that primarily relies on main memory (e.g. RAM) to execute processing tasks at very high speed.</td>
</tr>
<tr>
<td>JSON</td>
<td>Javascript Object Notation: mostly used to exchange data between web applications.</td>
</tr>
<tr>
<td>MIS</td>
<td>Management Information System: provides information needed to manage an organization efficiently and effectively. Examples are enterprise resources planning (ERP) and customer relationship management (CRM) systems.</td>
</tr>
<tr>
<td>MPP</td>
<td>Massively Parallel Processing: processing tasks by using different nodes (distributed computing).</td>
</tr>
<tr>
<td>NoSQL</td>
<td>Not only SQL: a new generation of horizontal scalable databases often compliant to the BASE ruleset and often capable to handle unstructured and multi–structured data.</td>
</tr>
<tr>
<td>SME</td>
<td>Small and medium–sized enterprises.</td>
</tr>
</tbody>
</table>
2. Introduction & motivation

According to IBM the amount of unstructured and multi-structured data within an average organization is about 80% (Savvas, 2011). Taking account the average data growth, annually by 59% (Pettey & Goasduff, 2011), this percentage will likely be much higher in a few years. Not only the volume of data is becoming a problem, also the variety and velocity are issues we need to look at (Russom, 2011). This phenomenon is called “big data” and is identified as one of the biggest IT trends for 2012 (Pettey, 2012). In the last decade people start publishing content on the Internet resulting in the presence of much more data. Also smartphones and other devices, often equipped with different sensors and connected to a network such as the Internet, generates a dizzy amount of unstructured or multi-structured data. More than ever this data is made public by many organizations, free to use for anyone interested in the data (e.g. open data), by providing API's and other distribution methods.

Data is the new oil, and like oil, data needs to be refined before its get value. Using big data analytics, organizations can extract usable information out of enormous, complex, interconnected and varied datasets leading to valuable insights (such as the state of the business and customer behavior). In 2009 only a handful of big data initiatives existed and total industry revenues were under $100 million (Seguine, n.d.). Deloitte estimates that at the end of 2012, more than 90% of the Fortune500 will likely have at least some big data initiatives leading to an increase of industry revenues up to $1–$1.5 billion (Lee & Steward, 2012). Venture capitalists are subsidizing big data projects with funding rounds of over $50 million (Berman, n.d.) while governments are subsidizing open data and big data projects (Harris, 2012). It indeed appears that data is a source for value.

This thesis is structured as follows. The content of my thesis starts with an explanation of my research approach and design in chapter 3. Chapter 4 introduces the term “big data” by identifying properties of big data and big data analytics and relates it to traditional data analytics. Chapter 5 explains how organizations can create value. This leads to a set of hypotheses indicating how big data analytics might act as a value creation driver. In the same chapter existing literature and reports containing information about existing big data analytics implementations are described, each examined in the light of at least one of the hypotheses. Continuing, in chapter 6, different cases are examined giving the possibility to test the hypotheses. Finally in chapter 7, I unfold my conclusions in which I answer the main question followed by extra relevant discussion material as a result of the observations. Also I describe what subjects can be initiated for further research.
3. Research approach

In this section the research design of this study is described. It consists of the research goals and its relevance followed by the definition of the research questions. Section 3.3 describes the research itself and how it is done. Finally, in section 3.4, limitations are provided which were defined before the study was conducted. These limitations are used in section 7.4 to look back and reflect on their impact.

3.1. Research goals & relevance

Research goals consists of personal goals, intellectual goals and practical goals (Maxwell, 2005). This section describes the personal goals (goals that motivates the researcher to do a study which are not necessarily important for others), intellectual goals (goals focussed on understanding something) and practical goals (goals focussed on accomplishing something) of this study.

3.1.1. Personal goals

After working a few years in the IT field, I think the value of data is underestimated, especially in smaller organizations. This is probably due to a lack of time and resources. I believe organizations can create more value by leveraging data they have and create, but due to a lack of knowledge, this is not happening on a large scale. But I believe this is changing due to certain developments in different fields, both technological and non-technological. Data is changing and new technologies make it possible to analyze this data easier, faster and cheaper. Hence, it becomes easier to extract valuable insights from data, even for smaller organizations not having significant knowledge in IT. But maybe more important is that topics such as big data and data governance increase in popularity in board rooms (Cotte, n.d.). My personal goal is to get knowledge about the state of affairs in the “data as an asset” topic. How valuable is data really, especially data classified as big data?

3.1.2. Intellectual goals

Is big data a marketing hype, a buzz word, or is it real in terms of value creation? If so, how do organizations create value using big data analytics? These are the main questions this study tries to answer. By doing so it contributes to existing literature, which in fact, is relatively scarce. This will be valuable for both scientific fields (e.g. computer science and information science) and organizations (e.g. determining if they should invest in big data analytics). A recent article showed that existing definitions of big data (analytics) miss an importing thing: the way it creates value for organizations:

“If you read a dozen articles on big data, there’s a good chance the 3 Vs [...] will be cited in at least half of them. [...] The V model is a way to normalize, within bounds, an industrywide discourse, but it does not add much value [...] if you are trying to figure out how your organization should approach big data.” (Swoyer, 2012)
Burgel (2012) also argues this by saying that existing definitions lack the implication-aspect of big data analytics. The three Vs gives a comprehensive description of the evolution of data as can be read in section 4.2 but simply misses the “consequence” part of big data analytics: how does it affect an organization? This study tries to fill that gap by showing connections between big data analytics and organizational value creation, something which has not been done before.

### 3.1.3. Practical goals

The main goal of this study is to show how organizations can benefit from big data analytics in terms of value creation. By introducing big data analytics, where it originates from, describing the relevant techniques and more important: how value is created, a good starting point is provided giving all sorts of organizations a head-start in leveraging (their) big data.

### 3.2. Research questions

The main research question of this study is:

> Which ways of value creation can we identify when an organization analyzes data referred as big data and hence, is succeeding in doing big data analytics?

In order to answer the main research question, three sub-research questions are formulated:

> What characterizes big data and how does big data analytics differ from traditional data analytics?

In order to distinguish between big data analytics and traditional data analytics, a clear definition of big data and big data analytics must be given. This definition is also needed in order to correctly identify whether big data analytics is happening within each case.

> What causes and affects value creation within organizations?

To answer this sub-research question different theories on value creation are examined leading to a clear overview about how value can be created within organizations (and what can lead to value creation).

> Which impact does big data analytics might have on value creation of an organization?

By analyzing existing literature about big data analytics, exploring different existing and fictional cases and linking these findings with the theories on value creation, the possible effect of big data analytics on organizational value creation can be identified. This
knowledge is used to examine two cases leading to the answer of the main research question.

3.3. Research method

3.3.1. Research type

The term “big data” is relatively new and besides that, the meaning of big data is subjective and unclear. In a study conducted by The Data Warehousing Institute (TDWI) in 2011 most participants were familiar with something resembling big data analytics. Yet, only 18% used the term “big data analytics” for this (Russom, 2011). In other words: what is called “big data analytics” in this study, will probably have other names within different organizations. In fact, many organizations may even be unaware of the fact that they are “doing” big data analytics. This is the main reason for doing a qualitative study: cases needed to be carefully examined in their context by myself in order to generalize it with other data. Hence, a case study is conducted: an in-depth study of a specific instance (or a small number of instances) within a real-life context (Lazar, Feng, & Hochheiser, 2009). A case study knows three different characteristics. Each characteristic will be described in terms of this study in the following paragraphs.

In-depth investigation of a small number of cases

Because the limited amount of time and resources, only a few cases are examined. For the sake of this study, it is better to examine a few cases more deeply than superficial study dozens of organizations. This is because identifying value creation drivers and underlying factors requires a more in-depth study. Another reason is the lack of known cases in The Netherlands. as big data analytics is relatively new and unexplored.

Examination in context

As will be described in chapter 5, value creation is a subjective topic. Adding the vagueness of the definition of big data, each organization has its own thoughts about how big data analytics will, in the end, add value to their organization. Codifying, categorizing and defining these ways of value creation is very context depending thus more suitable for qualitative research.

Multiple data sources

Due to the impact of most big data implementations (both on technical and organizational level) documents and other forms of information about these initiatives exists. These data sources may include information about goals, targets and other aspects and hence are valuable for this research. More than any other type of research, case studies rely upon multiple data sources (Lazar et al., 2009).

3.3.2. Emphasis on qualitative data and analysis

The main goal of doing a case study is to get a deeper understanding of a certain phenomenon. More than research based solely on quantitative methods, conducting a case study is a technique helpful to describe or explain behavior (Yin, 2003). Although some value creation theories solely focus on quantitative ways of defining value (such as value chain analysis), other theories are more context related (such as strategic networks theory) and hence, are more suitable for qualitative research methods. Case studies can include both research methods as Lazar et al. (2009) mentions: “Case studies can
certainly include quantitative components measuring traditional metrics [...] but these measures are not usually the sole focus of the investigation” (Lazar et al., 2009).

### 3.3.3. Research framework

Figure 2 gives an overview of my study and this thesis. Four main activities can be extracted from this overview: defining big data analytics (A1), defining organizational value creation (A2), connecting this with the definition of big data analytics resulting in a set of hypotheses (A3) and examining different cases to test these hypotheses (A4). These activities are strongly connected to the research questions previously described.

![Figure 2: Research framework](image)

### 3.3.4. Case selection

In this study a broader understanding of the topic of this thesis is developed. Therefor this case study is called an instrumental case study (Lazar et al., 2009). Some criteria used to select cases is developed. This criteria is based on the fact that this study studies the effects of big data analytics on organizations’ value creation. As a result, this criteria consisted of two rules:

- The case needed to be a “big data” case, and;
- The case already needed to be integrated in the organization in a functional way;

The first rule assures that a case is indeed a “big data” case by using characteristics described in chapter 4. The second rule is important too: since big data analytics is relatively new and unexplored, many organizations are still orienting and discovering its possibilities and impact on their organization. Since identifying the ways of value creation when succeeding in doing big data analytics is the main goal, it was important to find cases that already proofed its value in the organization. In addition to above criteria, all
cases need to be located in the Netherlands as it creates the possibility to conduct personal interviews making data collection easier and more reliable.

### 3.3.5. Data collection & analysis

The cases are examined by having at least one interview with responsible managers and/or involved employees. The interviews are semi-structured, meaning that there is a list with topics (based on findings described in chapter 4 and 5) which are particular important, but closed questions, not leaving any space for alternative observations, are avoided. The interviews are conducted according to the following predefined self–made protocol:

1. Arrange an informal introduction by email, telephone or real–life conversation;
2. Invite interviewees with a formal introduction on the subject, study and goal of the interview by email;
3. Write notes during the interview. Also, after asking permission of the interviewees, audio record the interview;
4. Work out the interview (in the language of the interviewees) in form of a report using the audio recording and written notes;
5. Ask the interviewees to review the report;
6. Adjust the interview with the feedback of the interviewees.

Although interviews were the main data collection method, also data from other sources (e.g. presentations and conferences) are used. Since each case has its own specific data sources, these data sources will be discussed individually in their corresponding section. Data is codified, which enables the process of finding connections between observations. The content analysis technique is used for this and is defined as:

> “Any technique for making inferences by objectively and systematically identifying specified characteristics of messages […] and not only applies to textual information, but also to multimedia material.” (Lazar et al., 2009)

By using existing theories about big data analytics and value creation, the “priori coding” approach can be used which uses predefined coding categories based on existing theories. An example of such a category is the technology used (e.g. Hadoop). Also data is codified in groups, which makes it easier to relate observations with each other.

### 3.4. Reliability & validity

The validity of this study can be influenced by different factors, each being described in the following paragraphs.

#### 3.4.1. Invalid/ill-conditioned definitions leading to codify errors

Big data is loosely–defined term and prone to context and understanding issues. Without a clear definition of this term (and the phenomenon itself), identifying whether big data analytics happens within each case will be impossible without risking codify errors. To
overcome this issue, I will define characteristics, making it easier to identify big data analytics, both for the interviewees as myself as the coder.

3.4.2. **Case selection**

Organizations must be willing to commit some of their own resources when being involved in a case study. Since cases will be studied by me as an outsider, organizations can be mysterious or taciturn about facts and other information.

3.4.3. **Limited to the Netherlands**

Only a limited number of cases will be studied leading to a possible incomplete answer of the main research question. Also, since only cases located in the Netherlands will be studied, the outcomes of this study might only represent the Dutch market, although I believe the results are replicable in other countries as well.
4. Big data analytics

4.1. Introduction

According to the Oxford dictionary, analytics is defined as “the systematic computational analysis of data or statistics”\(^2\) or, a better definition, also according to the Oxford dictionary: “information resulting from the systematic analysis of data or statistics”. Since information is the source for knowledge and even wisdom as described in chapter 1, analytics is very important in many different fields, both scientific and organizational, especially for decision making (Golfarelli, Rizzi, & Cella, 2004). For example, without analytics, the procurement department of a supermarket chain would have a hard time deciding what to buy in which numbers.

Many forms of traditional data analytics exist. Probably the best known is business intelligence, defined as:

“The process of turning data into information and then into knowledge. Business intelligence was born within the industrial world in the early 90’s to satisfy the managers’ request for efficiently analyzing the enterprise data in order to better understand the situation of their business and improving the decision process” (Golfarelli et al., 2004)

Many other forms of data analytics exist: from web traffic analysis to customer satisfaction analysis. The way it creates value differs for each form of data analytics but in the end, almost always try to achieve the same goal: increase revenues, reduce costs or both. For example, web traffic analysis can lead to insights which can be used to improve the usability of a website leading to more satisfied customers resulting in an increase of sales.

Continuing with big data analytics, you might ask what is different compared with traditional data analytics. First of all, big data analytics is not the successor of traditional data analytics. It can however support traditional data analytics, also mentioned by Adrian & Chamberlin (2012):

"Big data can enable richer reporting and analysis, complementing, rather than replacing, traditional data warehouses” (Adrian & Chamberlin, 2012)

Many definitions of big data can be found when reviewing the relatively scarce literature. This widespread of definitions in fact tells something about the current state of the global understanding of this theme. It appears that big data is an unclear term, probably because big data is not tangible nor a fixed property. Big data itself is not a disruptive innovation but a sustaining innovation making it harder to describe, define and distinguish big data (analytics) from traditional data (analytics). Sustaining innovations

\(^2\) Retrieved from the online version of Oxford dictionary on June, 7th 2012 from http://oxforddictionaries.com/definition/analytics?q=analytics
“are innovations that make a product or service perform better in ways that customers in the mainstream market already value” while disruptive innovations “create entirely new markets through the introduction of a new kind of product or service […]” (Christensen & Overdorf, 2000). What is now called “big data” is not the result of a specific innovation of a product or service but rather a combination of revolutionary and evolutionary changes and innovations in different fields (both technical and non-technical fields) (Gantz, 2011, p. 6). These events, including the relevant fields in which they occurred, will be discussed in this chapter leading to insights about what big data (analytics) is and how this phenomenon can be distinguished from traditional data (analytics) and hence, answer the first subquestion: “What characterizes big data and how does big data analytics differ from traditional data analytics?”.

4.2. The evolution of data

Introduced in 2001, “the three Vs”, (Laney, 2001) provides a comprehensive definition of big data and each of the three Vs has its own ramifications for analytics (Russom, 2011) which will be used to describe the evolution of the overall change of data we generate, collect and store and is responsible for the big data phenomenon. Change in terms of volume is most commonly used to define big data but gives an incomplete picture. Earlier studies have denoted a change in variety and velocity too.

4.2.1. Volume

The volume of data organizations and individuals own, grows rapidly. This growth can be quantitative measured by looking at the amount of bytes organizations collect and store, the number of records they have stored in their databases, the number of transactions (e.g. queries) being run or by counting the number of tables, files and other forms of data (Russom, 2011). In 2007, researchers estimated that the 13 most used groups of digital storage and 12 most used groups of analog storage, holds 295 exabytes\(^3\) of data (Hilbert & López, 2011). Back in 2000, this amount was 15.8 exabytes, indicating a major growth of the presence of data.

Also, in 2007, a study sponsored by information–management organization EMC and conducted by research–firm IDC, showed that the amount of digital data created in that year exceeded the world’s data storage capacity for the first time: more data was created and replicated than actually could be saved, as indicated in figure 3 (Gantz, Mcarthur, & Minton, 2007). Although it is not always necessary to store all produced data, the same data can hold valuable insights of the behavior of (potential) customers, markets and

---

\(^3\) 1 exabyte is equal to 1,000,000,000 gigabytes in this study
analyzing this data can extract these insights leading to potential value and exactly for this reason, big data analytics differs from traditional data analytics in terms of volume. With traditional data analytics, data to be analyzed is often carefully picked, sampled and hence, can be biased in comparison with the complete picture (the population) (Bloem et al., 2012, p. 6). By selecting pieces of data (e.g. rows in a database) or elements of data (e.g. columns in a database), the context, detail, relationship and identities of important (business) information can be destroyed leading to incomplete or even corrupted meaning of data (Blechard, Adrian, Friedman, Schulte, & Laney, 2011). Also, because data is carefully picked, meaning that data is highly structured and originates from a limited number of data sources, chances to discover new insights are small. Rather, historic events are illustrated and described in an easier way to read for human beings. This is how many traditional BI systems work: analyzing a limited number of transactional data (e.g. sales transactions for the last week) and illustrating this as a graph (or any other illustration method easy to interpret). Although such graphs often indicate what happened during a certain timespan, not often new insights will evolve, simply because the volume of data is limited.

So far only data at global level is discussed. More important for this study is the amount of data organizations create and replicate: a subset growing even faster than the digital universe as a whole (Gantz et al., 2007). Factors driving this growth are the computerization of small businesses, new privacy regulations, archiving and security policies and of course, the emerge of the Internet as distribution and communication channel. As an example, Walmart’s customer transaction database was reported to be 110 terabytes in 2000. By 2004 it increased to be over half a petabyte (Schuman, 2004) In 2009, McKinsey estimated that nearly all sectors in the US economy had at least an average of 200 terabytes of stored data per organization (for organizations with more than 1000 employees) (Manyika et al., 2011).

An important question needs to be answered: what is responsible for the data growth? This question is important since it might change the way how organizations analyze data in the future. It appears that there are a number of causes, starting with the
global increase of the number of Internet users as illustrated in figure 4. ITU (International Telecommunication Union) estimated that in 2011, 34.7% of the world population had access to the Internet while in 2007, only 17.5% had access. This growth is not only explicable due to the fact that developing countries provide (better) access to the Internet, also the number of Internet users in developed countries is raising significantly: from 53.5% in 2007 to 73.8% in 2011. This fact is important since particularly organizations based in developed countries have the ability to utilize big data. The increasing number of Internet users will directly influence the rise of the presence of data since more entries will be saved in log files of web servers and other coherent services (such as database servers). More important is the combination of the number of Internet users and what is called “web 2.0”. Although web 2.0 knows many definitions, it is commonly defined as the second generation of the Internet in which collaborative content creation and modification, reusing and combining of data and social networks take a central place (Murugesan, 2007). Examples are blogs, wikis and social networks. The main difference is the way how data is being provided: instead of an one-way-only web in which the owner of a website is the only content provider, now visitors have the ability to add, modify and reuse the content of the website resulting in a much richer Internet and an increase of the presence of data available online.

Not only an increasing number of human beings are connected to the Internet, also there is a significant increase in the number of physical devices connected to the Internet. This phenomenon is called the “Internet of things” (IoT):

“Sensors and tiny devices (actuators) embedded in physical objects—from roadways to pacemakers—are linked through wired and wireless networks, often using the same Internet Protocol (IP) that connects the Internet”. (Chui, Löffler, & Roberts, 2010)

In 2011, the GSMA (GSM Association) published a research report showing that the number of total connected devices (e.g. mobile phones, manufacturing sensors and modern cars) is expected to increase from approximately 9 million in 2011 to more than 24 billion in 2020 and within that, mobile connected devices will double from more than 6.6 billion (more than the world’s human population) today (Boogert, 2012) to 12 billion in 2020 (“GSMA Announces That the Proliferation of Connected Devices Will Create a US $1.2 Trillion Revenue Opportunity for Mobile Operators by 2020,” 2011). These devices
will include an increasing number of sensors creating raw data, including but not limited to: temperature, location, tilt, vibration, light, humidity and sound sensors. Kuil, innovation manager at the Rabobank, said that the growth of data is mainly caused by a growing number of devices customers use (Kuil, 2012). Also ATM’s are sending much more data, such as video, leading to an explosive growth of data.

Falling prices and increased performance of digital devices also help driving the data explosion. A well-known reliable indicator in technology is “Moore’s Law”, which states that the number of transistors on a chip doubles roughly every 18 months (Jovanovic & Rousseau, 2002). This basically means that the computational capacity of an electronic device doubles roughly every 18 months maintaining the same physical size and price. The same can be applied to storage capacity: magnetic areal storage density doubles roughly every 18 months. This phenomenon is known as the “Kryder’s Law” (Walter, 2005). Increasing storage capacity (which influences the selection of which data is being saved and which data is being discarded) and increasing computational power (which influences, for example, the choice which analyzing method is being used) per dollar spend, made all industries adopt data-intense applications (Gantz et al., 2007). Between 1997 and 2008 the Internet traffic doubled every 6 months, indicating that online data is being (re)produced and send in a much quicker way than technology is capable to handle maintaining the same number of (storage) devices (L. G. Roberts, 2000).

Besides the fact that users and organizations are generating and storing more data than ever before, more and more data is becoming accessible through (public) API’s and other data-sharing methods. ProgrammableWeb, a platform containing an extensive directory of public API’s, has reported a explosive growth in the number of API’s they have listed (DuVander, 2012). Another portal is the “open data” initiative run by the Dutch government. By sharing data using open standards, civilians and organizations have the ability to create useful applications (“Open standaarden, open source en open data,” n.d.). This leads to more data sources which may contain valuable information about customers and the market an organization is operating in.

Concluding, much more data is being generated, stored and made available for public use and due to various technical innovations this data can be analyzed more easily. This, in the end, leads to more knowledge and hence, valuable insights: “The more data that is created, the more knowledge [...] people can obtain” (Evans, 2011).

4.2.2. Variety

Although the increasing amount of data is the main driver for big data analytics, also the variety (also called diversity in some literature) of data being generated, stored and made public, plays an important role in describing the big data phenomenon. Data can be divided into two groups: structured and unstructured data. Traditional data analytics focusses mainly on structured data. Examples are most traditional (management) information systems (MIS’s) using data from enterprise data warehouses (EDWs) which often uses relational databases to hold data in a highly structured format, leaving out the possibility to store (and hence, analyze) data that does not fit in the predefined structure. Changing this structure is almost impossible without changing the dependable information systems and without risking dataloss. As a result, changing a MIS to include other forms of data (e.g. in another structure or even without knowing the structure) is (too) expensive and therefore this data is being ignored, leaving out the possibility to extract new knowledge and valuable insights. This problem, together with others, has led
to multiple initiatives to change the way digital data is stored, which will be discussed in section 4.3.2.

In order to store and analyze unstructured data, the need to dynamically add new attributes to support new forms of data without changing the structure of the database (Cattell, 2011) is increasing. Since storing unstructured data is becoming easier, analyzing this type of data will be more feasible too and this is important since, in 2003, more than 85 percent of all business information exists as unstructured data (Blumberg & Atre, 2003). In addition, in a recent study 80 percent of the respondents said that the amount of unstructured data within their organization will rise in the next three years but only 24 percent believe their current infrastructure will be able to manage it (Hayward, 2011). In most organizations, the amount of multi-structured data is growing at a considerably faster rate than structured data (C. White, 2011).

An important question needs to be answered: what is unstructured data? Simply said, unstructured data is all data that is not fully structured according to a pre-defined data model (e.g. relational tables). Since such a model is made by assuming a certain context and structure, data is always structured when it fully conforms expectations, whether these expectations are human or computer made.

"Unstructured information lacks defined data types and rules to enforce where that data is stored. In addition, unstructured data is created and generated across a broad cross-section of end users, and is stored in user-defined directories, beyond the reach of enterprise rules.” (McKendrick, 2011)

Unstructured data is often defined as multi-structured as data often has some form of structure (C. White, 2011). Take for example a Word document. Basic information such as the title and the name of the author are stored in meta data and hence, are structured since these fields conforms the expectations we have (apart from the fact whether this data is filled in correctly or not). However, the content itself (the document) is often unstructured: without analyzing the document, we are not able to find what we are looking for. It is even harder when you do not know what to look for, as with analytics is often the case, especially with big data analytics.

New forms of data sources such as web pages, social networks, instant messaging, email and other communication channels are, together with new media types such as audio, video and images, main responsible for the high amount of unstructured data (Manyika et al., 2011). IDC estimated in 2007 that one quarter of the data growth between 2006 and 2010 derived from cameras and camcorders (Gantz et al., 2007). Since more people worldwide will have access to cameras and camcorders connected to the Internet this portion is likely to grow. Increasing resolutions and camera capabilities (such as adding location meta data) are also responsible for this data growth since the amount of megabytes per image grow too.

And then there is video. In a study conduced by IDC in 2006, 77% of the digital camera users also had video recording capabilities (including 50% of the camera phones) (Gantz et al., 2007). Another great source of video is YouTube, indicating that in 2011 every minute more than 48 hours of video was being uploaded while two years before that, “only” 24 hours of video was being uploaded every minute (“Thanks, YouTube community, for two BIG gifts on our sixth birthday!,” 2012).
Also audio, taking up about 20% of the overall data being created in 2006 (Gantz et al., 2007), is responsible for the data explosion. Most help desks and call centers record audio conversation with (potential) customers for, among others, training purposes and can contain valuable pieces of information that possibly can lead to new insights.

Traditional data analytics tend to analyze textual and structured data only. The above observations indicate that this is in fact only a very small proportion of data being generated worldwide by both individuals and organizations. All other forms of data might contain valuable information too, depending on the industry: “financial services, administrative parts of government, and retail and wholesale all generate significant amounts of text and numerical data including customer data, transaction information [...] while other sectors such as manufacturing, health care and communications and media are responsible for higher percentages of multimedia data” (Manyika et al., 2011). For example, in the health care sector more than 95% of the clinical data is now video (Manyika et al., 2011).

Variety and volume tend to fuel each other (Russom, 2011) and until recently, analyzing unstructured forms of data was almost impossible within a reasonable time-span because of technical limitations. Recent developments within different fields have now made it possible to analyze varied datasets much faster which brings us to the third V: velocity.

4.2.3. Velocity

Insights and knowledge, and thus data are more and more becoming a indispensable asset for organizations in order to create competitive advantage and stay ahead of their competitors, as Teece (2003) also points out:

“The post-ware evolution of markets has powerful strategic implications for how and where firms position themselves to build competitive advantage. [...] It is no longer in product markets but in intangibles assets where advantage is built and defended.” (Teece, 2003)

Many possible opportunities lie in these intangible assets. As a business rule, the first who exploits the opportunity gets the lead and that is why extracting knowledge out of data as fast as possible is important. Taking account the massive growth of data (volume), mainly because of the rise of new data sources (variety), analyzing data as fast as possible is another challenge. Velocity is about two things: processing data taking account the speed in which it is created and the need to derive (business) value within a certain timespan from it, also mentioned by Gartner (Pettey & Goasduff, 2011). Real time data (also called stream data) focuses on what is happening right now and not on what has already happened, as it enables situational awareness (Blechar, Adrian, Friedman, Schulte, & Laney, 2011). A good example of real time data is clickstream data, being generated as visitors navigate through a website. Internet retailer bol.com analyzes clickstream data of all visitors to create a list with search suggestions as can be read in section 6.2. This clickstream data varies day to day so analyzing this data must be done as fast as possible: the list will be less valuable when it is a week old compared to a one-day-old list.
Two approaches of data analytics can be defined: the “analyze and store approach” and the “store and analyze” approach. The first approach means that data that flows through business processes, across networks and between systems, is being analyzed (almost) realtime while only important results are being stored (C. White, 2011). Analyzing data before storing it reduces latency. It also reduces storage, administration and security resources requirements, since important data is stored in a structured way while unnecessary data is being discarded. In order to process data before it is stored, it must be somehow part of the process where the data origins from. Two options are available: embedding the analytical processing in the business process or analyzing stream data as a separate process. Embedding the analytical processing in the business process “works well when implementing business process management and service-oriented technologies” (C. White, 2011) and is particularly useful for monitoring and analyzing business processes and activities in close to realtime. Also stream data can be analyzed as it flows across networks and between systems and hence, is a different process than the business process where the data origins from.

Traditional data analytics such as BI is a good example of the “store and analyze” approach. BI systems are often configured to use existing data sources (Seufert & Schiefer, 2005) in order to create reports to describe what happened during a certain period. This is valuable when there is a need to explain a particular event or for creating regular reports such as sales reports. Yet, although it can be a valuable source to indicate what is likely to happen in the future, it often does not. Therefore answering the question “what is likely going to happen tomorrow?” instead of “what is happened yesterday?” is more valuable and hence, is becoming more important. Also, the business value of an action decreases, the longer it takes from the occurrence of the event to taking action (as illustrated in figure 5) and therefore, (close to) realtime decision making is important in value creation (Hackathorn, 2002).

Concluding, big data can also be described by its velocity: the speed (frequency) in which it is being created and the need for realtime analytics to derive (business) value from it.

![Figure 5: Reducing action time leads to an increase of value](Hackathorn, 2002)

### 4.2.4. Complexity & connections

Analyzing traditional data often means analyzing data coming from structured and centralized data sources. While (complex) relationships with other data or data sources can exist, these relationships are often known and therefore relatively easy to include and analyze. With big data analytics, more exotic, unstructured and/or varied data sources are included, leading to relationships which are less obvious and hence more complex to analyze as Boyd & Crawford (2011) also mentions: “big data is notable not because of its size, but because of its relationally to other data”. According to Burgel (2012) this is an
important of big data analytics: new technologies have made it possible to go beyond internal data by using external data sources. A good example is linking a customer database with social media data sources such as Twitter: making relationships and preserving the context of what is being said is very challenging. Although this so called “linked data” often uses standardized ways of data transportation (e.g. HTTP), preserving context making it both usable for humans and computers, is difficult (Blechar et al., 2011).

4.3. A new generation of data-warehousing and analytics

Without the emerge of new data-warehousing and technologies, there would no big data phenomenon (Dijcks, 2012). Data will be more extreme in the future (e.g. in terms of the three Vs) and new techniques will be needed, making it possible to analyze this data. The last years the ability to store and analyze data in comparison with the amount of data that is being generated (as being illustrated in the previous section), lagged behind. New data-warehousing and database technologies have been introduced to address this problem. This section will elaborate on the developments in different (technological) fields making big data analytics possible.

4.3.1. The rise of the cloud

The rise of the cloud plays a significant role in big data analytics and likely this role will increase as the cloud is being adopted by a growing number of organizations (Armbrust et al., 2010). Cloud computing is an extremely successful paradigm of service oriented computing and provides services at different levels of IT, for example, Infrastructure as a Service (IaaS), Platform as a service (PaaS) and Software as a Service (SaaS) (D. Agrawal, Das, & Abbadi, 2010). Some advantages of cloud computing, compared to in-house computing, are:

- Infinite computing resources are available on demand;
- No up-front commitment by cloud users; users can start small but think big;
- Pay for use of resources on a short-term basis (e.g. more resources on peak hours);

These advantages are useful for big data analytics in several ways. In order to analyze data, there must be data available and as described earlier, data is being created in a much faster way than ever before. Therefore, a lot of storage space is necessary (especially with the “store and analyze” approach). A significant proportion of data organizations own, is created by end-users (such as visitors of the organization’s website) and hence, cannot be controlled by the organization itself. This indicates the need to easily demand more resources from the cloud provider when required.

Also, organizations tend to “start small but think big”. Vollaard, innovation manager at the Rabobank, mentioned that big data analytics is being rolled out in different phases trying to find out how big data analytics affects the organization and its processes (Vollaard, 2012). This principle can be found in many organizations as will be discussed in section 7.3.2. For this, the dynamics of the cloud is a big advantage: no expensive data centers needs to be purchased, installed and maintained leading to huge investment costs. Rather, necessary computing resources can be rent with the possibility to easily scale when needed. This is extremely beneficial for especially small
organizations since big data analytics is now also possible for SMEs as Agrawal (2012) also points out:

“A move to cloud with big data analytics as a service is not far-fetched as SMEs will look to eliminate costs from building in-house infrastructure to support big data analysis.” (T. Agrawal, 2012)

Gantz & Reinsel (2011, p. 2) also supports this. Another great advantage of cloud computing is the way how computing resources can be rented on very short-term basis, especially for analysis-related tasks. These tasks are often being done at specific moments. Even with realtime analytics, resources needed to analyze data streams often vary because the data stream is smaller at certain moments (e.g. factory machinery that’s running half-speed during the night). When this data is analyzed using an in-house data center, the data center needs to have as many resources as possible to fulfill the need when the data stream is at its maximum. Hence, when the data stream is smaller, many resources are unused leading to cost inefficiency. With cloud computing, the user generally pays for the resources he actually uses. This may lead to significant cost improvements, especially when data analytics happens only on specific moments (e.g. once per week). However, a calculation of costs should always be made as the opposite may be true. For example, bol.com is analyzing on such a high scale, renting capacity from a cloud provider would have been more expensive (Basjes & Mathijssen, 2012). Also privacy could be an issue, something which has been discussed often lately.

Amazon’s evolution of their database system will be used as an example to illustrate the possible benefits of big data analytics in the cloud and the need for similar systems. This database system became a well-known cloud service publicly available. “Amazon runs a world-wide e-commerce platform that serves tens of millions customers at peak times using tens of thousands of servers located in many data centers around the world” (DeCandia et al., 2007). The slightest outage has significant financial consequences and with so many servers world wide hardware fails continuously, highly decentralized storage technology was needed:

“Customers should be able to view and add items to their shopping cart even if disks are failing, network routes are flapping, or data centers are being destroyed by tornados. Therefore, the service responsible for managing shopping carts requires that it can always write to and read from its data store, and that its data needs to be available across multiple data centers.” (DeCandia et al., 2007)

Amazon has developed a number of online storage solutions, including Amazon S3 to store BLOB data (e.g. images). Another solution is Dynamo, a highly reliable NoSQL database technology partitioning and replicating data using consistent hashing over multiple servers. In order to be able to do this, only the primary key is indexed and hence, data can only be efficiently fetched using the primary key. In most cases, when using a proper data model, this is sufficient for most applications (e.g. a shopping cart). Traditional information systems often use relational databases to store data and besides
the scaling difficulties it has (as will be discussed later on), it provides applications with an excess of functionality, such as advanced querying and management functionality, making it an inefficient solution since it influences the performance of the system in a negative way. Since big data analytics also focuses on unstructured or multi-structured data and the amount of data is growing faster than the technology needed to store and analyze it, databases such as Dynamo will be used more often and will even replace relational databases in some cases, although it will not replace them completely (Leavitt, 2010).

In January 2012, Amazon released Dynamo as part of AWS (Amazon Web Services) under the name of DynamoDB. DynamoDB continues to build on the principles of Dynamo and also “builds on our years of experience with running non-relational databases and cloud services, such as Amazon SimpleDB and Amazon S3, at scale” (Vogels, 2012). Together with existing AWS services such as EC2 (computing power in the cloud) and Elastic MapReduce (Amazon’s implementation of Hadoop, which will be discussed in the next section), Amazon delivers a powerful big data analytics platform in the cloud.

Other well-known solutions providing decentralized, highly available and scalable (database) systems in the cloud are Google BigQuery, a service provided by Google making it possible to analyze very big tables and Windows Azure Tables, Microsoft’s NoSQL solution in the cloud.

4.3.2. The global introduction of NoSQL databases

New forms of databases have been developed, giving up at least one constraint of the ACID principle. ACID stands for atomicity (a transaction is “all or nothing”), consistency (the database will be in a consistent state before and after a transaction), isolation (transactions may not interfere with each other) and durability (a transaction is always permanent) (Haerder & Reuter, 1983). Since the amount of data is growing extremely fast compared with how technology evolves (e.g. Moore’s law and Kryder’s law) and the structure of data itself, scaling databases has become important. Since vertical scaling (e.g. moving the database to a more powerful system or increasing the capacity of the database system) is always limited to the fastest possible system available and relatively expensive, horizontal scaling (e.g. distributing the database, or its functions, across multiple nodes) is often preferred since relatively cheap commodity systems can be used and no physical limitations are in play. However, horizontal scaling databases has disadvantages comparable to many other distributed systems. First of all, not all constraints of ACID can be applied at the same time. A theory, known as the CAP theorem, says that if you want consistency, availability, and partition tolerance, you have to settle for two out of three (Browne, 2009). Partition tolerance refers to “no set of failures less than total network failure is allowed to cause the system to respond incorrectly” (Gilbert & Lynch, 2002). Since the ACID principle cannot longer be fully fulfilled when using a distributed database, a new alternative has been introduced, known as BASE. BASE stands for “Basic Availability”, “Soft-state” and “Eventual consistency” (Pritchett, 2008) indicating that, rather than requiring consistency after every transaction, it is enough for the database to eventually be consistent. A possible implication of moving from ACID to BASE is two customers buying the same book while there is only one copy available. Since two transactions can occur almost at the same time, both customers think they have bought the book and the organization needs to apologize to one of them. Although this is not an ideal situation, it is better to slowing down their site which will affect all customers.
Most new databases are NoSQL compliant where NoSQL is often defined as “Not Only SQL” or “Not Relational”. In his paper Cattell (2011) identifies NoSQL databases by the following six key features:

- The ability to horizontally scale throughput over many servers (nodes);
- The ability to replicate and to distribute data over many servers (nodes);
- A simple call level interface or protocol;
- A weaker concurrency model than ACID (e.g. BASE);
- Efficient use of distributed indexes and RAM for data storage;
- The ability to dynamically add new attributes to data records.

Most existing NoSQL databases can be categorized in four types of databases, namely key-value stores, document stores, extensible record stores and scalable relational systems (Cattell, 2011). The latter type refers to recent developments which makes it possible to horizontal scale traditional relational databases. However, benchmarks have showed that these databases cannot achieve scaling comparable with “real” NoSQL systems. This in turn indicates that there is a true need for highly scalable NoSQL databases, especially when dealing with complex data (e.g. in terms of the three Vs discussed earlier).

### 4.3.3. Hadoop, the open source heart of big data analytics

According to Forrester, Hadoop “is the nucleus of the next generation enterprise data warehousing” by delivering cloud-facing architectures, MPP, in–database analytics, mixed workload management and a hybrid storage layer (Kobielsu, 2012). Created by Doug Cutting, the creator of Apache Lucene, Hadoop provides a comprehensive toolset for building distributed systems, including data storage, data analysis and coordination (T. White, 2010). Hadoop originates from Apache Nutch, an open source web search engine. After realizing that existing architectures would not scale to the billions of pages on the web, the initiators wrote an open source implementation based on Google's distributed filesystem (Ghemawat, Gobioff, & Leung, 2003), called Nutch Distributed Filesystem (NDFS). In 2004 Google released a paper that introduced MapReduce, a parallel programming model and an associated implementation for processing, analyzing and generating large data sets across a cluster of commodity machines (Dean & Ghemawat, 2008), to the public. Nearly a year later all Nutch algorithms had been ported to use MapReduce and NDFS. In 2006, Nutch became a separate subproject under the name Hadoop and two years later it became a top–level project at Apache, confirming its success. In that year, Hadoop was used by many international organizations such as Last.fm and Facebook.

For many, Hadoop is a synonym for big data because of its capabilities to store and handle huge amounts of (unstructured) data within a smaller timeframe in an economically responsible way (Kuil, 2012). For this reason the Hadoop ecosystems plays a major role in big data analytics. Figure 6 illustrates the “mountain of data” commonly find within organizations. With traditional data analytics, only the peak can be analyzed and utilized in order to create value or support value creation. This peak often consists of highly structured data stored in traditional data warehouses. Since the amount of unstructured data is growing rapidly as described earlier, this peak is becoming relatively smaller. With Hadoop, it is possible to store and analyze unstructured data in a much smaller timeframe using the power of distributed and parallel computing on commodity
hardware. More important, the line indicating the boundary of data that can be utilized and data that cannot, is dropping, leading to a much greater peak and hence, in more possible value.

Together with its free license, huge community and open source techniques, many initiatives using Hadoop have been emerged, also indicating its success. Also, many big IT organizations started to distribute their own commercial version of Hadoop by adding enterprise support, additional functionalities and tools and even bundled with specific hardware (which contradicts on the fact that Hadoop is great because it runs on cheap commodity hardware). Examples are Cloudera, EMC GreenPlum, IBM InfoSphere BigInsights, Amazon AWS Elastic MapReduce and Microsoft’s release of Hadoop on its cloud platform Azure by porting Hadoop to Windows.

An important part of the Hadoop ecosystem is the Hadoop Distributed File System (HDFS) making the partition of data and computing across many nodes possible. Although often seen as a NoSQL database, it is not. Although both data storage systems use distributed storage across multiple nodes, there are significant differences. First of all, NoSQL databases are responsible for maintaining records by providing efficient ways to insert, modify or delete records using indexed attributes. HDFS does not index data, rather when Hadoop executes a job, it scans all data which is considered very inefficient within the field of NoSQL databases. This does not matter for Hadoop since its performance is mainly depending on the number of nodes and the sum of raw CPU power (Knulst, 2012). Another difference between NoSQL databases and Hadoop is the general purpose of the database: NoSQL database focusses on running small jobs as fast as possible (e.g. providing data for a website being requested) while Hadoop focusses on running big jobs using big files.

4.3.4. **In-memory analytics**

The techniques discussed so far are mainly focussing on analyzing huge amounts of unstructured data (both the volume and variety part of big data). Although these techniques both make use of disk storage and memory, another generation of data-warehousing runs primary in memory significantly improving the time needed to get the job done, sometimes even 100,000 faster than traditional techniques as Steve Lucas (2012) argued. Hence, in-memory analytics focusses on the velocity part of big data. In-memory analytics can make realtime analytics possible, even when dealing with large amounts of data. This in turns means that insights are more valuable, taking account the fact that some knowledge looses value when it becomes “old”, especially when also known by competitors.

In-memory analytics uses a so called “in-memory database” (IMDB) to perform analysis. Although this technique is not new (it originates from the 1980s⁴) it became an

---

4 In 1984 DeWitt, Katz, Olken, Shapiro, Stonebraker and Wood wrote about main memory database systems in their article called “Implementation techniques for main memory database systems”.

---

**Figure 6: New technologies make it possible to utilize more data (Knulst, 2012)**
important topic again as a result of the big data phenomenon. Especially for realtime analytics, in-memory analytics is important to achieve insights without a significant delay. With data becoming more complex, analyzing data without influencing the “realtime” aspect of realtime analytics using traditional analyzation techniques (e.g. relational databases) is becoming problematic. The main reason for this is that query data must be retrieved from a disk or network first (Plattner & Zeier, 2011, pp. 10–11). To give an example: accessing memory is 5.000 times faster than accessing a disk, and reading one megabyte of data from memory is 120 times faster than reading the same amount of data from a disk. Due to the architecture of modern computers, the CPU (which is responsible for analyzing data) communicates with the memory in a much faster way than the CPU does with the disk and hence, saves CPU cycles. The same applies to the GPU, an even more efficient and faster way to analyze data (Wu, Zhang, & Hsu, 2009).

Oracle developed their Exalytics In-Memory machine, an integrated BI solution consisting of collaborative software, middleware and hardware. Due to its 1 terabyte of RAM and Intel Xeon E704800 processor which consists of 40 cores, it delivers great capabilities regarding in-memory analytics. Also SAS delivers commercial in-memory analytics by providing SAS High-Performance Analytics that collaborates with Hadoop-based systems making it possible to analyze unstructured data using in-memory analytics and hence, making use of its advantages. This has a downside too as Basjes & Mathijssen (2012) points out:

“Hadoop is built in a way that the node containing a particular piece of data will also analyze the same piece of data. This keeps the distance between data and CPU small resulting in more performance. However, some solutions break this concept by separating the data storage from the analysis capacity because suppliers make more money with existing data analytics solutions.” (Basjes & Mathijssen, 2012)

At last, also non-commercially products offering in-memory analytics are available such as VoltDB, which can be run on commodity hardware.

4.4. Summary

To conclude this chapter, the first sub-research question “what characterizes big data and how does big data analytics differ from traditional data analytics?” can be answered in short as follows: data changed in terms of the three Vs: volume, variety and velocity, due to several causes and required the need of new analysis techniques to handle these changes. Next, this answer is elaborated.

Data created by users and organizations is changing rapidly in terms of the three Vs: volume, variety and velocity. In addition to that, organizations started to combine exotic and external data sources and started to analyze forms of unstructured or multi-structured datasets in order to extract valuable insights. These statements indicate that big data analytics is focusing on a new level of data; rather than using structured and sometimes even sampled data originating from traditional EDWs, big data analytics is about combining all sorts of data sources to extract unique insights that would have been hidden without the set of technologies making big data analytics possible. These technologies enable organizations to analyze much more data, especially unstructured,
multi-structured and multimedia data. Also, due to the introduction of distributed and parallel analysis techniques such as Hadoop and the global acceptance of the cloud, analyzing this data happens in a much smaller timeframe leading to more added value as most insights lose value when they become older, also indicating the importance of real-time data analytics.

IDC defines big data analytics as “a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis” (Villars, Eastwood, & Olofson, 2011). Since this definition is not specific enough, this definition is extended by giving some characteristics making it easier to recognize the presence of big data analytics when examining the cases in chapter 6:

- The volume or velocity of data being analyzed within a required timeframe, requires the need of distributed computing power or an in-memory database, or;
- At least one of the used data sources uses data which can be identified as unstructured or multi-structured data, or;
- No fixed or predefined connection between datasets exist.
5. A potential value creation driver

In this chapter different value creation theories are examined. Each theory sheds its own light on how organizations create value in general. By connecting these theories in section 5.2, hypotheses are generated acting as an answer to the second subquestion “what causes and affects value creation within organizations?”. These hypotheses make it possible to identify whether a big data analytics case is creating organizational value and hence, will be used in chapter 6. In the last section of this chapter, existing reports and literature is used to argue why big data analytics might be seen as a value creation driver itself and answers the third sub-research question “which impact does big data analytics might have on value creation of an organization?”.

5.1. Value creation theories

In their research, Amit & Zott (2001) introduced a model illustrating how value is created within e-business organizations. Besides examining dozens of organizations, they used existing theories on value creation to define their model. In order to define big data analytics in terms of value creation, the same theories are used to describe how value can be created when an organization is successfully performing its value activities. Next, these theories will be discussed.

5.1.1. Value chain analysis

Value chain analysis identifies the activities of the organization and then studies its economic implications (Amit & Zott, 2001). Value is defined as the amount that buyers are willing to pay for a product or service. The organization is profitable if the value it commands exceeds the cost of performing these value activities. Hence, an organization becomes more valuable when either the value activities can be performed at lower costs or results in more sales.

An organization's value chain is a system of interdependent activities connected by linkages (Porter & Millar, 1985). These linkages not only connect value activities within an organization but also connect value chains between organizations. This means that the value of an organization can also be influenced by its strategic partners, for example a supplier.

Value chain analysis explores both primary activities (activities involved in the physical creation of the product, its marketing and delivery to buyers and its support and servicing after sale, hence activities having a direct impact on value creation) and support activities (activities supporting primary activities through their impact on the performance of the primary activities).

5.1.2. Schumpeterian’s innovation theory

In Schumpeterian’s innovation theory, innovation is the source of value creation, emphasizing the importance of technology and research & development (R&D). According to this theory, value can be created by introducing new goods, new production methods, new markets, new supply sources and the reorganization of industries through technical innovation (Amit & Zott, 2001) resulting in economic development (Moran, 1997). Existing products, services and technologies can be combined resulting in new products or services, which in turn can be the start of the creation of another product or service.
This vicious circle is the foundation of the rise of complete new markets and hence, creative destruction (also called “Schumpeter’s gale”) referring to the creation of entire new markets destroying old markets.

An important aspect of Schumpeterian’s innovation theory is the way how existing products, services and technologies can be combined resulting in a new product or service. Schumpeterian’s innovation theory originates from the 1930’s, an industrial period in which “combining existing products and services” referred to the combination of tangible and physical goods. Galunic & Rodan (1998) extends Schumpeterian’s innovation theory by arguing that innovation in non-technical areas is a possible value creation driver too:

“\textit{Innovation, more broadly, can occur in a wide variety of areas, including the underlying business processes, marketing image and strategic positioning of the firm, all of which may be impacted by how the underlying resource of a firm are combined and configured.”} (Galunic & Rodan, 1998)

According to Galunic & Rodan (1998), new products or services are created by combining and reconfiguring resources. These resources can be categorized in two groups: input resources and knowledge–based resources. Input resources mostly refer to tangible resources like people, equipment, property rights and capital while the latter refers to an intangible ingredient needed to combine the resources productively and in a value-creating way: knowledge. Knowledge–based resources operate on the input resources “providing both specialized understanding of the separate inputs as well as coordinative understanding of how the various input resources fit together to provide value to the firm” (Galunic & Rodan, 1998).

5.1.3. Resource-based view

The resource–based view is closely related to Schumpeter’s innovation theory, especially Galunic and Rodan’s extension described in the previous section. In the resource–based view, resources are defined as “anything which could be thought of as a strength or weakness of a given firm” (Wernerfelt, 1984) and according to this theory, an organization is a bundle of resources and capabilities (Amit & Zott, 2001). An organization is creating value when marshalling and uniquely combining a set of these resources and capabilities in order to reduce the organizational costs or increase its revenue “compared to what would have been the case if the firm did not possess those resources” (Amit & Zott, 2001).

5.1.4. Strategic networks

Strategic networks are “long-term, purposeful arrangements among distinct but related for-profit organizations that allow those firms in them to gain or sustain competitive advantage” (Jarillo, 1988) and can exist in forms of strategic alliances, joint ventures, long-term buyer–supplier partnerships and other ties (Amit & Zott, 2001). The relationships between organizations within a strategic network are essential to their competitive position. With strategic networks, organizations often outsource existing functions to other organizations. Strategic networks theory is arguing that networking can create value by enabling access to resources (information, markets, technologies,
activities, etc.). Also strategic networks offer risk sharing, the opportunity to scale and scope more easily, knowledge sharing and learning (Amit & Zott, 2001).

5.1.5. Transaction cost economics

In transaction cost economics theory, value creation occurs when increasing the efficiency of a transaction as enhanced efficiency reduces costs. Hence, this theory is more focussed on cost and effort reduction. Value can be derived “from the attenuation of uncertainty, complexity, information asymmetry, and small-numbers bargaining conditions” (Amit & Zott, 2001). Also reputation, trust and transactional experience can increase transaction efficiency. Important is the broad definition of a transaction: it is defined as a transfer of a good or service across a technologically separable interface (Amit & Zott, 2001).

5.2. Linking value creation theories

So far the different theories on value creation are introduced. In this section connections between these theories are exposed leading to a comprehensive view of how a value activity might create value. Figure 7 illustrates the connections between the theories discussed. Next these connection will be discussed in more detail including examples.

According to the Schumpeterian’s innovation theory, innovation is the ultimate source of value creation. This innovation comes from the combination of existing physical goods or knowledge-based resources (Galunic and Rodan’s extension). Hence, the connection (A) between Schumpeterian’s innovation theory and the resource-based view becomes clear:
both argue that combining resources is a major value creation driver. This might have another goal than just creating new products and services: it also can improve primary or secondary activities leading to transaction efficiency improvements resulting in a decrease of costs and hence, indicates a connection with the transaction cost economics theory (B & G). This also indicates a connection between the transaction cost economics theory and the value chain analysis theory since, according to the latter, value creation occurs when the value activity can be performed at lower costs. An example would be an organization creating an improved version of its factory machines leading to energy savings and an increase in capacity per hour. This organization might also invest in a long-term buyer–supplier partnership by exclusively renting these machines from another organization leading to the same efficiency improvements. This indicates a connection (C) between the strategic network theory and the transaction cost economics theory. Strategic networks also influence the set of resources for a particular organization, since a strategic partnership with another organization opens a new range of resources that can be used to create unique combinations. This indicates a connection between the strategic network theory, Schumpeterian’s innovation theory (e.g. by creating a new supply source) (H) and resource–based view theory (e.g. access to specific knowledge) (I). Since, according to the value chain analysis theory, an organization’s value chain is also depending on inter–organizational connections called linkages, a clear connection between this theory and strategic networks theory exist (D). Schumpeterian’s innovation theory might influence the value chain of the organization (E) by improving organizational efficiency through the combination of (new) resources. Also, by innovating, organizations have the ability to sustain or increase the amount that buyers are willing to spend for a (new) product or service, leading to added value. Such an innovation is often temporary as competitors will try to catch up. By protecting innovations, a (temporary) monopoly emerges, restraining competitors to do so. The resource–based view theory also connects with the value chain analysis (F) as combining resources might reduce organizational costs (e.g. due to innovation or strategic networks) or increase sales (e.g. due to new products or a new marketing strategy).

By connecting the earlier described value creation theories it becomes clear that there is a significant overlap between these theories, also indicated by Amit & Zott (2001). To summarize and answer the second sub–research question, value creation within organizations is caused and affected when an activity acts as a value creation driver. This happens when it:

- Creates, supports the creation of, or improves a product or service, of which the amount buyers are willing to pay is higher than the cost to create and leverage it (value chain analysis), or;
- Creates, supports the creation or improves a product or service (possibly leading to entire new markets) by combining existing resources (Schumpeterian’s innovation theory and Galunic and Rodan’s extension), or;
- Combines resources in an unique way reducing organizational costs or increasing its revenue (transactional cost economics theory), or;
- Is a result of a strategic partnership leading to competitive advantage (strategic networks theory), or;
- Reduces costs by improving transaction efficiency (transaction cost economics theory).
Remarkable but quite obvious is the fact that most theories directly link value with profit, either by decreasing costs or increasing sales. Only Schumpeterian’s innovation theory is not directly referring to this. It indeed argues that technological innovation is the source of economical development (Moran, 1997) but not necessarily for that particular activity. Many new products and services failed to be profitable but yet, value still might have been created; perhaps in forms of knowledge-based resources or technical developments. These resources can be used for the creation of a new product or service which eventually can be profitable, hence leading to economic development.

5.3. Value creation hypotheses

In order to judge if a big data analytics case in chapter 6 is acting as a value creation driver, the characteristics of the earlier discussed value creation theories are aggregated and expressed in three hypotheses. Each hypothesis connects to at least one value creation theory in order to proof its value for the organization discussed earlier.

\( H1: \) Big data analytics improves transaction efficiency leading to a reduction of costs or an increase in sales;

By improving the efficiency of the value chain of the organization, remaining the amount that buyers are willing to spend for a particular product or service, costs are reduced, hence the organization becomes more profitable. In addition, by improving the transaction efficiency between supplier and customer, costs are saved, hence value is created. This benefits either the supplier, the customer or both, depending on the activity which has been improved and its implications.

\( H2: \) By utilizing an exclusive set of resources through a strategic network, making a specific big data analytics case possible, the amount buyers are willing to spend increases;

As stated earlier, the relationships between organizations within a strategic network are essential to their competitive position as it creates an unique set of resources in which the buyer is interested (directly or indirectly) to invest more of its own resources (often money). In other words, strategic networks increase the amount that buyers are willing to spend which in turn, according to the value chain analysis, creates value for the organization.

\( H3: \) Big data analytics creates, supports the creation or improves products, services, markets or channels leading to significant advantages for customers.

Although innovation can be a main driver for transaction efficiency improvements leading to a reduce of costs for both the customer and organization (as stated in the first hypothesis), this hypothesis focuses primarily on the customer. As described earlier, an innovation can be valuable in many ways. Yet, since identifying the value of such an innovation is hard or even impossible, an innovation is considered valuable if it offers
significant advantages for the customer. A good example is the Apple iPhone. When Apple launched its iPhone in 2007, people queued up for days in order to be among the first owners worldwide (“Where would Jesus queue?,” 2007). The reason? Its innovative multitouch screen and its easy-to-use operating system, although it lacked some basic features as 3G Internet speed and video recording capabilities (Vogelstein, 2008). Value created from innovation is often short-lived as competitors often easily have the ability to copy this innovation. This applies mainly to technical innovations. For this reason, in the resource-based view, the attractiveness of a technical resource is considered doubtful:

“On the one hand, a technological lead will allow the firm higher returns, [...] The followers, on the other hand, will often find the reinvention of your ideas easier than you found the original invention.” (Wernerfelt, 1984)

Wernerfelt states that, in order to see a technical innovation as a valuable resource, an organization needs to keep innovating:

“You need to keep growing your technological capability in order to protect your position. This should, however, be feasible if you use your high current returns to feed R&D” (Wernerfelt, 1984)

Once again, Apple is a good example of this. Every year after the iPhone was released, Apple continued releasing new versions with attractive new features that kept customers interested in this phone. The introduction of the App Store, which in March 2012 contained more than 550,000 applications (“Apple's App Store Downloads Top 25 Billion,” 2012), is such an innovation. Patents often help organizations to protect certain innovations, something that is also the case for Apple in their patent–war with Samsung (Bonnington, 2012).

5.4. Potential value creation driver

In the previous section, different hypotheses are defined indicating how big data analytics might affect organization’s value creation. This section starts by examining different examples. The results, together with other data sources, will be used to elaborate on each hypothesis.

5.4.1. Examination of both existing and non-existing examples

Next, eight different examples of big data analytics implementations are examined of which each example either exist as a real case or represents a fictitious case made up by the corresponding author indicated as a reference. The selection is based on their generality and the general ease of understanding. Per case, its “big data factor” and how it affects organization’s value creation is described. The examples will be used in the next sections to imply specific hypotheses.
## 5.4.2. Effects on transaction efficiency

Porter & Millar (1985) argues that information is an important asset and due to the information technology revolution, organizations changed the way they operate. Data forms the fundament of information as described in section 1.1 and the last decades the costs to store, manipulate and transmit data are dropped and the error rate compared to manual data entries has been decreased significantly (Porter & Millar, 1985). As a result data organizations have changed significantly as discussed in detail in section 4.2. Due to emerged technologies, this data can be analyzed in less time. Table 1 contains different

---

<table>
<thead>
<tr>
<th>#</th>
<th>Description (source)</th>
<th>Big data analytics case</th>
<th>Impact on value creation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Analyzing patient characteristics in combination with outcomes of treatments (Manyika et al., 2011)</td>
<td>Analyzing large datasets in a timely manner</td>
<td>Reduce over-treatment and under-treatment</td>
</tr>
<tr>
<td>2</td>
<td>Analyzing physician entries and compare them with guidelines to alert for potential errors (Manyika et al., 2011)</td>
<td>Analyzing different data sources including unstructured data sources such as X-ray images</td>
<td>Reduce adverse reactions and lower treatment rates and liability claims</td>
</tr>
<tr>
<td>3</td>
<td>State-of-the-art cross selling (Manyika et al., 2011)</td>
<td>Analyzing different (exclusive) data sources including multi-structured data sources</td>
<td>Easier access to additional customer information eventually leading to an increase in sales</td>
</tr>
<tr>
<td>4</td>
<td>Sensor–driven operations in manufacturing (Manyika et al., 2011)</td>
<td>Realtime analyzing granular data originating from sensors (IoT)</td>
<td>Improved transaction efficiency due to ubiquitous process control and factory optimization</td>
</tr>
<tr>
<td>5</td>
<td>Customer ad targeting based on behavior (Schmarzo, 2011)</td>
<td>Sets of users based upon behavior, demographics, etc. is formed</td>
<td>Targeted sales improves the transaction efficiency</td>
</tr>
<tr>
<td>6</td>
<td>Automated crime intelligence (Genovese &amp; Prentice, 2012)</td>
<td>Analyzing different (exclusive) data sources including multi-structured data sources</td>
<td>More–accurate intelligence and reliable solution</td>
</tr>
<tr>
<td>7</td>
<td>Listening to thousands of customers (IBM Big Data Success Stories, 2011)</td>
<td>Analyzing different big datasets with feedback originating from web, email and text messages</td>
<td>Automated tagging of feedback saves field managers work, hence time</td>
</tr>
<tr>
<td>8</td>
<td>Fraud detection on population (Bloem et al., 2012)</td>
<td>Analyzing all bank transactions for criminal behavior</td>
<td>Better insights in criminal behavior and the ability to prevent it</td>
</tr>
</tbody>
</table>

Table 1: Existing and non-existing big data analytics examples and their impact on value creation
examples all relying on the presence of data. Not all examples were primarily implemented to improve transaction efficiency, yet one might find reasons it does. Examples 1 and 2 focusses on service innovation with the customer (the patient) in mind. However, it can also improves transaction efficiency in different ways. For example, it can help the doctor making certain decisions based on the outcome of the analysis. In addition, it also might improve the efficiency of communication between the doctor and the patient as it saves time to go through the patient’s medical history.

Example 3 is a nice example indicating the advantages of the analysis of external data sources as it might contain very interesting data when looking at purpose of the example: finding new ways to sell products or services to new or existing customers. By utilizing the features web 2.0 brought us, this example might profit from social networks as Twitter and Facebook to collect behavioral data of a specific person or a group of people. Analyzing this behavioral data might lead to additional insights improving the sales activity, hence enhances transaction efficiency. For example, by monitoring social media, feedback about products can be collected more easily, which might improve the efficiency of the sales department as they have a better understanding of their customers.

According to McKinsey, by making data more easily accessible to relevant stakeholders in a timely manner, tremendous value can be created (Manyika et al., 2011). This is particularly the case between otherwise separated departments. Transparency leads to more information which could have a positive impact on transactions’ efficiency and implies that big data analytics is a value creation driver as it opens up the possibility to exploit more data leading to information, as illustrated in figure 6. A good example of this is the Dutch Health Club, a program initiated by Almere DataCapital. The amount of data generated by hospitals is increasing explosively, especially multimedia leading to an increase of costs (Manyika et al., 2011, p. 21). The Dutch Health Club started a consortium consisting of hospitals helping each other by sharing data and knowledge to handle this data (Gibbels, 2012, p. 1). As a result costs can be reduced and unlocks a new set of (knowledge) resources possibly leading to added value as argued by the strategic network theory in section 5.1.4. Another example of the advantage of transparent information is illustrated as example 4 in table 1. Analyzing the manufacturing process in a automated way using sensors embedded in factory machines opens up different possibilities to enhance transaction efficiency. For example, the sales department could be informed automatically if the manufacturing process has problems due to defects recognized by machines’ sensors. Hence, information is shared in a transparent way between otherwise separated departments.

Comparable with example 3, example 5 illustrates a case in which a specific group of people is targeted to show ads. On a specific website this can be easy (e.g. targeting pet owners on a pet forum). Yet, when a specific group of people needs to targeted on a general website (e.g. Facebook), a smarter way of targeting should be used. By collecting and analyzing behavior, demographics, personal details, etc., a pattern is created that can be matched with other visitors. The results can be used to target ads in an automatic and simple way leading to a reduction of advertisement costs.

The 6th example indicates a case in which the population is analyzed instead of a sample. Due to the improved capacity of data analysis systems, mainly due to distributed analysis as described in section 4.3.3, this is now possible even with tremendous amounts of data. Kuil, innovation manager at the Rabobank, also argues this by saying that big data analytics is especially valuable for fraud detection as the entire population can be analyzed instead of samples (Kuil, 2012).
Big data analytics also opens up a range of possibilities within the field of automated decision making as it delivers more knowledge-based resources faster than before (Bloem et al., 2012, p. 5; Schmarzo, 2011, p. 6). Automated decisions are based on algorithms and can minimize risks and unearth valuable insights that would otherwise remain hidden (Manyika et al., 2011, p. 99). In a study conducted by TDWI in 2011, 37% of all respondents indicated that “automatic decision-making for realtime processes” was one of the possible benefits when implementing big data analytics (Russom, 2011).

Example 7 in table 1 illustrates an example of a job that can be made automatic using big data techniques. Hertz receives thousands of emails, text messages and feedback-messages through their website. Before, managers had to tag these unstructured messages manually. Using content analytics software, this is done automatically saving an incredible amount of time, hence led to transaction efficiency. Yet another example of an automated decision system is illustrated as example 8 in table 1. It describes a crime intelligence system in which big data analytics plays an important role as it is better capable of recognizing patterns resulting in a more accurate and reliable system (Genovese & Prentice, 2012, p. 2). By using different (exclusive) data sources (e.g. speed camera photos), it is capable of detecting potential dangers and sends police as a precaution.

5.4.3. Effects by partnering through a strategic network

Data, like many other knowledge-based resources, often has no clear ownership. This leads to conflicting arguments: one might argue that owning data (e.g. Facebook having data about its customers) is becoming more valuable but on the other hand data is becoming more available for anyone to use, often for free, through public API’s and open data platforms. It appears that the ability to use data in an unique way is becoming more valuable compared to just owning data, something which is supported by the resource-based view and Schumpeterian’s innovation theory.

When looking at table 1 and its examples, the 2nd hypothesis requires another approach to identify its impact on the cases and visa versa as this hypothesis says something about big data analytics rather than on organizations’ value creation. All examples might benefit from a specific partnership (e.g. the use of a specific data source or technology). Yet, it is likely that activities directly influencing organization’s competitive position, benefit most from strategic partnerships as these partnerships prevent competitors to utilize the same advantages. Examples 3, 5 and 7 are likely private organizations as it concerns the acquisition and retention of customers, something that has a direct impact on the organization’s competitive position. Furthermore, the effects of a strategic network can be considered doubtful when looking at examples 1 and 2 because internal resources will mostly be used due to, for example, policies.

As more data is publicly accessible, it is harder to leverage these resources in an unique way and according to Wernerfelt (1984), an organization wants to create a situation where its own resource position directly or indirectly makes it more difficult for others to catch up. “Competitive advantage equals the difference between the value created by the company and the potential value created by its competitors” (Spulber, 2009, p. 233). Hence, in order to create competitive advantage, an organization should make it difficult for competitors to catch up. Owning resources others do not have is an easy way to achieve this, but as stated before, this is becoming more difficult. For this reason big data analytics might act as a value creation driver when it increases the
amount buyers are willing to spend for a particular product or service as a result of partnering through an strategic network, which excludes competitors from utilizing the same advantages. The micro-blog Twitter is providing exclusive paid access to their complete realtime data feed called the “Firehose” (“Enabling A Rush of Innovation,” 2010). With this, organizations have the ability to analyze all status updates which can be valuable in many ways, for example targeting ads or improving products leading to competitive advantage. Not only partnering with data suppliers can be valuable, also strategic partnerships with software suppliers can be valuable as it gives the organization the ability to extract value from data in an unique way through the use of exclusive software.

5.4.4. Effects on organization’s innovation level

McKinsey argued that big data analytics “enables companies to create new products and services, enhance existing ones, and invent entirely new business models” (Manyika et al., 2011, p. 5). The third hypothesis focusses solely on the customer. To what extent does the customer benefit from this new or enhanced product, service or even business model? Most examples in table 1 indicate enhanced products or services and all of them will, in the end, offer advantages to its customers (e.g. example 4 might lead to notifications to the customer when the production process is delayed due to defects), but only the first two examples are designed with the customer in mind and implies support for the third hypothesis arguing that big data analytics is a value creation driver since it may result in a new or improved product or service with significant advantages for customers.
6. A true value creation driver

An understanding of the definition of big data analytics and how it might affect organizations’ value creation, is formed. Using the criteria provided in section 3.3.4, two cases have been studied: X FACTOR Heartbeat, an application showing live statistics about the participants of the music show X FACTOR and bol.com, an Internet-based retailer. Each case is discussed separately in this chapter. The possibility exist, that multiple hypotheses can be supported by the same case. For example, as section 5.4.2 also implies, a new service as a result of a big data analytics implementation, can offer advantages for both the customer and the organization as well as improve transaction efficiency.

6.1. Case I: X FACTOR Heartbeat

6.1.1. Introduction

X FACTOR is a music competition started being displayed on national television in 2006 via RTL Nederland. The goal is to find new singing talent. The show consists of different stages. During each stage, participants must audition in order to get to the next stage. If they fail to impress, they are out of the show. The selected finalists are asked to do a few live performances. Each performance is rated by both judges and by the public using televoting (e.g. SMS). During the last season, Lost Boys Mobile (a partnership between LBi Lost Boys and Triple IT) and RTL Nederland developed X FACTOR Heartbeat: a second screen web application which displays the heart rate and social activity around a particular participant. This participant is wearing multiple heart rate sensors sending data to one of the three receivers used. Each receiver is sending the data to a web server. Hence, this case is a nice example of the “Internet of things” phenomenon discussed in section 4.2.1. Data originating from Twitter is also collected and analyzed. Figure 8 illustrates the web application and gives an impression of the functionalities the application offers.

Figure 8: The X FACTOR Heartbeat web application
6.1.2. **Data collection**

As an employee of Triple IT, I easily had access to data regarding this case. Data originated from the application itself, pseudocode\(^5\) and lists containing words to calculate the sentiment of a tweet as will be described in the next section.

In addition to these data sources, also an interview was held with Lippes, responsible for technical matters such as the development of the web application and delivering support. When studying this case, I also had different email conversations with Lippes and Zonneveld, project leader and acting as the bridge between development and the business case.

The following online article which summarizes an interview with Gooijer (director Lost Boys Mobile) and Bruitsman (Creative Director RTL Concepts), was also used:

http://www.frankwatching.com/archive/2011/05/06/x-factor-heartbeat-webapp-live-hartslag-van-kandidaten

6.1.3. **X FACTOR Heartbeat as a big data case**

Since the web application was only used during the live shows, not much computational power was required between the live shows because almost no visitors visited the web application. However, during the show, the web application was used very intensively: during the second live show (each live show runs between one and two hours, once a week) 200.000 page views and 41.000 unique visitors were recorded, each visitor using the web application for more than 7 minutes on average. In order to handle these peak-moments, the web application used the cloud as infrastructure. By using Microsoft Azure, computational power could be increased easily by adding instances (virtual nodes) in order to handle traffic in an economically way. However, the analytics part of the application (collecting, aggregating, analyzing and storing data from both the heart rate sensors and Twitter) was being done by only one instance which does not make this case a big data case according to the definition provided in chapter 4:

*The volume or velocity of data being analyzed within a required timeframe, requires the need of distributed computing power or an in-memory database.*

It is the analysis of the Twitter streaming feed what makes this case a big data case:

*At least one of the used data sources uses data which can be identified as unstructured or multi-structured data.*

Using the Twitter streaming API, tweets\(^6\) about a particular participant were collected. This was done by making a connection to the API which was kept open as long as data was required in contrast to a normal request, which closes after a predefined amount of bytes has been sent. The application subscribed to a selected number of hashtags (e.g. the #heartbeat hashtag) in order to receive tweets containing this hashtag. Also the

---

\(^{5}\) Human-readable programming language to describe the operating principle of a computer program or algorithm

\(^{6}\) A tweet is a short message send using Twitter
application subscribed to the hashtags of the participants enabling the application to distinguish between participants. Each tweet was analyzed by detecting the gender of the author. This was done by looking up the name of the author in predefined list containing names of each gender. After that, the sentiment of the tweet was calculated using a more complex algorithm. Each word of the tweet was looked up in predefined lists with words (negatively charged words, positively charged words, reinforcing words and words indicating a negation) that may (help) determine the sentiment of the tweet. A score was calculated depending on the presence, combination and order of these words. When this score was higher than zero, the tweet was considered positive and when the score is below zero, the tweet was considered negative. When the score was zero, it did not influence the overall sentiment of all tweets at all. The sentiment of the tweets appeared to be corresponding with the (virtual) audience: when a participant was told to leave the show and the audience did not agree, the sentiment indicated a negative charged tension. The opposite was also true: when a participant was told to leave the show and the audience did agree: the sentiment indicated a positive charged tension.

Both with determining the gender of the author and the sentiment of a tweet, a chance existed that either one or both could not be determined correctly. This happened for example when the name of the author did not exist in the list with names of each gender. With big data analytics, especially when analyzing unstructured or multi-structured data, this sometimes occurs. These missing or false interpretations are often influencing the outcomes in a negative way but due to the amount of data the magnitude of this effect is often considered negligible, something that has been mentioned often in the big data debate as “quantity beats quality”, as with X FACTOR Heartbeat is the case too. However, this may be considered doubtful as bigger data is not always better data; the size of data should fit the question being asked (Boyd & Crawford, 2011, p. 8).

6.1.4. Effects on organization’s value creation

In chapter 5, three hypotheses assuming how big data analytics might affect value creation within organizations, hence acts as a value creation driver, are described. In this section each hypothesis is examined in the light of the X FACTOR Heartbeat case.

\begin{itemize}
\item \textbf{H1:} Big data analytics improves transaction efficiency leading to a reduction of costs or an increase in sales;
\end{itemize}

In section 5.1.5, a transaction is defined as a transfer of a good or service across a technologically separable interface. Technologically separable interfaces defines the boundaries of tasks between which transactions may occur (Foss, 1998). In this case, the service would be the ability to see the heart rate and social activity around a particular participant while the web application would represent the technologically separable interface between the user and front–end of the web application.

Like examples 1 and 2 from table 1 in the previous chapter, this case is also created with the customer in mind. As discussed in section 5.4.2, these examples might also improve the efficiency of the transaction between the patient (customer) and the doctor (organization). Since the case discussed in this section is representing a new transaction, no transaction is made more efficient. In addition, Lippes mentioned that the efficiency of the show might be affected in a negative way as participants had to be provided with sensors what took a considered amount of time. Also, during the show, the
participant changed cloths multiple times. This caused extra delay as a certain amount of
time was needed to hide the heart rate sensor as good as possible. In addition,
sometimes batteries needed to be replaced.

\[ H2: \text{By utilizing an exclusive set of resources through a strategic network, making a specific big data analytics case possible, the amount buyers are willing to spend increases;} \]

Section 5.4.3 argues that many big data implementations will probably benefit from a
specific partnership by, for example, using a specific data source or technology. This case
also benefits from such a partnership by collecting data from Twitter. This indicates a
collaboration between RTL and Twitter but cannot be identified as a strategic partnership.
In chapter 5 a strategic network is defined as a “long-term, purposeful arrangement”.
According to the Oxford dictionary, an arrangement is defined as “an agreement with
someone to do something”. In order to use the Twitter streaming API’s, RTL had to agree
with certain terms and conditions indicating some sort of agreement. Yet, this agreement
can be made by other organizations too, leading to a loss of value. In his paper Jarillo
(1988) extends his definition of a strategic network by arguing that:

“If a firm is able to obtain an arrangement whereby it ‘farms out’ activities to the
most efficient supplier, keeps for itself that activity in which it has a comparative
advantage, and lowers transactions costs, a superior ‘mode of organization’ emerges: the strategic network” (Jarillo, 1988)

In other words: in order to create a valuable strategic network, organizations need to
create an arrangement and exclude other organizations preventing them to utilize the
same advantages. With X FACTOR Heartbeat, every individual or organization is able to
monitor the public streaming API’s and hence, create a similar functionality. You might
say that the combination of the heart rate information (which is not shared with other
organizations or is made public in a reusable way) and tweets around a particular
participant, leads to an unique product and hence, to competitive advantage, but this
data is not aggregated in a computational way (e.g. analyzing by finding correlations in
data originating from the different data sources). In other words: no network is created in
this sense.

\[ H3: \text{Big data analytics creates, supports the creation or improves products, services, markets or channels leading to significant advantages for customers.} \]

The main goal of the web application was to involve more people with X FACTOR and the
participants by using new technologies and innovations. This included the use of Twitter,
the cloud as infrastructure and the “Internet of Things” phenomenon by connecting
heart-rate sensors to the Internet. The combination of these innovations, all originating

\[ 7 \text{ Retrieved from the online version of Oxford dictionary on July, 16th 2012 from http://oxforddictionaries.com/definition/english/arrangement?q=arrangement} \]
from the 21st century, made it possible to develop this web application. According to the Schumpeterian’s innovation theory and resource-based view, these innovations are resources that can be combined in order to create a new product or service, as described in section 5.1.2 and 5.1.3 respectively, and act as a source of value.

The amount of viewers indicates the success of the application. More than 55% of the users used this application on a mobile device. Although no proof exist, it is likely that most users used this application as a second screen experience meaning that they watched the show on their television and used this web application as an extension to gain access to additional information. This exactly indicates the additional advantages for the customer. Lippes also mentions that, even when using the web application as a first screen experience, advantages are present:

“You can follow the show without actually seeing the show on television. By paying attention to the heartbeat and the social activity (including the sentiment of the tweets) around a particular participant you are able to see if he is about to come on stage, when he is performing, if it is any good and when he goes back stage again.” (Lippes, 2012)

6.1.5. Conclusion
This case clearly indicates that big data analytics is not only about volume. Although a lot of data was analyzed, it was analyzing Twitter’s multi-structured streaming feed that makes this case a big data analytics case. The sentiment of each tweet and its implication on the overall sentiment of a participant’s live show was calculated using list of words.

In this case, big data analytics acted as a value creation driver as it offered significant advantages for the customer and was a result of the combination of (technical) innovations seen as resources by both the Schumpeterian’s innovation theory and the resource-based view. Hence, it supports the third hypothesis and failed to support the first and second hypothesis.

6.2. Case II: bol.com’s search suggestions

6.2.1. Introduction
The online retailer bol.com started in 1999. In 2011, it had more than 3.4 million customers in both the Netherlands and Belgium, and is the market leader in the field of online books, entertainment, electronic devices and toys (“Over bol.com,” n.d.). It is also the largest Internet organization in the Netherlands and Belgium. In January 2011, more than 3.4 million unique visitors visited bol.com (Azevedo, 2011), leading to an impressive amount of web footprints: data containing customer’s web behavior on bol.com’s website. Google launched “Google Suggest” in 2004 as part of their experimental environment called “Labs” (Gibbs, 2010). When visitors type in the first characters of their search query, Google provides a list with most relevant search queries based on historical data analyzed earlier. This makes searching on the Internet more efficient and less time consuming. Although a few years later, this experiment inspired Basjes, IT architect at bol.com, to create a similar functionality on their website. In 2010, after a period of developing, testing and tweaking, the search suggestion functionality went live.
6.2.2. **Data collection**

Data regarding this case was collected by using the following data sources:

- An interview was held at bol.com’s headquarter in Utrecht on July 10th, 2012 with Basjes, IT architect at bol.com and initiator of this case, and Mathijssen, site search specialist at bol.com;
- Presentation by Basjes given at the BIG DATA Forum 2012 in Almere on April 17th, 2012, the first independent big data conference forum in the Netherlands;
- Presentation by Basjes given at the Webanalytics Congress on March 15th, 2012 in Utrecht. For this study, the published presentation was used and can be found at: http://www.webanalyticscongres.nl/presentaties-2012/295-klanten-helpen-klanten-met-webdata

In addition to above sources additional sources were used to verify and lookup details. These sources are indicated as references and can be found in the bibliography.

6.2.3. **bol.com’s search suggestions as a big data case**

Seen by many as one of the pioneers in the area of big data analytics in the Netherlands, this case clearly indicates the new possibilities when using emerged technologies capable to handle evolutionized data in a different way. When a visitor visits bol.com, anonymous data is being collected. This data includes click data (e.g. when the visitor clicks on a product) and search data (e.g. when the visitor searches for a product). Every night this data is analyzed on a Hadoop cluster containing 5 commodity server nodes, each containing 32 gigabyte RAM. The result was a list with search suggestions for different possible alphabetical character combinations. Initially these search suggestions were based on just one week of historical search data. This data was analyzed every night to identify popular search terms. Since visitors could also type in products or categories bol.com does not sell, some search suggestion were created not leading to any product and were in fact a dead end. To overcome this issue, the algorithm was improved multiple times and was tested by looking at its impact on the website’s conversion rate and overall sales. Currently the algorithm is based on two principles: the depth of a visitor in the sales funnel and the age of behavioral data. Starting with the first, clicking an item to see its details happens more than actually buying the item. Hence, behavioral data consisting a purchase represents a more valuable piece of behavioral data. This principle is illustrated in figure 9: due to the rarity and relevance of purchases compared to searches, behavioral data originating from purchases is more valuable.
Secondly, since historical behavioral data is analyzed and information loses value when it becomes older (as described in section 4.2.3), the algorithm also takes the age of data into account. This principle is illustrated in figure 10. Currently 100 days of historical behavioral data is analyzed each night in order to create a list with search suggestions. Each day around four gigabyte of compressed behavioral data is collected (during peak season this increases to eight gigabyte), hence theoretically each night around 400 gigabyte of compressed data needs to be analyzed. This indicates the need for a Hadoop cluster and the “evolutionized” aspect of data as previously written, as it indicates big data in terms of volume. It qualifies itself as a big data case according to this characteristic provided in chapter 4:

*The volume or velocity of data being analyzed within a required timeframe, requires the need of distributed computing power or an in-memory database.*

In order to save CPU cycles, hence costs, so called preprepared “intermediates” are generated every night of the previous day. These intermediates make it possible to analyze only the part which need to be recalculated on daily basis.

### 6.2.4. Effects on organization’s value creation

In chapter 5, three hypotheses assuming how big data analytics might affect value creation within organizations, hence acts as value creation driver, are described. In this section each hypothesis is examined in the light of the bol.com’s search suggestions case.
**H1:** Big data analytics improves transaction efficiency leading to a reduction of costs or an increase in sales;

As stated in section 5.1.5, according to Williamson (1989, p. 142), a transaction occurs when a good or service is transferred across a technologically separable interface. Such a transaction consists of different activities. By making these activities more efficient, value is created as enhanced efficiency reduces costs (Amit & Zott, 2001). According to Amit & Zott (2001), efficiency enhancements relative to offline businesses can be realized in a number of ways. One is by reducing information asymmetries between buyers and sellers by delivering up-to-date and comprehensive information. This case does not necessarily provides more up-to-date or comprehensive information compared to the same website without the search suggestion functionality. When a product is added, it can be found immediately on the website but behavioral data regarding this specific product will be analyzed during the next night (Basjes & Mathijssen, 2012). As a result information originating from this analysis cannot be considered up-to-date compared to the website before the search suggestion functionality. Hence, in this sense, lacks added value. This can be best described using an example: when Amy Winehouse died on July 23th, 2011, many visitors start buying her music on the website of bol.com. Since behavioral data representing this behavior was not analyzed until the next night, the next morning the search suggestion functionality start displaying “Amy Winehouse” when entering an “A” in the search field.

Another way is reducing customer’s search costs by improving the search activity’s efficiency. This has been studied marketwise multiple times and can even cause the creation of a new market or the destruction of an existing one (Bakos, 1997, p. 1677; Lucking-Reiley, 2001). For example, electronic marketplaces are easier to explore for customers since more information is available making it possible to search for the cheapest supplier. Yet, searching for a product within a shop (this can either be a traditional retailer or an online retailer) is also part of the transaction as a whole and by improving the efficiency of this process, according to the transaction cost economics theory, creates value too as it reduces costs. The great number of users of the search suggestion functionality and, more importantly, the increase in sales it caused, indicates the advantages it offers, hence supports the first hypothesis.

**H2:** By utilizing an exclusive set of resources through a strategic network, making a specific big data analytics case possible, the amount buyers are willing to spend increases;

No strategic network is created for the sake of this functionality. Although software and hardware is used from a specific vendor, no exclusive arrangement was made. Also maintenance, support and inquiries are handled internally. Since all data is collected from

---

8 Costs does not always refer to money but can also refer to other resources such as time as The Free Dictionary indicates: “The expenditure of something, such as time or labor, necessary for the attainment of a goal” (retrieved from http://www.thefreedictionary.com/costs at August 7th, 2012).
their own sales channel, no long-term agreements regarding data sources were made. Hence, this case is not supporting this hypothesis.

\[ H_3: \text{Big data analytics creates, supports the creation or improves products, services, markets or channels leading to significant advantages for customers.} \]

Innovation is, according to Schumpeterian’s innovation theory and the resource-based view described in section 5.1.2 and 5.1.3 respectively, a source of value creation (Amit & Zott, 2001, p. 497) as it acts as the foundation of new products, services, production methods, etc. The search suggestion functionality is a result of the combination of different resources such as the Hadoop cluster, IT staff, applications but more important: the behavioral data created as visitors navigate through bol.com’s website.

According to Amit & Zott (2001, p. 508), value creation occurs in traditional businesses when introducing new products or services, new methods of producing, distributing, or when tapping new markets. With e-businesses, in addition to these, value creation also occurs when innovating in the way they do business, that is, in the structure of transactions (Amit & Zott, 2001, p. 508). Previously is described how the search suggestion functionality improved the efficiency of the (internal) search activity, hence leads to enhanced transaction efficiency. Again, the great number of users and the increase in sales indicate the advantages and added value for both the customer and organization, hence this hypothesis is supported.

6.2.5. Conclusion

This case represents a big data analytics case as many will recognize: the data involved is simply put “big”. It requires the need of a Hadoop cluster as the volume of data to be analyzed, within the required timeframe of a night, requires the need of distributed data analysis.

This case study showed that big data analytics acts as a value creation driver as it improves transaction efficiency of the transaction between bol.com and its customers by improving the search activity within the shop, also part of the transaction as a whole. It also confirms that big data analytics supports the creation of new or improved services, hence innovation, with significant advantages for customers. The number of users indicated its popularity and the increase in sales its value for the organization (Basjes & Mathijssen, 2012). Hence, this case supports both the first and third hypothesis and fails to support the second hypothesis.
7. Conclusions & beyond

In this chapter the research conclusions are presented in which the main research question will be answered. In addition, additional observations during this study are discussed and an overview of topics for further research, are presented.

7.1. Research conclusions

In chapter 3 the main research question has been defined as follows:

\[\text{Which ways of value creation can we identify when an organization analyzes data referred as big data and hence, is succeeding in doing big data analytics?}\]

The answer to this question is: big data analytics might create value in two ways: by improving transaction efficiency and by supporting innovation leading to new or improved products and services.

In order to provide above answer, the definition of big data analytics is provided in chapter 4. Due to the overall change of data we generate, collect and store digitally, a potential source of information, knowledge and even wisdom is created, as described in section 1.1. This change becomes explicit when studying this change using the 3 Vs: volume, variety and velocity. Change in terms of volume is most commonly used to describe big data but variety and velocity are causes too, also help driving the volume characteristic. As a result of this evolution of data, new technologies are needed to handle this amount of data. In addition, realtime data analytics and the analysis of multi-structured data sources are receiving an increasing amount of interest, also driving the need for high performance solutions. Yet, these solutions must be economically feasible.

To study the effect of big data analytics on organizations' value creation, different value creation theories are examined in section 5.1. According to these theories, organizational value originates from different sources. By aggregating these sources with respect to this study, different hypotheses are defined. Each hypothesis indicates how an organization might create additional value when succeeding in doing big data analytics.

The first hypothesis argues that value can be created when big data analytics improves transaction efficiency leading to a reduction of costs or an increase in sales. This study shows that big data analytics indeed can result in improved transaction efficiency. bol.com’s search suggestion functionality increased the ease of use of bol.com’s internal search activity used by customers to find products. It made it easier for customers to find products matching their criteria resulting in an increase of sales (Basjes & Mathijssen, 2012).

The second hypothesis argues that by utilizing an exclusive set of resources through a strategic network, making a specific big data analytics case possible, the amount buyers are willing to spend increases. Although section 5.4.3 implies this assumption by arguing that a strategic network opens a set of resources, which in turn make big data analytics valuable as it excludes other organizations from utilizing the same resources, both cases examined in chapter 6 did not support this hypothesis.

The latter hypothesis, argues that big data analytics creates, supports the creation or improves products, services, markets or channels leading to significant advantages for
customers. The study shows that big data analytics is indeed an innovation that acts as the source of a new or improved product (X FACTOR Heartbeat) or service (bol.com’s search suggestion functionality), both offering significant advantages for the customer.

7.2. Achieved goals and relevance

In chapter 3, three different types of goals are defined: personal goals, intellectual goals and practical goals. Starting with the first, I said that by doing this study, I wished to get knowledge about the state of affairs in the “data as an asset” topic. By visiting conferences, speaking with people active in this field, and of course, by conducting this study, I did. One of the things I found interesting is the fact that organizations are aware of the possible value of their data. Yet, they are struggling to extract this value in a feasibly way. Moreover, it appears that mainly large organizations are interested in big data analytics. This can have multiple reasons: they simply may have more (internal) data or they have more resources (e.g. budget in R&D) to experiment with new technologies.

Intellectual goals are goals focussing on understanding something. These goals are very important regarding this study, as they indicate the level of contribution to different (scientific) fields. This study contributed in several ways, also indicating its relevance. First of all, by examining existing literature about big data (analytics), a comprehensive definition is provided. Moreover, this definition was extended by defining explicit characteristics making the current “big data analytics” more discernible. This improves the generalisability of this and future studies and helps “solving” the subjectiveness–issue of big data analytics as described in section 3.4.1. Next, this study also contributes to existing literature about value creation. Different theories on value creation were studied and linked resulting in aggregated hypotheses which can be used to study other possible value creation drivers too. More importantly, this study contributes to computer sciences, information sciences and business sciences by linking value creation with information and different technical subjects. I believe this helps bridging IT and business, an issue often discussed under the name of “business–IT alignment” (Luftman, 2004, p. 99).

Practical goals focusses on accomplishing something and hence, more than the previous goals, focusses on practical implications. This study is useful for organizations interested in big data analytics. First of all, this thesis can act as an eye-opener as it shows the potential of big data analytics in relation to traditional data analytics, especially when including exotic data sources. It also introduces different techniques which can be used to analyze these data sources. More importantly, it helps organizations defining expectations on how big data analytics will affect their organization, especially their value chain. For example, this study shows the added value of big data analytics for both itself and its customers as it can improve transaction efficiency and supports the creation of new or improved products and services. This, including the examples of chapter 5 and the examination of the cases in chapter 6, gives organizations the ability to reflect on their own (data) situation and enables them to identify whether big data analytics will be valuable to them.
7.3. Additional observation notes

7.3.1. Big data’s hype factor

During several meetings, seminars and conferences one thing became clear: big data analytics is still relatively new and unexplored. Many organizations in different sectors are very interested in this subject and see a potential source of value in their data, external data or a combination of both. Due to a lack of practical examples and knowhow, for many organizations it stays by “seeing” instead of “experiencing” this source of value. This was the reason why “only” two cases were studied in this study: many big data analytics implementations did not satisfy the requirement described in section 3.3.3: it already needed to be integrated in the organization in a functional way. Many organizations visited conferences because their competitors did or found out that their competitors already started experimenting with big data analytics. In order to prevent the competitor to take advantage of big data analytics solely, they were forced to take action themselves. This might be a reason for the explosion of the big data phenomenon in terms of global attention. This caused another hype: both large IT organizations and entrepreneurs started activities focussing on big data analytics by starting new initiatives. This in turn lead to even more attention making big data become a hype in different fields and will continue to do so (Burgel, 2012).

7.3.2. Big data analytics as an experiment

It was also noticeable that organizations start small. This has multiple reasons. For example, bol.com initiates IT projects using the Scrum method: start small and improve iterative by incremental enhancements, this is part of their policy (Basjes & Mathijssen, 2012). Also for organizations the added value and the impact on the organization is not completely clear. The Rabobank is a good example of this. In January they started experimenting with Hadoop and its possibilities by running different experiments on roughly a month of production data, consisting of transaction data. Since the Hadoop cluster is concealed and secured from external influences, experimenting is possible with less legal obstacles. This makes it possible to study its capabilities and future implications when the cluster will be an integral part of the organization in the future (Kuil, 2012). It appears that organizations are, consciously or unconsciously, following Forrester’s (Hopkins & Evelson, 2012) and Computable’s (Sluis & undefined author, 2012) advice by starting small. The third identified reason is that, in order to use new technologies such as Hadoop and extract valuable insights from datasets identified as big data, expertise is needed. This is also mentioned by Manyika (2011):

“A shortage of people with skills necessary to take advantage of the insights that large datasets generate is one of the most important constraints on an organization's ability to capture the potential from big data.” (Manyika et al., 2011)

By experimenting with big data on small scale and in an separate environment, the organization acquires expertise, as is the case with bol.com and Rabobank too (Basjes & Mathijssen, 2012) (Kuil, 2012). Another example is Red Data, an organization active in
the fields of data management, BI and sourcing. Red Data offers a Hadoop cluster to its customers. This gives Red Data the opportunity to acquire expertise in this new technology. (Burgel, 2012).

7.3.3. **Realtime data analytics**

Another observation is the rising interest in the realtime aspect of big data analytics. According to Basjes (2012), the “next thing” in the field is realtime analysis. Many big data implementations are batch-oriented, like the bol.com case described in section 6.2. When behavioral data is analyzed in realtime, search suggestions are up-to-date. This results in more potential value as delivering up-to-date information reduces information asymmetry (Amit & Zott, 2001, p. 503), hence to more efficient transactions. In section 4.2.2 the potential loss of value when data becomes older, is explained. Also Kuil (2012) mentions that speed is one of the greatest advantages of new big data analytics technologies. It enables Rabobank to extend current analysis tasks by simply adding more transaction data (e.g. historical data) as the analysis task will be finished much faster than before. For example, future rogue bank transactions can be prevented before money is actually transferred by checking its validity.

7.4. **Limitations**

This study knows different limitations which are important to mention as further research and organizations might depend on the outcomes of this study. Next, these limitations are described in a random order.

7.4.1. **Subjective theme**

This thesis focusses on a phenomenon which is often considered a hype as described in section 7.3.1. Partly for this reason, big data is a loosely-defined term and prone to context and understanding issues as mentioned in section 3.4.1. This section argues that by using characteristics defining big data analytics, this limitation will be tried to avoid. Although these characteristics were defined after a comprehensive literature study on this topic, both researchers and organization may use a slightly other definition of big data analytics as it is subjective topic as described in section 3.3.1. This might lead to incompatible results as the results of this study rests on the definition provided in this thesis.

7.4.2. **Limited number of (Dutch) cases**

Only two cases were studied increasing the chance of invalid results as partly mentioned in section 3.4.3. In addition, the reasons described in section 7.3.1 are in play: big data analytics is still relatively new and unexplored. In the worst case, this led to a type I error, in which a hypothesis is supported but in fact, should not (Lazar et al., 2009). This can have dramatic consequences for organizations investing money in the implementation of a big data analytics solution, as they may rely on the outcomes of this study. By using multiple data sources as much as possible and by asking constant feedback, this situation was attempted to avoid. In addition, the cases studied were located in the Netherlands, hence, the outcomes of this study might only represent the Dutch market. Yet, it is likely that the results are replicable in other countries as the
structure of data and the technologies to analyze this data, are not bound to specific countries.

7.4.3. Qualitative study based on a limited number of data sources
As mentioned in the previous section, as much data sources as possible were used. However, this still may be considered limited when looking at the impact of some facts. For example, the first hypothesis (big data analytics’ impact on transaction efficiency) is only supported by bol.com’s search suggestion case by relying on only one data source arguing that this specific case led to an increase in sales (Basjes & Mathijssen, 2012). By using more data sources, the limitation described in section 3.4.2, arguing that organizations can be mysterious or taciturn about facts and other information, is less present as it decreases the influence of a specific fact. Extending this topic to the next section, this study should be repeated in the future (when more cases are known), possibly using quantitative methods resulting in statistical signification confirming (or not) the outcomes of this study.

7.5. Further research

7.5.1. Focus on specific markets or types of big data analytics
By studying more (international) cases, the results will likely be more generic, hence could improve the validity of the results. Another way to improve the level of generalization is by focussing on a specific market such as health care. This is particularly valuable for organizations working in that specific market. Also, further research might focus on a specific part of big data analytics such as realtime data analytics or the analysis of a specific sort of multi-structured data. Especially the first is interesting, as it receives an increasing amount of attention as mentioned in the section 7.3.3.

7.5.2. Repeat study in the future
As big data analytics is considered new and relatively unexplored, the same study should be done in a few years to see how organizations currently struggling with conceptualization and implementation, have succeeded (or not) in turning their big data chaos into big data opportunities. The same design, characteristics of big data analytics and hypotheses should be used in order to create comparable results.

7.5.3. Big data analytics’ impact on competitive advantage and beyond
In this study, the effects of big data analytics on organization’s value creation is studied. Value creation and competitive advantage are closely related as “competitive advantage equals the difference between the value created by the company and the potential value created by its competitors" (Spulber, 2009, p. 233). As technologies are considered doubtful regarding their added value according to the resource–based view as discussed in section 5.3, and data is becoming more publicly accessible as discussed in section 5.4, you might argue how this “difference” is expressed in the “big data analytics as a value creation driver”–topic. Mainly for this reason, the second hypothesis was formed: partnering through a strategic network, excluding competitors from utilizing the same benefits, could lead to added value, hence to competitive advantage. Yet, as this study
failed to support this hypothesis, the effect of big data analytics on competitive advantage is considered doubtful and hence, requires additional research. As mentioned in section 7.3.1, many organizations visited relevant conferences as they discovered that competitors already were experimenting with this theme. This phenomenon makes future research even more interesting.

Furthermore, as big data analytics affects organization’s value creation and possibly affects competitive advantage too, it also can lead to reinvented or new business models, especially to leapfrog competitors (Voelpel, Leibold, & Tekie, 2004).

7.5.4. The value of a data source

With big data analytics, due to improved technologies capable of storing and handling unstructured and multi-structured data, and the increased capacity, more data sources can be analyzed (Burgel, 2012). Yet, which data sources are most valuable and how can we identify whether it is constantly improving the quality of the outcomes of the analysis? Hence, studying the value of a data source is interested as it gives organizations the ability to decide if investing in the implementation of a particular data source is worth the effort. A good starting point is Boyd & Crawford’s (2011) article arguing that bigger data is not always better data when analyzing data sets classified as big data.
8. Bibliography


Agrawal, T. (2012). 1/3 of US Mid-Market Businesses Using Business Intelligence are Interested in Big Data Analytics; Lack of Expertise is Main Barrier. Techaisle.


Thanks, YouTube community, for two BIG gifts on our sixth birthday! (2012, November 25). Thanks, YouTube community, for two BIG gifts on our sixth birthday! YouTube. Retrieved from http://youtube-global.blogspot.nl/2011/05/thanks-youtube-community-for-two-big.html


