Professional MBA
Entrepreneurship & Innovation

Status Quo of Big Data analysis in small and medium size enterprises in Austria

A Master's Thesis submitted for the degree of
"Master of Business Administration"

supervised by
Prof. Dr. Christian Lüthje

DI Elmar Schamp

9326425

Vienna, 30.06.2016

TU WIEN CONTINUING EDUCATION CENTER

WU EXECUTIVE ACADEMY

Affidavit

I, **ELMAR SCHAMP**, hereby declare

1. that I am the sole author of the present Master's Thesis, "STATUS QUO OF BIG DATA ANALYSIS IN SMALL AND MEDIUM SIZE ENTERPRISES IN AUSTRIA", 56 pages, bound, and that I have not used any source or tool other than those referenced or any other illicit aid or tool, and
2. that I have not prior to this date submitted this Master's Thesis as an examination paper in any form in Austria or abroad.

Vienna, 30.06.2016

Signature

ABSTRACT

Commentators around the world have emphasized the importance of data with statements like ‘data is the oil of the next century’ or ‘data is the new oil’ for the past decade. Has it come true? If data is so important, how important is Big Data?

What is Big Data and what is the status of Big Data in Austria? This thesis is aimed to analyze the status quo of Big Data for small and medium sized enterprises in Austria. Therefore, the value chain of Big Data is used to define a basement for each single step and oppose this to actual findings that were extracted from surveys. The difference between traditional data analysis and Big Data Analytics are blurred and especially for SMEs irrelevant. It is not so much about the magnitude of data volume, velocity of processing and the variety of data (3Vs) but rather the value that can be generated out of data. New techniques like predictive analysis, machine learning and network analysis and technologies like analytical databases and parallel programming models are not yet in focus for most SMEs, although in terms of the best solution for a certain problem the enterprises will use the best model available. Use cases and success stories for Big Data in the manufacturing field like smart production lines and predicting service intervals are not sufficiently sound for SMEs to embark on the Big Data paradigm. The major challenges for Big Data initiatives are a lack of knowledge within the companies, a lack of skilled workforce, privacy concerns and data security issues. The latter problems should be addressed in a data strategy which should be developed upfront a Big Data initiative and has to be seen interconnected with the business strategy.

To avoid missing business opportunities, SMEs should carefully analyze the potential impact that a deeper analysis of data could have on their customers and processes regardless of where the data will be acquired, how they will be stored or processed. For these technical issues, SMEs can easily find the right service facilitated by a Big Data platform service provider. Ventures that are not yet dependent on data like companies were on oil years ago certainly have further potential to improve productivity and service quality for their customers by starting Big Data initiatives.

Table of contents

List of tables

List of figures

1 Introduction

In this chapter, an outline of the thesis is drawn and a brief description of Big Data provided. A context for understanding the current development regarding Big Data for small and medium enterprises is created.

1.1 Context of the thesis

Big Data is a very prominent password that is currently used for almost every task related to storing and analyzing big amounts of data. Big Data has developed over the past ten years and it is one of the growth drivers of the current IT industry. Based upon success stories like Google, Amazon or Facebook, several CEOs are increasingly aware that the large amount of data at their hand could potentially be very valuable. They try to find ways how to monetize the information that is available within their company and is provided by users, sometimes for free. Furthermore, CEOs are frightened by the possibility that a competitor is faster in finding a competitive advantage thanks to Big Data or that a start-up is suddenly starting to disrupt the industry having found a new way of doing business based upon the usage of new technology. Even politicians preach the golden age of Big Data, claiming that data is the oil of the next century and that this will bring prosperity to the country. There are several reasons why Big Data is one of the most important topics in recent years; however, is this also valid for medium-sized or small enterprises? Why don´t we read much in the newspaper about companies that grow their business on the results of data analytics? The motivation for this paper is to analyze the status quo of Big Data in Austria especially for small and medium-sized companies in the manufacturing industry and identify the challenges and problems that hinder companies from using this new technology.

1.2 Aim and objectives

Big Data has become a major topic in recent years, being seen as one of the fundamental growth drivers of the next decade. However, is this a valid assumption or is it simply a topic that is very popular because the producers of such tools are generating hype around this password? The overall aim of this master thesis is to ascertain the status quo of the Big Data paradigm in Austria, especially for small and medium-sized companies. However, what is Big Data and what is the major difference from traditional data processing like we used to do it? Is Big Data only applicable to large companies or is it also a driver for small businesses? One of the goals of this master thesis is to obtain some clarity concerning these questions based upon existing studies and papers. Based upon definitions collected from primarily economics literature - as it should not be a technical study - available surveys are used to evaluate the status quo of Austrian-based companies regarding Big Data aspects. If necessary, the focus for certain topics is limited to the manufacturing field.

What is Big Data?

The phrase Big Data has been used very frequently and the simple question 'What is Big Data?' is difficult to answer. This thesis provides a brief overview of possible definitions and determines the Big Data term in terms of how it has to be understood for this thesis. The Big Data universe is explained along the steps required to analyze data. The very broad spectrum of Big Data methods, technologies and techniques are discussed without necessarily being exhaustive. The idea is to provide a high-level Big Data overview, which subsequently sets the stage for the comparison of publicly available surveys.

What is the current status of Big Data in Austria?

The idea was to determine the actual status of Austrian companies using Big Data based upon a survey or with interviews among Austrian SMEs. A web research revealed that several studies exist regarding this topic and that the response rate especially from SMEs was very low. In fact, all of these surveys deal with different parts of Big Data, different geographical regions and different target groups. To make them somehow comparable, the results of these studies had to be assigned to a holistic definition of Big Data. With this approach, it is possible to compare the results of different studies and reflect the overall result to a literature-based plan scenario. Furthermore, the findings have to be put into perspective in terms of the enterprise size of survey participants and studies with worldwide aspects were put into perspective with more regional and SME related statements. This should help to somehow verify and explain certain aspects of this complex topic.

Is Big Data only applicable to big companies?

One theorem for this thesis is that SMEs are not yet using Big Data, although generally the topic of Big Data is relevant for SMEs. However, which parts of Big Data are relevant for SMEs? Is Big Data a product that SMEs can use to enhance their business? Is it the data volume that motivates them to switch their IT systems? Is it the speed of decision-making that their organizations cannot or will not be able to cope with? Alternatively, has Big Data changed data analysis radically, whereby new methods will be the reasons for SMEs to use this new technology? Beside the questionnaires, a basic research of Big Data use cases and success stories should help to answer these questions. This part of the thesis will focus on SMEs in the manufacturing field.

What are the main challenges and hurdles?

A basic question that should also be investigated concerns the hurdles in terms of why Big Data initiatives are not implemented. Which challenges are listed in the literature and do they correspond with the challenges that the surveys reveal? What are the major success factors for a Big Data project?

Which products are available for SMEs?

However Big Data is defined and whatever the status for Big Data for SMEs, the thesis should briefly investigate available solutions for SMEs. If a SME decides to run a Big Data initiative, is there one preferred solution or which solutions are available?

1.3 Course of investigation

For this study, a conceptual analyze method was chosen. A quantitative approach served the purpose of revealing the key facts of Big Data surveys regarding the goal of the thesis. With the purpose of the thesis in mind, literature related to the Big Data basics, surveys, use cases and products was collected and analyzed. The first step taken in the data collection process was to search academic databases and the World Wide Web to verify that sufficient articles are available to justify a literature review. The search revealed an acceptable amount of literature.

As source academic databases, the Vienna University of Technology Libraries (http://www.ub.tuwien.ac.at/eng), Vienna University of Economics and Business (http://www.wu.ac.at/en/library/), SpringerLink (http://link.springer.com) and Google Scholar (scholar.google.at) were used. Additionally, a review of online resources led to websites maintained by organizations that provide material with a high degree of relevance to this thesis and help to frame the research problem. For the purpose of this study, resources made available through the websites of BARC, Frauenhofer IAIS, IDG, Gartner Inc. (www.gartner.com) and the Information Systems Audit and Control Association (www.isaca.org) were very useful. Furthermore, the archive of Management Information

Systems Quarterly (MISQ), Journal of the Association for Information Systems, Journal of Information Technology and Harvard Business Review were conducted and several home pages and white papers of Big Data platforms or software providers like IBM, Microsoft, Tableau, SAS, CNC, Oracle, Google, Apache, Amazon, etc. were taken into account.

To describe the status quo, several surveys were conducted. The following table lists the surveys including the interviewee group, the performing organization and the year of publication.

Survey	Organization	Year of publication
BIG DATA SURVEY EUROPE	BARC	2013
Big Data Analytics	BARC	2014
Datenmanagement im Wandel	BARC	2014
Big Data Use Cases	BARC	2015
BIG DATA – Vorsprung durch Wissen INNOVATIONSPOTENZIALANALYSE	Frauenhofer IAIS	2012
The Deciding Factor: Big Data & Decision Making	CapGemini, Economist Intelligence Unit	2012
Big Data Survey	IDG Enterprise	2014
Big Data Go Big or Go Home?	Supply Chain Insights LLC	2012
Data Analytics im Mittelstand	Deloitte	2014
Big Data Executive Survey 2016	NewVantage Partners LLC	2016
BIG DATA UND DATA-DRIVEN BUSINESS FÜR KMU	DIGITAL NETWORKED DATA	2015
#Big Data in Austria	Köhler, Martin; Meir-Huber, Mario	2015
Big Data in kleinen und mittleren Unternehmen	Westfälische Wilhelms-Universität Münster	2015

Table 1 list of surveys considered

All these surveys have a common subject around Big Data, data analysis or decision support systems and still they are very different. Aspects like time of investigation, number of participants, different regions and the industry representation are listed in Table 2.

Survey	time of investigation	type of questionaire	participants	participating managers	Region	Industry Representation	size of enterprise
BIG DATA SURVEY EUROPE	second half of 2012	online user querstionnaire	274	274 paricipants 56% LOB 44% IT	64% DACH 26% France 11% Great Britain	very broad 21% service, 19% IT, 17% finance	24% <250 26% between 250 and 2500 50% > 2500 employees
Big Data Analytics	September to Decmber 2013	online questionnaire	373	373 paricipants 33% IT 29% finance & controlling 14% manager	DACH region	24% manufacturing, 21% service, 14% IT	28% < 250 43% between 250 and 5000 32% >5000
Datenmanagement im Wandel	July to September 2014	online user querstionnaire	341	341 participants 44% IT, 20% fincance & controlling 12% BI not within IT	DACH region	27% manufacturing 16% service, 15 finance and 14% IT	25% < 250 46% between 250 and 5000 31% >5000
Big Data Use Cases	December 2014 to February 2015	worldwide online user querstionnaire	559	559 participants 42% IT 17% finance & controlling 14% manager	37% DACH, 22% North America, 8% South Europe, 7% France, 7% Asien/Pacific.	16% IT, 15% manufacturing, 3% consulting, 8% retail	29% <250, 38% between 250 and 5000 32% >5000

Survey	time of investigation	type of questionaire	participants	participating managers	Region	Industry Representation	size of enterprise
BIG DATA – Vorsprung durch Wissen INNOVATIONSPOTENZIALANALYSE	October, November 2012	online questionnaire	80	18% manager, 17% R&D, 15% IT, 12% marketing, 10% sales	Germany	70% service, 18% retail, 12% manufacturing	37%<250, 45% between 250 and 1000 18%>1000
The Deciding Factor: Big Data & Decision Making	February 2012	Questionnaire supplemented with interviews	607	43% senior management 57% manager	Worldwide, 38% Europe, 28% North America, 25% Asia-Pacific	20 different financial service, technology, manufacturing healthcare	rather huge companies
Big Data Survey	3rd Quater 2013	online survey among audiences of CIO, Computerworld,…	750	54% IT manager 23% LOB manager		20% high tech, 13% financial, 8% healthcare, 8% government, …	47% SME < 1000 50% >1000
Big Data - Go Big or Go Home?	April through June 2012	interviews are supplemented by data from an online, quantitative research survey	53	53% supply chain 47% ITmanager	USA	25% consumer packaged goods 11% food and beverage 6% Automotive 6% Industrial remaining retail	26% <1000 25% between 1000 and 5000 49% > 5000

Survey	time of investigation	type of questionaire	participants	participating managers	Region	Industry Representation	size of enterprise
Data Analytics im Mittelstand	2014	Questionnaire supplemented with interviews	70			Germany	SME <3000 (av. 400)
Big Data Executive Survey 2016	2016		44	36% C-Executives 29% Head of Big Data 12% Head of Analytics 13% Senior Technologist	44 of Fortune 1000 USA based global companies	73% Financial services 18%Life sciences 9%Other sectors	100% >5000
BIG DATA UND DATA-DRIVEN BUSINESS FÜR KMU	2015	individual interviews			Austria		
#Big Data in Austria	Nov.13		150	IT managers	Austria		
Big Data in kleinen und mittleren Unternehmen	Apri and May 2014	online questionnaire supplemented with interviews	24				64% < 250 33% > 250

Table 2 Meta data of surveys

Due to the focus of this thesis on SMEs, the size of the company is an important differentiator. The target groups of the surveys underline that the Big Data paradigm is primarily relevant for large companies at present. Indeed, only three studies focus on SMEs and for those surveys where the participation rate is disclosed, it has to be stated that the number of participants is not representative. Therefore, is it necessary to consider findings regardless of the size of the organization and – where applicable – conclusions for smaller enterprises have to be derived. Some survey results are separated according to the venture size, although more often the results have to be assessed based upon the participation rate of small enterprises.

This meta-analysis has to be conducted based upon the summary statistics and cannot use the raw data. This makes it difficult to compare certain findings with each other. This is particularly relevant for all sorts of categories; for example, use case categories. These categories are listed as the result of a question or the interviewee was asked to rank categories, whereby the ranking is the result. Sometimes the question is not disclosed by the author and in some instances it is unclear whether the list of values was part of the questionnaire and the respondent had to pick or rank single or more items or if the list of values was generated by the author based upon answers to an open question.

It has to be disclosed that most of the surveys are sponsored by the government or a group of Big Data software developers. However, this fact is fully disclosed within the studies and thanks to this sponsoring especially the BARC studies were able to reach a broader interview base.

The BARC Big Data Survey Europe study addresses European companies and is focused on Big Data topics like the central drivers for Bit Data, usage of Big Data, Big Data problems and which types of data and technologies are used? This study is a fundamental pillow of the meta-analysis. The BARC Big Data Analytics study emphasizes Big Data applications. Questions about the implementation status of Big Data initiatives, Big Data goals and the technical implementation are raised. The BARC study “Datenmanagement im Wandel” concentrates on the change from Data Warehousing (DWH) to Big Data and questions the current business intelligence (BI) situation in the organization. The BARC Big Data use cases study like the Big Data Survey Europe focuses on the usage of Big Data. This worldwide study helps to put regional aspects into perspective. The BARC studies are based on so called user questionnaires. The authors did not explain which users were invited to participate but it is assumed that they refer to users of products of the survey sponsors and not to only to Big Data users, otherwise all enterprises would have already had Big Data initiatives. The survey “BIG DATA – Vorsprung durch Wissen“ was realized by Frauehnofer institute in October and November 2012. Based on a desk-research and individual interviews an online questionnaire was executed. CapGemini commissioned the Economist Intelligence Unit to conduct a survey to find out more about how organizations are using Big Data, where and how Big Data is making a difference, and how Big Data will be used. IDG´s Big Data Survey is a worldwide study that gives a very good high-level overview about Big Data related issues. Although the survey claims to be worldwide it is not disclosed where the participants come from which makes it impossible to use the findings for regional discussions. The “Big Data – Go Big or Go Home” survey is based on discussions with 32 industry leaders currently working on Supply Chain Big Data initiatives. Beside insides from the supply chain perspective on Big Data the survey contributes with interesting inputs concerning data sources and Big Data techniques. “Data Analytics im Mittelstand” is a survey that has the focus on decision making. For this thesis only some aspects relating to general Big Data issues could be used. The Big Data Executive survey 2016 from New Vantage Partners is a high level summary of Big Data aspects but among 44 of the Fortune 1000 companies. As the focus for this thesis is on SMEs in Austria this survey has to be seen as a destination board to a possible future scenario. “Big Data und Data-driven business für KMU” is an Austrian based survey. The number of participants is not disclosed. The findings are not representative and only some aspects are used in the general Big Data chapter. “#big Data in Austria” is a study sponsored by the Austrian government and the results that are used for this thesis are based on a survey executed by IDC Austria in November 2012 and 2013. The proportion of participating SEMs could be determined. “Big Data in kleinen und mittleren Unternehmen“ is a survey among German based SMEs with 24 respondents. Even though the number of respondents is not representative the findings are very helpful as the survey was conducted in 2015 and the target group is very similar to the target group for this thesis.

Survey	General				value chain				Implementation				Budget / Investment	decision making	Vendor
	3Vs	goals	use cases	Initiatives	generation and acquisition	data recording and integration	data processing	data interpretation	Success factor	Challanges	Data strategy	Organisation			
BIG DATA SURVEY EUROPE	x	x	x		x					x	x	x			x
Big Data Analytics		x		x			x			x		x			
Datenmanagement im Wandel	x										x				
Big Data Use Cases		x	x	x			x	x		x	x	x	x		
BIG DATA – Vorsprung durch Wissen INNOVATIONSPOTENZIALANALYSE	x	x	x		x	x	x			x					
The Deciding Factor: Big Data & Decision Making	x			x	x	x				x		x		x	
Big Data Survey	x			x					x	x					x
Big Data - Go Big or Go Home?	x		x	x	x				x	x					
Data Analytics im Mittelstand	x	x										x	x	x	
Big Data Executive Survey 2016	x	x		x	x				x	x		x	x		
BIG DATA UND DATA-DRIVEN BUSINESS FÜR KMU	x			x						x					
#Big Data in Austria		x	x	x						x					x
Big Data in kleinen und mittleren Unternehmen	x	x	x	x			x			x		x			

Table 3 Surveys contribute to listed Big Data fields

To analyze and compare these studies in a systematic way, a blueprint with the survey hotspots was defined (see Table 2). Almost all surveys include questions about Big Data or data analytics that are relatively general and relate to the 3Vs, the aims and goals of Big Data, the use cases or the status of Big Data initiatives, including whether they are already implemented, planned or considered not relevant for the ventures. Surveys that are more focused on Big Data provide a deeper insight into the technical status quo. Comparing this with the Big Data value chain (see 2.2), it can be highlighted that the surveys are more focused on the data recording, integration and processing rather than the data interpretation and visualization. Almost all surveys present results concerning the challenges and hurdles that can be faced when implementing a Big Data initiative. The positive side – namely the success factors of Big Data projects – were less commonly cited. Special types of success factors include the availability of a data strategy and the support that a Big Data initiative has within the organization. Only three surveys deal with the aspects of software vendors or consultants for Big Data.

The results of the surveys of these different categories are outlined, compared and discussed in the following chapters. For every category, a literature-based definition helps to put the findings into perspective.

1.4 Structure of the thesis

The thesis is divided into four chapters with Chapter 1 being the introduction. The remaining chapters are briefly described below.

Chapter 2 deals with the Big Data paradigm and provides a definition for Big Data, as it has to be understood for this thesis, and explores technologies and techniques that are commonly summarized within the Big Data universe. For the purpose of matching Big Data and small and medium-sized companies a comparison of small and Big Data is drawn.

Chapter 3 extends the literature analysis to general aspects of Big Data and details the results of different surveys regarding the criteria that define Big Data, the goals of Big Data and some use cases to illustrate the advantages of Big Data. Findings of several surveys are combined and compared to provide a big picture of these aspects of Big Data. An important question for Big Data is which data sources are available and a single section is dedicated to summarize the findings of the surveys referencing the different types of data. Finally the usage of Big Data techniques and technologies is investigated.

Chapter 4 relates to implementing Big Data. Success factors as well as challenges and hurdles for Big Data initiatives are carved out of the available literature and are compared to findings of different surveys.

Chapter 5 gives a brief oversight of the Austrian Big Data market and Big Data platforms appropriate for SMEs are described.

Chapter 6 summarizes the key findings of the research and contains conclusions.

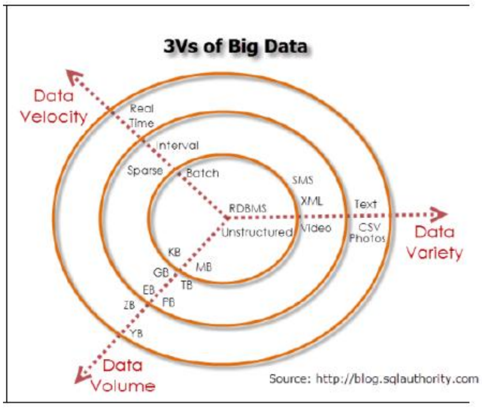
2 Big Data

In the following chapter the term Big Data is defined for this thesis and the Big Data universe is fragmented or detailed by using the Big Data value chain.

2.1 Defining Big Data for this thesis

The term Big Data was coined in a paper by Doug Laney as early as 2001. Doug Laney, at that time an analyst of META, defined challenges and opportunities brought about by increased data with a 3Vs model, i.e. the increase of volume (great volume), variety (various modalities) and velocity (rapid generation) in a research report (Laney D, 2001).

Although such a model was not originally used to define Big Data it was used to describe the characteristics of Big Data. This model was over the time extended with the term value and this was widely used as the 4Vs model . Such a 4V definition highlights the meaning and necessity of Big Data, i.e. exploring the huge hidden values and indicates the most critical problem in Big Data, which is how to discover values from datasets with an enormous scale, various types and rapid generation. As Jay Parikh, Deputy Chief Engineer of Facebook, said, “You could only own a bunch of data other than big data if you do not utilize the collected data” (Chen, et al., 2014, p. 173) . Sometimes the fourth V - namely value - is substituted with the term veracity. Veracity refers to the level of reliability associated with certain types of data (IBM 4Vs). However, such a description of the term does not really define what Big Data is. It somehow sets four dimensions to describe characteristics of data sets. NESSI (Networked European Software and Services Initiative) specifies in a White Paper in December 2012 that “Big Data is a term encompassing the use of techniques to capture, process, analyze and visualize potentially large datasets in a reasonable timeframe not accessible to standard IT technologies. By extension, the platform, tools and software used for this purpose are collectively called Big Data technologies” (NESSI, 2012). The Business Application Research Center (BARC) in Berlin also uses a more holistic approach to define Big Data. In their opinion Big Data should include methods and technologies for high scalable recording, storing and analyzing unstructured data (BARC, 2015). As there are different definitions rather than one precise



Source: http://blog.sqlauthority.com

Figure 1 Illustration of the 3Vs of Big Data

definition of Big Data, an alternative perspective is to view Big Data more as the approach about how to deal with data rather than how much data is available or which types it contains. Finlay wrote about four tenets of this idea: 1. Seek, 2. Store, 3. Analyze, 4. Act (Finlay, 2014, p. 14), which reflects the basic idea how companies have to deal with Big Data. It is necessary to proactively search for and obtain new data. Therefore, by analyzing the data, it is usually necessary to store them, whereby organizations try to obtain value out of the data and in best case the system will act based upon this outcome. These tenets can also be seen as the necessary phases of the process Big Data. For this thesis, an adequate definition for Big Data is given by a broad agreement of authors of Wikipedia:

"Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying and information privacy. The term often refers simply to the use of predictive analytics or certain other advanced methods to extract value from data, and seldom to a particular size of data set." (Wikipedia, 2016)

It would be wrong to assume that survey participants had this definition in mind when responding to Big Data topics, e.g. 60% of interviewees for a survey of Austrian SMEs stated that the term "Big Data" is not sufficiently described for them, although they presumably answered the questions related to Big Data (Digital Networked Data, 2015, p. 7) and 20% of interviewees of the Frauenhofer IAIS have not heard of the term to date (Frauenhofer IAIS, 2012, p. 44). Out of 24 SMEs, more than one-third evaluated themselves concerning their own Big Data knowledge with 0 or 1, where 0 stands for 'no knowledge' and 5 for 'very good knowledge', and almost three-quarters with below 4 (Vossen, et al., 2015, p. 41).

2.2 Big Data´s value chain

The Big Data process or the value chain of Big Data is a good way to illustrate on a high level the necessary steps that have to be taken to implement a Big Data project (Chen, et al., 2014) (Bajaj & Ramteke, 2014, p. 1878 f.). It can be seen as part of a project but the value chain can also be used in terms of an external view, like outlined by BITCOM (BITCOM, 2013), as a change for ventures to offer services and products in this special area.

2.2.1 Data generation and acquisition

Nowadays data is generated everywhere. The task of acquiring them is more a technical issue and a question of entitlement. In an organization several information about their customer base and their suppliers like purchase details, billing information, e-mails, server logs, texts, transcripts of phone calls, complaint letters, notes taken by staff in branch, credit reports, GPS data, etc., are stored in a database. All this data gathered for three million customers in the last five years can be considered as Big Data. The data is available and generally the company is allowed to use the data (Chen, et al., 2014, p. 15)

At present, main sources of Big Data are the operation and trading information in enterprises, logistic and sensing information in the Internet of Things (IoT), human interaction information and position information in the Internet world, and data generated in scientific research, etc. (Chen, et al., 2014, p. 179). A huge amount of content is nowadays produced in the web by anybody willing to and is not related to a specific company but is closely related to people's daily life. Such human generated content reveals useful information such as habits and hobbies of users that enables to forecast users' behaviors and detect emotional moods. Such data sources include sensors, videos, clickstreams, and/or all other available data sources.

Significant amounts of data are generated in production lines that are increasingly equipped with sensors to measure single steps of the production. The content generation of users of social networks is currently a steady stream of information waiting to be analyzed, although the acquisition is a little different in respect to the privacy issue and the content is very "noisy" as the real information is difficult to extract. The Big Data approach would be to store the raw data. However, sometimes it is not feasible to store all raw data and it is useful to filter and compress the data by orders of magnitude but the challenge is to define these filters in such a way that they do not discard useful information (Bajaj & Ramteke, 2014).

Data collection techniques are used to collect raw data from the basic data sources and the most common are log files, Sensing and Network data. Log files are files written by IT systems like a web server with the aim of recording activities mostly in ASCII text file formats for subsequent analysis, e.g. log file formats are public log file format (NCSA), expanded log format (W3C), and IIS log format (Microsoft). Sensing nowadays is everywhere, the Internet of Things (IoT) and Industry 4.0 are based upon the idea that devices, e.g. in a production line or small wearable devices, submit information to the data collection point sometimes in the cloud and the collection of this information is the origin for Big Data analytics. Sensory data may be classified as sound wave, voice, vibration, automobile, chemical, current, weather, pressure, temperature, etc. wired sensor network is a convenient solution to acquire related information. In addition to the measurement result the position and the time of the measurement is gathered and sometimes this information is the most accurate and valuable one. For acquiring network data more sophisticated methods have to be used like a combination of web crawler, word segmentation system, task system and index system, etc. A web crawler works almost in the same way as human do when searching for information. It digs from one web page identified by a uniform resource locator (URL) to linked web pages and stores all these retrieved URLs and sequences them. These relations are used in search engines or for web caching (Chen, et al., 2014, p. 181)

New business opportunities also from small and medium-sized enterprises emerge by digitizing data that was hitherto only available in analog form, by generating or gathering general data (like weather, traffic or demographic data) or providing services for data acquisition (BITCOM, 2013).

2.2.2 Data recording, data integration, data quality management

Once the raw data are collected an efficient transmission mechanism is essential to send the data to a proper storage management system that is capable of storing very huge amount of data to support different analytical applications (Chen, et al., 2014, p. 180). Such applications normally require a high rate of redundancy over different nodes. Frequently, the information collected will not be in a format ready for analysis, e.g. collection of pictures of produced components. An information extraction process is required that extracts the useful information from the underlying sources and expresses it in a form suitable for analysis. Simply storing the data might be insufficient. The heterogeneity and the flood of data require an adequate metadata management and data integration to have the change to generate value out of the data. Data integration is one of the cornerstones of modern commercial informatics, which involves the combination of data from different sources and provides users with a uniform view of data. It is expected that this part of the Big Data process will be outsourced and several new services will emerge (BITCOM, 2013).

Big Data techniques as they are listed in the MCGI report include categories of techniques applicable across a range of industries and not all of these techniques strictly require the use of Big Data—some of them can be applied effectively to smaller datasets (McKinsey Global Institute, 2011). Some important terms are briefly explained below.

- Data cleaning is a process to identify incomplete or incorrect data and if possible to clean meaning to modify or substitute such data to improve data quality. Data cleaning - e.g. as a part of data warehousing extract, transform and load (ETL) process - holds vital importance to keep the data consistency (Chen, et al., 2014). For Big Data, this is more complex due to the enormous amount and speed of data.
- Data fusion and data integration: Means a set of techniques to integrate data from different sources. The data fusion tries to complete missing data with an appropriate and useful representative. For example, data from web site logs can be combined with real-time sales data and social media data to determine the effect of online marketing.
- Data warehousing: In 2012 in a column of Roland Berger it was highlighted that one of the difference between a relational model, I would call it a classical data warehouse, and Big Data model is that Big Data models do not guarantee consistency but large sequential read access (Roland Berger, Sept. 2012). The Hub & spoke architecture is the most commonly used DWH architecture (BARC, 2014, p. 7). This is a mature research field for traditional database. Historically, two methods have been widely recognized: data warehouse and data federation. Data warehousing includes a process named ETL (Extract, Transform and Load).

2.2.3 Data management, data aggregation, data processing

This is the most IT technical part of the Big Data process as the data is distributed over several servers and the question of data modeling, data management and query performance is essential for the response time of the system. Therefor Big Data technologies are used and they are briefly described in this chapter. Big Data technologies are majorly classified into three bottom-up levels: (i) file systems, (ii) databases, and (iii) programming models (Chen, et al., 2014, p. 186 f.).

- File systems in general are responsible for the organization, reading, writing and the protection of files. Databases store their data in files on the file system. File systems for Big Data use cheap commodity servers to distribute data across them to achieve faulttolerance and parallelize activities to provide high performance services. Google File System (GFS) and the successor Colossus and derivatives of these open-source solutions like HDFS and Kosmosfs are well known Big Data file systems. Microsoft developed Cosmos, Facebook utilizes Haystack and Taobao also developed TFS and FastDFS.

- Traditional relational databases systems (RDB) storing the data row by row for each entity and normally support a structured query language (SQL) cannot meet the challenges brought about by Big Data. NoSQL databases (i.e. non-traditional relational databases) are used for these purposes. There are three main NoSQL databases (a) Key-value, (b) column-oriented and (c) document-oriented databases.
    - Key-value databases store data corresponding to key-values and are based upon a simple data model. The key has to be unique queried values can be found according to this key very fast. Amazon Dynamo DB was one of the first key-value products and influenced Redis, Tokyo Canbinet and Tokyo Tyrant, Memcached and Memcache DB, Riak and Scalaris, all of which provide expandability by distributing key words into nodes. More RAM (random access memory) based technologies are Voldemort, Riak, Tokyo Cabinet and Memecached.
    - The column-oriented databases store and process data according to columns other than the RDBs to rows. To realize expandability columns and rows are segmented in multiple nodes. The column-oriented databases are mainly inspired by Google’s BigTable, which is designed to process the large-scale (PB class) data among thousands commercial servers. Indexes of mapping are row key, column key and timestamps and every value in mapping is an unanalyzed byte array. By lexicographical order, rows are stored and continually segmented into Tablets (i.e. units of distribution) for load balance. Facebook´s Cassandra adopts the ideas and concepts of both Amazon Dynamo (key-value) and Google BigTable (column-oriented). It is a distributed storage system to manage the huge amount of structured data distributed among multiple commercial servers. Tables in Cassandra are in the form of distributed four-dimensional structured mapping, where the four dimensions including row, column, column family, and super column. Open-source projects based upon BigTable such as Hbase and Hypertable are available. Apache´s Hbase is a part of Hadoop of Apache’s MapReduce framework and replaces the basic file system GFS with HDFS.

- Document Database: Three important representatives are MongoDB, SimpleDB and CouchDB. MongoDB is an open-source product and stores documents as Binary with one document identification attribute as the primary key. Indexing of queryable attributes helps to enable rapid query. MongoDB supports horizontal expansion with automatic sharing to distribute data among thousands of nodes by automatically balancing load and failover. Amazon´s SimpleDB is organized into various domains which include different properties and name/value pair sets. These domains store the data and may be acquired and queried. Apache´s CouchDB is organized into documents comprising fields named by keys/names and values and every document is provided with a unique identifier.

Big Data are generally distributed over hundreds of servers. To secure a fast response when querying these data, parallel programming models have been invented and became cornerstones for massive data analysis. MapReduce is a simple yet powerful programming model with only two functions, i.e. Map and Reduce. The user needs to program these two functions over time to improve the programming efficiency, whereby some advanced language systems have been developed, e.g. Sawzall of Google, Pig Latin of Yahoo, Hive of Facebook and Scope of Microsoft. Dryad is a general-purpose distributed execution engine for processing parallel applications of coarse-grained data. The operational structure of Dryad is a directed acyclic graph, in which vertexes represent programs and edges represent data channels. All-Pairs a system specially designed for biometrics, bio-informatics, and data mining applications and which focuses on comparing element pairs by a given function (Chen, et al., 2014, p. 189).

2.2.4 Data analysis, data interpretation

Data analytics is the final and the most vital part in the value chain of Big Data. Like for the analysis of smaller data statistical technologies and data mining have been used the Big Data analysis involves the same methods but enhanced, whereby they can cope with huge amounts of data and that are able to visualize them. Some relevant terms are explained briefly:

- Data mining / predictive analysis: “A set of techniques to extract patterns from large datasets by combining methods from statistics and machine learning with database management. These techniques include association rule learning, cluster analysis, classification, and regression" (McKinsey Global Institute, 2011, p. 27). These techniques include association rule learning (discovering interesting relationships, i.e. "association rules," among variables in large databases), classification (identify the categories in which new data points belong, based upon a training set containing data points that have already been categorized) and cluster analysis. The most frequently used tools are (Chen, et al., 2014, p. 193):

  - R, an open-source programming language and software environment, is designed for data mining/analysis and visualization.
  - Excel, a core component of Microsoft Office, with plug-ins, such as Analysis ToolPak and Solver Add-in.
  - Rapidminer, also an open-source software used for data mining, machine learning and predictive analysis.

- Machine learning: "A subspecialty of computer science (within a field historically called "artificial intelligence") concerned with the design and development of algorithms that allow computers to evolve behaviors based upon empirical data. A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based upon data" (McKinsey Global Institute, 2011, p. 27). Natural language processing (NLP) is a very powerful example of machine learning which enables Amazon to launch Amazon Echo.

A frequently used term is Big Data Analytics. This term can be defined as the process of analyzing the data and respectively Big Data. The term doesn´t classify which techniques are used. Traditional analysis techniques and mathematical models are combined. A few major big data analytics application areas are discussed in the succeeding texts (Kune, et al., 2016, p. 87). Traditional data analysis techniques are (Chen, et al., 2014, p. 191):

- Cluster analysis: is a statistical method for grouping objects with same features.
- Factor analysis: is essentially targeted at describing the relation among many elements with only a few factors.
- Correlation analysis: is an analytical method for determining the law of relations, such as correlation, correlative dependence and mutual restriction.
- Regression analysis: is a mathematical tool for revealing correlations between one variable and several other variables.

- Statistical analysis: classical statistics based upon the statistical theory.

Whereas Big Data analytic methods include:

- Bloom Filter: Hash values of data are stored by utilizing a bit array.
- Hashing: it is a method that essentially transforms data into shorter fixed-length numerical values or index values.
- Index: Indexing is a technique to reduce the expense of disk reading by ordering index values at a disadvantage that it has the additional cost for storing index files.
- Triel: The main idea of Triel is to utilize common prefixes of character strings to reduce comparison on character strings to the greatest extent to improve query efficiency.
- Parallel Computing: a task is split into threads that compute parallel on different nodes. Traditional databases like Oracle, Teradata, etc., are capable or MPI (Message Passing Interface), MapReduce and Dryad.

Sometimes Big Data analysis is also classified based upon the necessary resources in:

- Memory-level analysis: The fastest accessible storage is the central memory. Nowadays this can be very huge (TB). Memory-level analysis is if the total data volume is smaller than the maximum central memory. It is extremely suitable for real-time analysis. MongoDB is a representative memory-level analytical architecture.
- BI analysis: is for the case when the data scale surpasses the memory level but may be imported into the BI analysis environment.
- Massive analysis: analysis surpassing the scale of traditional relational databases. Most massive analysis use MapReduce and HDFS of Hadoop.

The data analytical research can be divided into six key technical fields, i.e. structured data analysis, text data analysis, web data analysis, multimedia data analysis, network data analysis and mobile data analysis. Such a classification aims to emphasize data characteristics (Chen, et al., 2014, p. 195 f.).

- Structured data analysis: relational database management systems (RDBMS), data warehouse and OLAP use structured analysis over the past 30 years.
- Text data analysis: Test is the most common format of storing information, e.g. e-mails, business documents, etc. Text mining is based upon natural language processing (NLP).

- Web data analysis: Can be further divided into web content mining, web structure mining and web usage mining. Web content mining uses text and hypertext to discover useful knowledge from web pages. Models for web structure mining focus on topological structures provided with hyperlinks and reveal the similarities and correlations among different websites and are used to classify website pages. Web usage mining aims to mine auxiliary data generated by web dialogues and include access logs at web servers and proxy servers, browsers' history records, user profiles, registration data.
- Multimedia data analysis: Mainly related to images, audio and videos and include multimedia summarization, multimedia annotation, multimedia index and retrieval. Summarization can be accomplished by extracting the prominent words, phrases or video content sequence multimedia annotation inserts labels to describe contents of images and videos at both syntax and semantic levels. Multimedia indexing and retrieval involve describing, storing and organizing multimedia information and assisting users to conveniently and quickly look up multimedia resources.
- Network data analysis: link-based structural and content-based analysis of online social networks. The first group focuses on link prediction, community discovery, social network evolution, and social influence analysis whereas the latter one is also known as social media analysis and includes text, multimedia, positioning and comments mining.
- Mobile data analysis: Mobile data has characteristics, e.g. mobile sensing, moving flexibility that helps building and maintaining mobile communities. Mobile communities are defined as that a group of individuals with the same hobbies based upon geographical locations combined with same interests (e.g. the latest Webchat).

The value chain of Big Data helps to define the single steps that are necessary to consider when starting a Big Data initiative. It is not helpful when distinguishing data analysis for SMEs whether it should be defined as Big Data or not. This question is discussed in the following chapter.

2.3 Big Data for SMEs or small vs. Big Data

Companies like Google, Amazon and Facebook helped Big Data to grow to its actual importance. Their business models are based upon analyzing, interpreting and acting on data which customers of their service are sharing with them. They are very successful pioneers who disrupted industries and changed the life of many of us. However, there is more to come in the years ahead. SMEs will be affected at the latest when a direct competitor is becoming more competitive through better usage and smarter analysis of internal and external data. The availability of data and the technological progress forces mature enterprises to change from an experience and gut feeling based to a data-driven decision-making culture. The near real-time processing capabilities enable small organizations to serve more customers and scale up their business.

According to the definition for small and medium-sized enterprises from the EU from 2003 medium-sized companies have between 50 and 249 employees and an annual turnover below 50 MM €. This definition correlates with the definition of Austrian Chamber of Trade. Below 50 employees there exists a differentiation between small and smallest ventures but this is not relevant for this paper (WKÖ, 2015).

Big Data for small business sounds contrarian and considering the definition of Big Data it is questionable whether there is a fit at all. A small business will presumably not generate significant amounts of data, nor will its social media presence generate a data tsunami that can only be handled with techniques that are attributed to Big Data. However, a small business could be linked to Big Data if it uses – for example – predictive analysis or if its business model relates to Big Data. This could be achieved by selling a Big Data product, Big Data consulting or using the benefits of Big Data to innovate and generate a competitive advantage in the industry. The latter type of small companies will possibly grow and like Google or Facebook invent new solutions. Medium-sized companies could have the size to generate significant amounts of data and are probably more exposed to different types of unstructured data and should be considered as potential Big Data users.

To evaluate the question whether the Big Data philosophy is applicable on SMEs it is helpful to concentrate on the traditional analysis methods or sometimes also called small data. Small Data in this context is a synonym for traditional decision support systems (DSS) or data warehousing (DWH). DWH is defined (Inmon, 2005) as subject-oriented, integrated, time-variant and non-volatile collection of data. A DWH is a separated database where data from different sources, in general the organization’s operational database, are stored. This database is specially designed to fulfill analysis and reporting requirements. Therefore, the data is stored in either a stare or snowflake schema. The transformation task from the core operational systems into the DWH is called ETL. This means that the data has to be extracted from a system, transformed e.g. to standardize attributes or integrate them and load them into the DWH. These loads are executed on regular bases which can mean every month, day or minute. Sometimes it is necessary due to performance issues to develop data marts that store only parts of the available data or process them in a special way (Coleman, et al., 2016, p. 83).

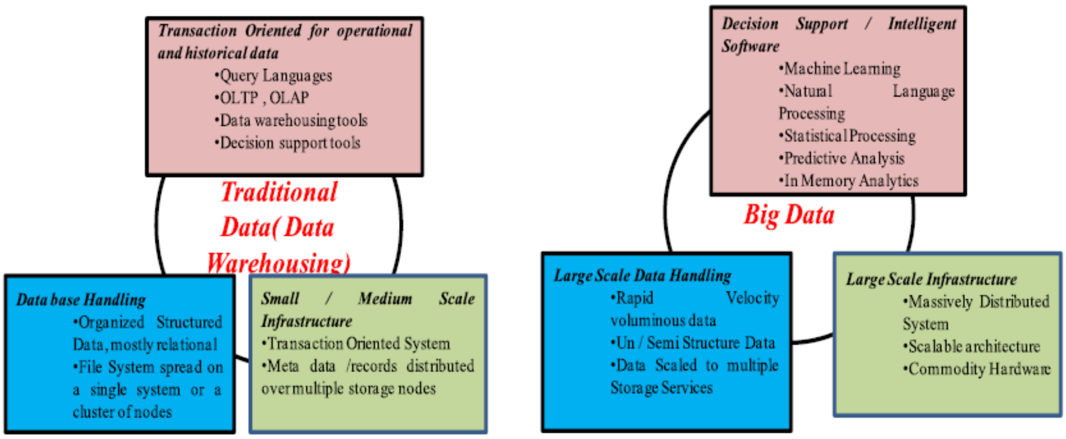

Transaction Oriented for operational and historical data
•Query Languages
•OLTP , OLAP
•Data warehousing tools
•Decision support tools
Traditional Data(Data Warehousing)
Data base Handling
•Organized Structured Data, mostly relational
•File System spread on a single system or a cluster of nodes
Small / Medium Scale Infrastructure
•Transaction Oriented System
•Meta data /records distributed over multiple storage nodes
Decision Support / Intelligent Software
•Machine Learning
•Natural Language Processing
•Statistical Processing
•Predictive Analysis
•In Memory Analytics
Big Data
Large Scale Data Handling
•Rapid Velocity voluminous data
•Un / Semi Structure Data
•Data Scaled to multiple Storage Services
Large Scale Infrastructure
•Massively Distributed System
•Scalable architecture
•Commodity Hardware

Figure 2 traditional DWH vs. Big Data source Coleman, et al.

However, when does a traditional DWH stop and a Big Data application start, given that both serve similar purposes and a near real-time DWH with some 100 TBs can already be called Big Data particularly if data mining tools are used for consumer behavior analysis? Accordingly, there is no strict answer to this question. A comparison of main characteristics of Big Data and Small Data (Delibašic, et al., 2014, p. 27) provides a brief overview of some aspects. The fundamental difference concerning the data acquisition is that Big Data tries to store all available data in its origin structure whereas the Small Data approach tries to store only necessary data to answer specific requirements. For sure the boarder is very blurred but the goal of a Big Data project is more flexible and less context based to one problem. The data sources for Small Data are in general more from inside the organization and for a Big Data platform the data is gathered from different media on different servers also outside the company. The Big Data server architecture comprises a distributed computing where multiple servers work together to store and process information. Concerning the used data structures Small Data uses primarily structured data (transaction data or XML) and Big Data processes and analyzes also unstructured data (e.g. free text, images…). For the most important part of the data environment, the preparation and analysis - namely the value generation – it is arguable that users of a Small Data environment are in charge of preparing their own data for their own purpose and that in most cases it is possible to analyze all the data at once. The Data preparation for Big Data is normally not task of the user and that the analysis is usually undertaken in incremental steps with different methods. This simplified differentiation is not meant to be exhaustive but it is a helpful guideline to distinguish traditional data environments from Big Data environments (Cecere, 2012, p. 9).

A survey in 2014 questioned the importance of classical data warehousing versus Big Data for the interviewed ventures. 88% identified the DWH as important or critical whereas one-third of the interviewees graded Big Data of the same importance (BARC, 2014, p. 10). This means that a big portion of the enterprises have a DWH and at least 30% also have a Big Data environment and that it is not a question of either or. A survey among German SMEs in 2015 revealed that 20% use Big Data software on regular basis and for 80% it is not planned in the near future. Almost all the surveyed enterprises are using traditional small data with technologies like Excel and 54% use a DWH. This questioning revealed also that in almost 70% of the organizations the department itself is in charge for data analytics and only for 15% the IT and 15% a separate department is doing data analytics (Vossen, et al., 2015, p. 41).

This is an indication that at the present the majority of SMEs have to be considered as users of small data. Especially for small and medium-sized companies, it is essential to obtain information quickly and without too much overhead. To derive value out of analytic models or DSSs, you have to invest lot of time and money in the implementation of these systems. It is questionable whether it is necessary or helpful for small organizations to obtain more data faster possibly near real time and whether this leads to better decisions. According to key findings from a data analytic SME study, the IT is not the core problem (Reker & Andersen, 2014):

1) It is not a question of Big Data or Data analytics or other passwords it is the problem of decision-making. Small and medium-sized companies have to honestly analyze the decision-making processes and identify relevant optimization and improvement potential to enhance the quality of decisions.
2) It is not an IT topic it should be driven by the top management and the technology is only a tool that should help to improve the decision-making process.

3) It should not mean that decisions are made only fact based or automatically even if this is a theoretical approach but the decision-making situation and the systems should be harmonized.

These findings emphasize the necessity for SMEs to work on possibilities to generate value with the help of data processing. The improvement of decision making or a more general change in the business model towards a more data driven one can bring a competitive advantage regardless if the underlying processing is considered as Big or Small data.

3 Big Data general aspects

This chapter is dedicated to survey results relating to the Big Data definition and value chain. The status quo of 3Vs discusses the underlying Big Data definition and how important the Vs are. The “Big Data initiatives” section reflects the willingness of the surveyed ventures to start a Big Data project. The remaining chapter provides an idea concerning what Big Data is used for and details the data acquisition- and processing-related findings of the studies.

3.1 Status quo – 3 Vs

Do the surveys reflect the definition of Big Data (see 0) in terms of the 3Vs and are they equivalent important for a company’s decision to start a Big Data project? A survey among SMEs in 2012 revealed that 70% of participating SMEs fulfilled at least one of the 3Vs and 20% fulfilled all three, although the author did not explain the criteria that led to this fulfillment (Frauenhofer IAIS, 2012, p. 44). Should this be taken for granted? The data volume is pretty easy to measure and data storage capabilities increase every year. Is the volume or the volume growth rate the basic criteria for enterprises to start working on Big Data? In 2014 an international survey revealed that on average respondents currently manage about 164 terabytes of data. Although the number varies considerably by size: larger companies report an average of 291 terabytes, while smaller companies average about 57 terabytes. Furthermore respondents expect that average to almost double over the next year to 18 months (IDG Enterprise, 2014, p. 3). However, still this growth in volume is not the basic driver behind Big Data investments. According to the NVP study 40% of firms cite “variety” (more sources) as the primary driver of Big Data investment, only 14.5% name “volume” (more data) or “velocity” (faster data) (New Vantage Partners LLC, 2016). These results significantly differ from the findings of the recent BARC survey, where the main reasons for Big Data initiatives are volume (57%), variety (50%) and velocity (46%) (BARC, 2015, p. 12). Presumably the difference can be explained with the higher participation rate of very large companies involved in the NVP survey. These organizations have already invested in data storage and Big Data initiatives and once a Big Data platform is implemented it is easy to increase the data storage. Big Data projects will also increase data growth themselves, as 60% of those surveyed predict that these initiatives will drive increased growth in unstructured data, presenting IT with further data management investments (IDG Enterprise, 2014, p. 4). Thereafter, the investment focus will shift to the other factors, namely variety and velocity, whereas smaller companies lack this trend. The results of a BARC survey published in 2014 – where participants were asked to classify the growth rate of their data volume of 2011 to 2012 – outline that 19% detected a strong and only 7% detected a very strong growth (BARC, 2013, p. 14), thus hinting at a manageable volume increase. Keeping the definition of Big Data in mind, where it is stated that it has to exceed the capabilities of traditional data processing applications to be classified as Big Data, the question is whether this data increase has to be qualified as such. A study by Deloitte identified that 87% of participating SMEs recognized a volume increase, while only 14% agree or strongly agree to the statement that the data volume exceeds the data processing capacity of their IT systems (Reker & Andersen, 2014, p. 10). Hence, only 14% of SMEs would see the volume characteristic as a criterion for their Big Data investment.

A survey in 2012 among international acting companies derived the finding that 85% of respondents felt that the issue was not so much about volume but rather the need to analyze and act upon Big Data in real time (CapGemini, 2012). The BARC Big Data survey reveals that in 2013 companies were able to analyze 4% of their data near real time, meaning with a delay of less than 5 seconds. Another 4% of the data is available for analysis within one minute. The main part (45%) of data is refreshed once a day. The companies plan to improve this situation and want to have 26% ready for analyzing within one hour and 10% near real time (BARC, 2013, p. 31). The data is not explicitly split according to the size of enterprises, although 24% SMEs participated in the survey and there is no significant difference between best-in-class and laggards, which supports the assumption that the requirements concerning velocity are the same.

The sources for Big Data are detailed in 3.5 and in general firms are seeking to integrate more sources of data, including new sources as well as legacy sources (New Vantage Partners LLC, 2016). Beside the usage of business data, log data and sensor data, an increased usage of social media and correspondence is planned. The surveys provide no indication that SMEs use different data sources compared with large companies. The usage of Big Data sources is more a question of the data analytical skills and the data-driven decision-making than the size of the organization. For SMEs, it is presumably the variety rather than the volume or the velocity that is the most important driver for Big Data activities.

3.2 Big Data initiative

Without proceeding further into detail about what a Big Data initiative precisely is – because it is also not detailed by the surveys – a general status quo of Big Data initiatives should be provided to derive a better understanding of who is using Big Data at present and whether there is a general trend towards using these technologies. A survey published by NVP in 2016 representing 44 of the Fortune 1,000 firms posted that it is now clear that Big Data reached mainstream and even the most cautious firms have adopted a Big Data strategy of some form. This conclusion is supported by several key findings of the survey (New Vantage Partners LLC, 2016, p. 4):

- Big Data in place: In 2014, almost 63% of firms reported that they had at least one instance of Big Data in production. This is almost double the 31.4% who reported the same result in 2013.
- Planned investments are expected to rise sharply: 26.8% of firms report that they will invest more than $50MM in Big Data by 2017, up from only 5.4% of firms that invested more than $50MM in 2014.
- Perceived Importance for business: 69.6% of firms now view Big Data as very important or critical to their business success.

According to a survey of more than 750 IT decision-makers in the third quarter of 2013 the interest in Big Data rose as 49% of respondents had been implementing or planned to implement Big Data projects within their organizations – a 5 percentage point increase from 2012. Almost half (48 %) of respondents expected that Big Data usage will be widespread across the company in three years, while another 26 percent expected mainstream use at one or more business unit, department or division (IDG Enterprise, 2014, p. 2). Comparing these outcomes with the status quo from NVP the conclusion seems likely that the size matters. However, beside NVP´s focus on very big companies the figures have to be put into regional perspective. According to a BARC worldwide survey from December 2014 to February 2015 with 559 participants only 18% have a Big Data initiative in production. The study shows a clear backlog of European companies with only 16% versus 28% of northern American companies in terms of implementing Big Data initiatives (BARC, 2015, p. 10 f.) and this is only half of the 49% of IDGs and far less than 63% of the Fortune 1000 enterprises. Once again this could be explained with the company´s size as only 30% of the BARC survey exceeds 5000 employees. Looking within Europe, there remain differences to be found. The BARC Big Data Survey Europe from 2013 divided the participants in best-in-class/average/straggler. This categorization was based upon a self-assessment of the data handling of the company in comparison with its peers. The largest group of stragglers with

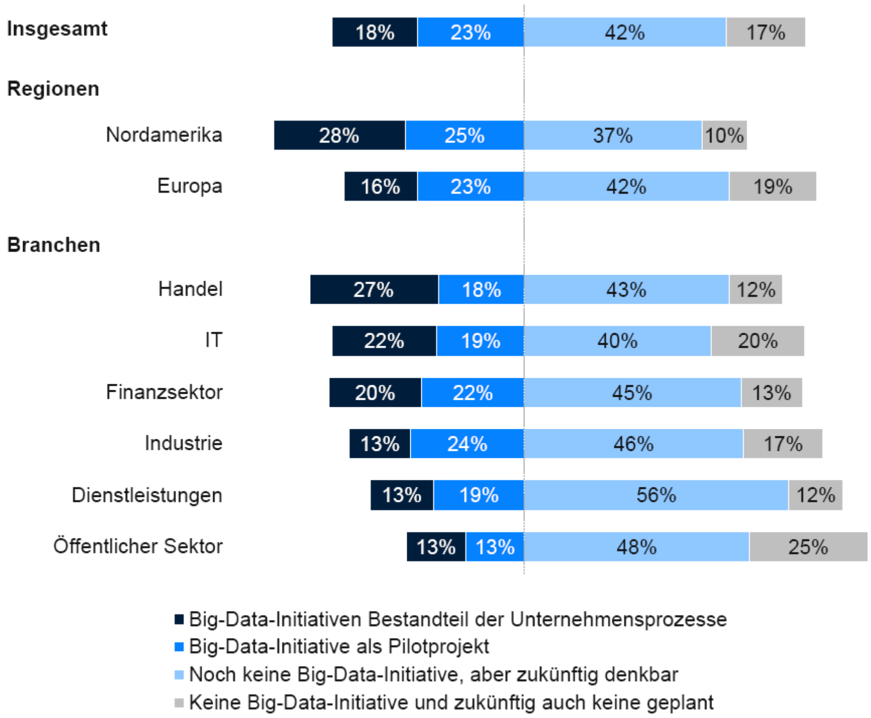


Figure 3 Big Data initiatives according to BARC

25% identified firms with more than 2,500 employees and in geographical term with 27% UK based (BARC, 2013, p. 10). Even within the group of larger companies there are 25% of them knowing that they lag behind their pears and this could somehow verify the 60% of NVPs paper. The average cluster consists to 58% of the average ventures employee between 250 and 2,499 employees (BARC, 2013, p. 10) and 40% of these best-in-class group were ventures with less than 250 employees and 33% of them out of DACH (Germany, Austria or Swiss) region. The DACH region somehow outperforms the UK in terms of data usage but this is no indication whether they use Big Data or not. Another survey reported that UK-based SMEs are lagging behind in the usage of big data analytics and that the adoption rate was 25% for businesses with over 1,000 employees but only 0.2% for SMEs. Even if the expected growth rate of 42% would come true, SMEs would still lag behind (Coleman, et al., 2016). A German based survey among SMEs reported that 20% use Big Data tools on a regular basis (Vossen, et al., 2015). An Austrian-based study for SMEs highlights that 80% of the interviewees perceived the topic Big Data as relevant for the future, although only 50% thought that it was relevant in 2015 and that only a small portion of them plan to implement a Big Data initiative (Digital Networked Data, 2015). This hasn´t changed much since 2013, when 38% didn´t consider Big Data at all, almost 50 % discussed the topic, 3% planned to and 9% implemented a Big Data initiative. Focusing on the SMEs of this survey 52% cited not to consider Big Data, 40% discussed and only 8% implemented Big Data initiatives (Köhler & Meir-Huber, 2015, p. 78).

Considering these findings, it seems that Big Data remains a topic for large companies, whereas only about 10% of SMEs are truly investing in this topic.

3.3 Big Data goals

Most of the surveys include questions about the Big Data aims and goals. This section gives a brief summary about Big Data goals presented in literature and lists finding of the studies. One goal that can be found within the definition of Big Data is to derive value from the data (see 2.1) According to McKinsey Global Institute, the five top findings where Big Data creates value across sectors are as follows (McKinsey Global Institute, 2011, pp. 91 - 109):

1. Creating transparency: Value can be created by providing information faster to relevant stakeholders and make it more easily accessible.

2. Enabling experimentation to discover needs, expose variability, and improve performance: Through the increase in analysis performance and the rising amount of available data points, from assembly-line equipment to automobiles to mobile phones that measure processes and human behavior, companies are enabled to base their decisions concerning the outcome of controlled experiments to make better decisions. For example, retailers are adjusting prices and promotions in a bid to experiment with which combination best drives traffic and sales.
3. Segmenting populations to customize actions: Big Data helps to target services or meet individual needs with personalized offers in near real time.
4. Replacing/supporting human decision-making with automated algorithms: Sophisticated Big Data-based analytics including rule-based systems, statistical analyses and machine learning techniques can substantially improve decision-making and minimize risks. Big Data provides the raw material necessary and Big Data enables those algorithms to operate.
5. Innovating new business models, products and services: Big Data enables enterprises to create new products, like data obtained from sensors embedded in products to create innovative after-sales service offerings such as proactive maintenance and invent entirely new business models. For example, with real-time location data, an entirely new set of services has emerged, like pricing property casualty insurance based upon where and how people drive their cars.

These key findings are relatively general but concisely express the advantages of Big Data. These findings will be labeled as M1 to M5 to compare them with survey results explained in the following paragraphs.

With 30% the second ranked goal of the NVP study is “faster time-to-answer, faster time-to-decision, and faster speed-to-market” and this relates to M2. Firms clearly see Big Data as providing an opportunity to gain and act on insights quickly to seize market advantage. Furthermore 9.3% cited greater analytics capabilities and the opportunity to create a data-driven culture as Big Data goals (New Vantage Partners LLC, 2016, p. 4) and 51% (second ranked) in BRAC survey is “better steering of operational processes” followed by 50% for “faster and more detailed analysis” (BARC, 2013, p. 25). Also 32% of SMEs expect better data analysis.

M3 is supported by following results: The NVP survey among big multinational firms revealed that 37% cite the ability to develop greater insights into their business and customers as the single largest driver of Big Data investment (New Vantage Partners LLC, 2016, p. 4) and according to #big Data, 25% deemed the goal to detect trends in customer behavior in second place in 2013.

59 % of respondents of the IDG survey cited to improve the quality of decision-making and 53 % deemed making quicker decisions as the primary business driver (IDG Enterprise, 2014). Both top ranked findings refer to M4. As well as the top ranked goal of the BARC study where 59% of participants ranked strategic decision-making first (BARC, 2013, p. 25).

For the categories M1 and M5 no equivalent aspects were found in the surveys. Following aspects could not be classified: The Deloitte survey obtained that 54% of participants want to improve processes, 47% see necessary improvement by quality assurance and data cleansing and 43% increased IT implementation as a top priority. Investment in workforce or more intensive usage of external data sources were occasionally mentioned (Reker & Andersen, 2014). Lower costs were selected by 28% as a Big Data benefit (BARC, 2013, p. 25).

The studies give no reason to differentiate the goals according to the size of the companies. The questionnaire that is most focused on Austrian based SMEs heads with 32% of questioned SMEs better data analysis in first place, followed by the goal to detect trends in customer behavior (25%), to optimize process (18%), to reduce cost (14%) and conduct profitability analysis (11%) (Köhler & Meir-Huber, 2015).

It is difficult to extract a clear picture of main goals of Big Data initiatives due to the different categories that are included in the questionnaires. Every survey has to be seen in its context but trying to come up with a global summary leads to following points:

- improving the quality of decision-making and making quicker decisions
- develop greater insights into their business and customers
- planning and steering operational processes
- more detailed and faster analysis
- monetary aspects (cost cutting, increase in turnover)

3.4 Big Data use cases

Many different Big Data use cases are documented and publicized by software developers and consultancies to emphasize the value that Big Data can have for companies. These use cases can help to understand what Big Data can be used for. IDC's top use cases are listed in Figure 7 (IDC, 2012).

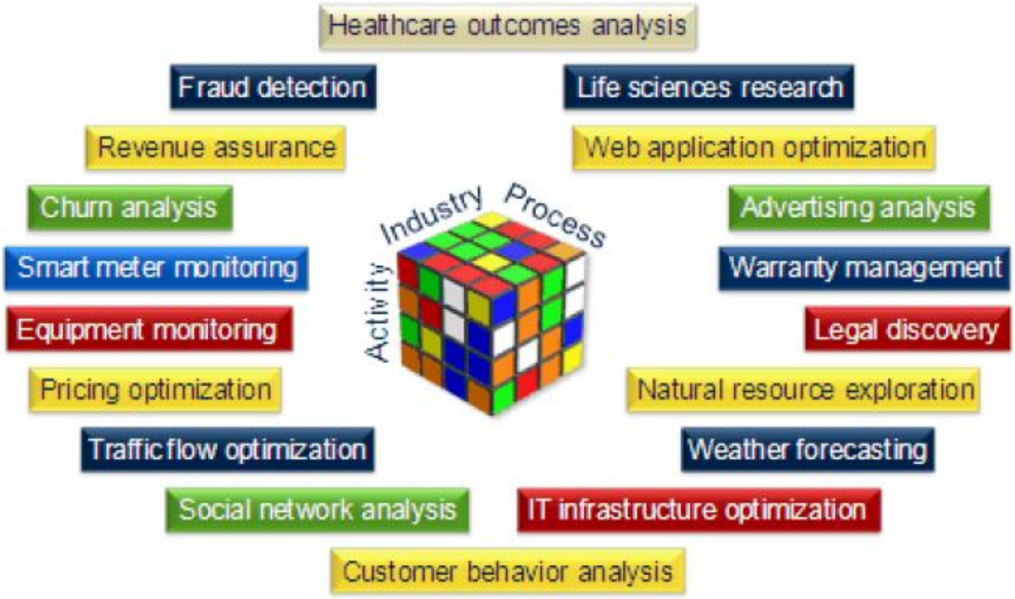


Figure 4 Top Use cases for Big Data according to IDC, 2012

Observing these use cases, it is questionable whether they are primarily Big Data-related because some of these topics have been around for years and were already solved with traditional analytical tools. Once again, the border between conventional ware housing and Big Data is blurred (see 2.3). However, these examples have to have some special characteristics of the 4Vs to be determined as Big Data use cases. For example, customer analysis can be considered as a state-of-the-art use case for every data warehouse. Enhancing this use case in terms of Big Data, it is necessary to have millions of customers, use the location information of millions of wearable devices, use billions of transaction data or the analysis is based upon a very broad range of sources from interaction with the point of sale, social media feedback to log files from web servers, all in near real time. In reality, companies nowadays like to refer to their application as a Big Data application. In general, Big Data applications in enterprises replace existing BI, OLAP and data mining applications (Chen, et al., 2014, p. 198).

Figure 7 already shows categories of use cases rather than use cases. The comparison of use cases across different surveys is difficult because several different use case categories were formed. Only three surveys included questions about use cases. The BARC Big Data Use Case study featured open questions and the results were ranked by how many times a single term was mentioned. The BARC Big Data Analytics survey lists use cases, although it is unclear whether these cases were chosen by the author and the interviewee had the possibility to rank them or the use cases express a clustering of free-text answers. The Frauenhofer IAIS started a web analysis and identified certain use case categories, subsequently asking the interviewee to rank them.

In general predictive analysis (see 2.2.4) is perceived as a very important topic whether it is about trends in customer care or customer behavior (12%), churn prediction, the prediction of service windows (41%), the calculation of spare parts availability (30%) or guarantee analysis (BARC, 2013, p. 30 f.).

According to the Frauenhofer results companies are most interested in marketing related topics like predicting the efficacy of advertising, monitoring the brand perception, monitoring competitors, competitive pricing or buying interest in the web. The results of the BARC 2014 survey also identify that the heavy users are marketing and sales but BARC use cases are more customer-oriented than those revealed by the Frauenhofer survey, like identifying customers with high customer value, more detailed customer segmentation and optimizing campaigning (BARC, 2014, p. 30).

For the manufacturing sector the potential usage of Big Data has to be differentiated more closely as passwords like IoT and Industry 4.0 drive data generation. It was an early and intensive user of data to drive quality and efficiency, adopting information technology and automation to design, build, and distributes products since the dawn of the computer era (McKinsey Global Institute, 2011, p. 76). MGI highlighted that manufacturers can use Big Data across the value chain and they identified seven levers to improve performance. Even if these findings are very general they help to gain a better idea of what Big Data can be used for. The first lever is related to R&D and should enable a concurrent engineering with more rapid simulation by building a consistent interoperable, cross-functional R&D and product design database. This relates to creating transparency and this is one of the most important values that can be created by using Big Data. The R&D departments especially in large companies can benefit from interoperable cross-functional data pools that would help to reduce redundancies and focus on potential inventions that haven´t been successful on the market to date. Such a system developed for R&D relevant aspects could easily be used to interact with people outside the R&D departments and could lead the company to an open innovation approach. The second and third lever are related to R&D and marketing and sales and are also aimed at sharing data and be more transparent by aggregating customer data and share data through virtual collaboration sites or idea marketplaces. For the supply chain management, MGI identified an advanced demand forecasting and supply planning across suppliers as most valuable. For the production, a digital factory that creates process transparency and the development of dashboards increases lean manufacturing. More generally, in terms of operation it is mostly about efficiency and optimization to reduce costs or improve quality. Another important lever is to implement sensor data-driven operations analytics to improve throughput, optimize the internal logistics, identify bottlenecks and enable mass customization. The sensor-driven operation and Industry 4.0 will generate more and more data and this is very closely related to Big Data. The last lever is related to the Internet of Things (IoT) and this suggests collecting after-sales data from sensors and using this information to trigger maintenance, after-sales service or detect design flaws. The marketing and sales tasks can contribute to Big Data platforms with customer data, market prediction, the demand forecasting and capturing market relevant data. The after-sales service can collect data from sensors that are included in the products and help to analyze the usage, triggering after-sales services and detecting manufacturing or design flaws (McKinsey Global Institute, 2011, p. 78 f.).

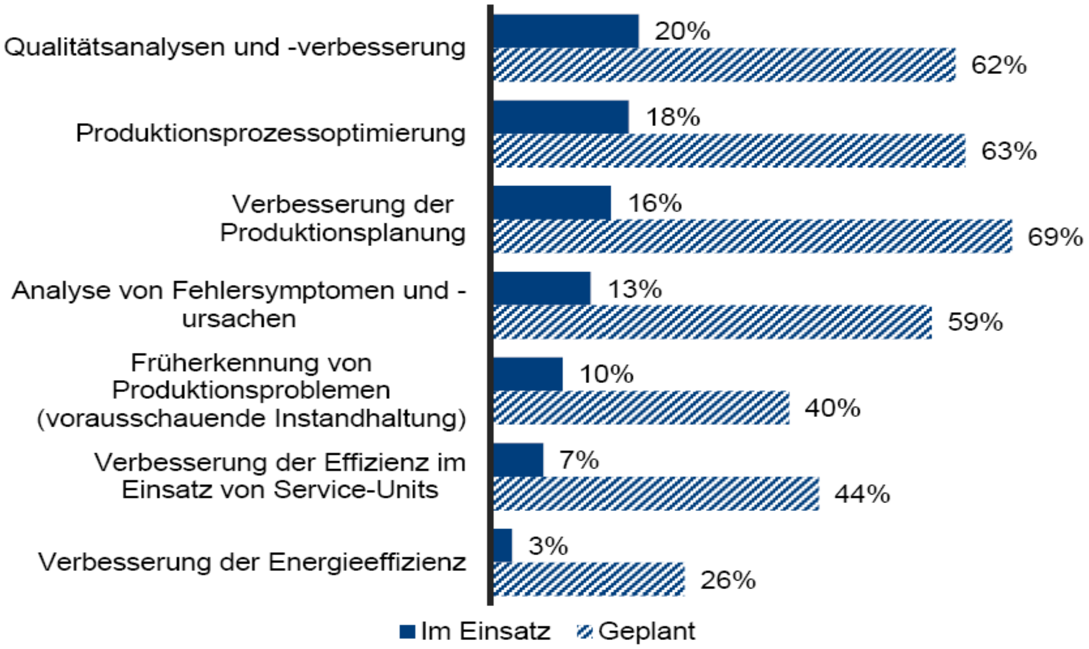


Figure 5 use cases in production according to BARC 2013

For production the top use case is quality analysis and – enhancement followed by production optimization and optimizing the production planning. Especially in production, you can find significant amounts of structured and unstructured data, particularly sensor data. The authors experienced that data is usually joined with data from other parts of the supply chain to gain a holistic view of the processes and understand cause and effect of - for example - quality defects. The related topics process optimization and production planning use the rising data volume from machinery, transportation and loading equipment or the production line itself with describing or predictive analysis (BARC, 2014, p. 32 f.). The most interesting use cases according to Frauenhofer IAIS for the manufacturing industry are predicting maintenance, operational optimization and product innovation (Frauenhofer IAIS, 2012, p. 46).

The supply chain can use Big Data for inventory optimization, logistic optimization, supplier coordination and involvement (Chen, et al., 2014). Looking at the supply chain of global enterprises, it is fair to assume that today´s supply chains are more complex than ever. Thinking of a company selling products comprising parts of different vendors that have to be shipped all over the world to assemble the product in one part of the world where the labor costs are cheap and subsequently have to be shipped all over the globe to sell them, at present all these different parts can be tracked in near real time, which has increased the data in volume, variety and velocity. Big Data enables managers to store all this data and postpone the analysis. The common way is to define the purpose of the analysis and based upon these requirements enable the IT systems to collect and analyze the data. One major advantage of the possibility to store huge amounts of data is the possibility to extend the supply chain from the supplier´s supplier to the customer´s customer. This enhances supply chain managers to overcome previously existing system boundaries and better manage information, cash and product flows (Cecere, 2012, p. 10).

To proceed into further detail for Austria, some information was found for the producing industry. The potential in agriculture and forestry is seen to be limited even if huge American companies have shown that it is possible to optimize the output by considering several aspects like landscape, weather and the sort of grain (see #big data, page 84 f.). Big Data value in the mining industry is also seen to be rather small. The possible use cases are focused around production optimization and monitoring of heavy equipment which presumably will be equipped with sensors in the future. For good manufacturers, almost 25,000 enterprises in Austria, additional benefit could be generated by reducing waste production thanks to the possibilities of near real-time analytics. In the energy industry Big Data will help to manage new challenges that come along with smart grids and smart meters (BITCOM, 2013, p. 22). This will have an impact on the pricing and the optimization of grid usage. The construction industry in Austria is considered not very IT-affine and thus the outlook for using Big Data is negative even though the general potential in the industry to gain competitive advantage is intact (Köhler & Meir-Huber, 2015, p. 94).

As these findings remain very global and to demonstrate more practical use cases especially for SMEs three success stories related to the manufacturing industry are listed below.

A German-based steel producer used a Big Data platform to enhance the steel production. Almost one-third of the steel production is scrap and the problem is that two-thirds of this waste can only be detected at the end of the process. The production line is controlled by employees of the quality assurance and it is additionally equipped with a network of sensors like video cameras, laser based and ultrasonic instruments, surface inspection and several temperature and vibration measurements. All this different information generates a stream of data that conventional systems were unable to process, e.g. the video examination alone generates hundreds of TB a year. This Big Data platform enables the company to react in near real time to recognized deviations and reduce the amount of produced scrap. Furthermore the time buffer, that has to be included in the process to reproduce the scrap, was optimized. Due to Big Data technology it is possible to store the data and use it for an ex-post and ex-ante analysis. This helps to recognize new patterns in an ex-post data mining and predict certain process conditions. Prescriptive analytics is used to inform the process owner that it is necessary to take a decision and certain information is presented in a dashboard (BITCOM, 2015, p. 35).

A company of the machine building industry also used data of the production line sensors to control 100.000 data points with up to 1,000 updates per minute. The system stores data structured and unstructured data like quality assurance protocols, event logs, images and video sequences (BITCOM, 2015, p. 59).

A global producer of elevators developed a platform to store and analyze information that is sent every minute by sensors of 1.1 million elevators. Based upon the IoT, a cloud-based solution is used to predict maintenance with the help of machine learning algorithms. This helps the company to gain a competitive advantage by offering a better maintenance service and streamlining the operation for this service (BITCOM, 2015, p. 32).

To summarize this chapter a rather global group of use cases is given (Power, 2014):

- customer analysis;
- data-driven products and services, including personalization systems;
- operational analytics, quality assurance and monitoring business operations;
- fraud detection, compliance, security/intelligence extensions;
- enterprise data warehouse optimization and data warehouse modernization.

These Big Data use cases are relevant fo all companies regardless of their size. The necessity of sound use cases and success stories is obvious, although until now potential users obviously neglect them. At least almost 50% of contributing SMEs have stated that they do not perceive an identifiable advantage in Big Data (Vossen, et al., 2015).

3.5 Big Data sources and types of data

This chapter gives a short overview of relevant data sources for Big Data platforms and reviews the survey results based on these findings.

In 2013 IBM issued a report “the Applications of Big Data to the Real World” which indicates that the internal data of enterprises are the main sources of Big Data. Those data are created and collected within the company and the organizations are familiar with the data structure and the incorporated information. This corresponds with findings of the Deloitte survey where respondents when asked which IT systems are used to gather information for decision-making, problem analysis, developing new products and controlling, answered 94% use heavily financial reporting systems, 70% ERP and data exchange systems with customers and suppliers and 51% management information systems (MIS) (Reker & Andersen, 2014). This is not very detailed regarding the type or source of data because a MIS can include all different types of data but the core message is that essentially business data are used. The Frauenhofer IAIS structured possible types of data for their survey into business data, machinery generated (sensor, logs, etc.) or human generated (social media, email, etc.). 75% of the participants use business data, 53% machinery and 42% content data. 30% of them use machinery and business data and only 10% all three categories (Frauenhofer IAIS, 2012). In 2012 a survey among 32 worldwide active industry leaders (like 3M, BP, Carlsberg, …) with the emphasize on supply chain showed that geo-location and mapping data (47%) and product traceability data (42%) are heavily used. In the survey this type of data was categorized as business data and data coming from IoT and sensor data are used by 30% and sentiment data is used by 20% (Delibašic, et al., 2014, p. 13). Respondents of IDG survey cite e-mails (52 percent) and customer databases (49 percent) as the current most common Big Data sources (IDG Enterprise, 2014). Similar results can be found in the BARC Big Data Survey Europe where transactional data are used by 70% of the companies. In general, this type of data is well structured and thus easy to process. Moreover, log data (55%), sensor data (44%) and unstructured data (40%) are already used for analysis. Until 2014 data from social media is not very often used (14%) (BARC Big Data Survey Europe, page 29). The NVP survey among large companies obtained the need to integrate more data, from new sources as well as legacy sources without further information. Although only 3.6% of firms cite the need for unstructured data (e.g. documents, text) and only 1.8% cite social media data as a priority (New Vantage Partners LLC, 2016, p. 10). However, this could be explained with the assumption that companies of the Fortune 1000 have Big Data and probably these sources are already included, while the same argument was used for the volume disparity. Analyzing the data usage of best-in-class companies, it can be witnessed that best-in-class use more types of data and especially the usage of transactional data (48% vs. 32%) and social media data (20% vs. 6%) is significantly higher (BARC, 2013). The surveys that are more focused on SMEs do not include questions related to data sources.

The Frauenhofer study asked to state whether the representatives of the companies consider explicitly named data types, i.e. data service, audio/video, patents, static data, images, sensor data, free text, web content, log data, social media posts, correspondence (email/letter/fax), CRM data and transactional data as very important (4) or not important (1) for the success of the company. Assuming that data that is important for the success of

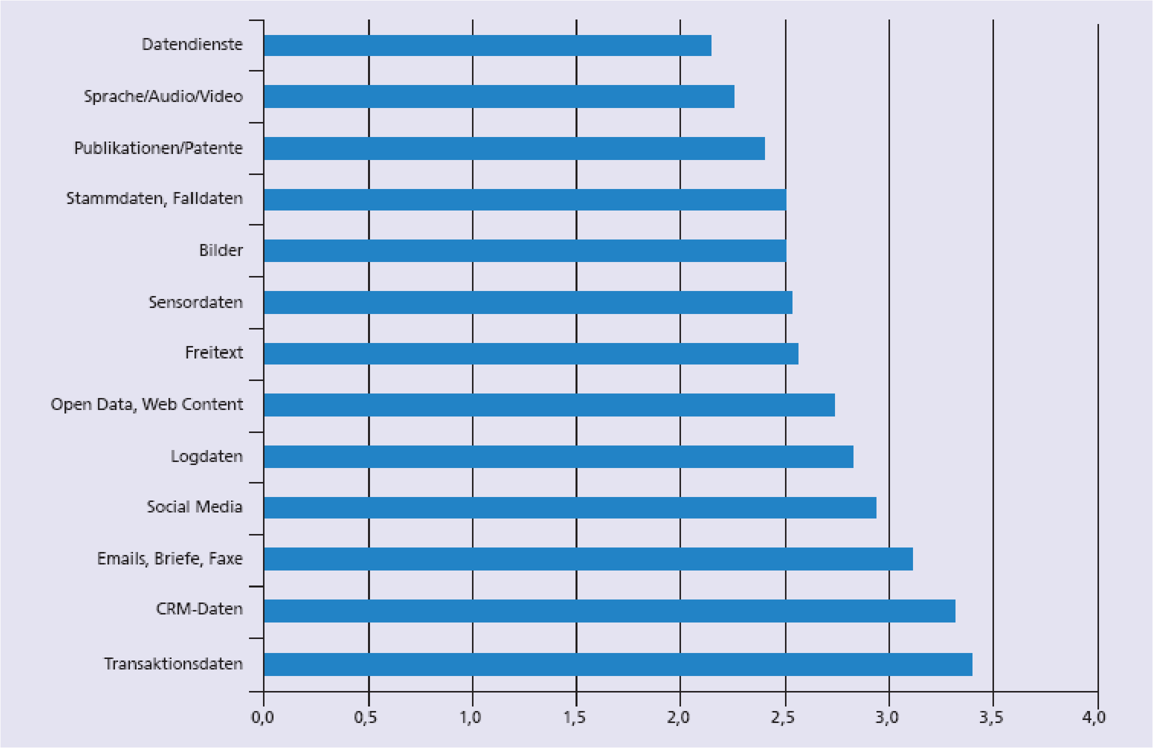


Figure 6 Importance of data for business success according to Frauenhofer IAIS

the company is also a relevant source for Big Data, the top information is provided by transaction information followed by CRM data and e-mails (Frauenhofer IAIS, 2012, p. 44). These findings are also obtained by a survey of Capgemini questioning which types of big data sets adding the most value to the organization, where 69% selected business data, 32% correspondence, 28% social media data, website clickstream data 22%, RFID 19% and geospatial and telecommunication data with 16% and images with 8% (CapGemini, 2012, p. 10). The category CRM Data was not available in the latter survey.

The future potential of different data sources is considered to be very high, which is reflected in very ambitious planning. Beside the top ranked transactional data, companies plan to use documents/text (50%), social media data (47%) and log data from IT (45%) and the web (44%). The better part of companies recognized the additional value of data (BARC, 2013, p. 41) but also 42% of survey respondents say that unstructured content is too difficult to interpret (CapGemini, 2012, p. 11). This could explain why categories that are perceived as adding significant value to the business success like social media and correspondence are not yet used as a data source.

A classification of Big Data sources can also be found in literature. For example it is distinguished between the operation and trading information in enterprises, logistic and sensing information in the Internet of Things, human interaction information and position information in the Internet world, and data generated in scientific research, etc. (Chen, et al., 2014, p. 179). Table 4 summarizes the findings well-rounded according to these categories.

Type of data	Frauenhofer	Supply chain	BARC 2013	CapGemini 2012
Operation and trading	75%	45%	70%	70%
Log and sensing	50%	30%	50%	20%
Human interaction	40%	20%	40%	30%

Table 4 summary of % Usage of data sources

The surveys reveal that two third of Big Data initiatives use business and transactional data, approximately the half use email and customer databases and 40% source log files and sensor data. High potential is seen in social media data, although it is not so frequently used yet. A differentiation for SME was based on the survey findings not possible and it is probably not necessary to do so as SMEs can generate and acquire the same sources like bigger companies presumably the amount of data is smaller.

3.6 Big Data Techniques and technologies

The literature about Big Data includes several different sometimes well known techniques and technologies like data warehousing, data mining or different statistical analysis methods (see 2.2.2 and 2.2.3). Big Data techniques draw from traditional information technology fields and extend them to the next level in term of volume, velocity, variety and vale (McKinsey Global Institute, 2011, p. 27).

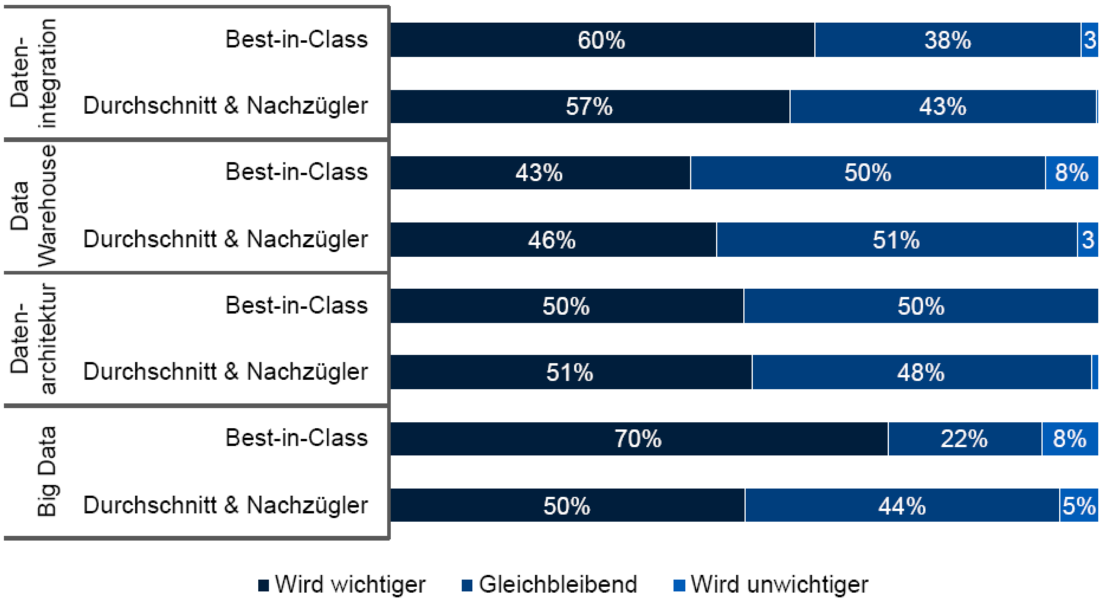


Figure 7 development best-in-class according to BARC

Classical data warehousing was identified by 88% as critical or important (BARC, 2014) and among SMEs in 2015 still 56% reported to have a DWH in place (Vossen, et al., 2015). Taking a closer look at the best-in-class category it is revealed that the prospect of the importance of Big Data is the major difference to the not best-in-class and it is expected that DWH will lose importance (BARC, 2014, p. 11). The best-in-class category was defined by the author on behalf of a self-assessment of the interviewee of the company's ability in data management compared to their peers (BARC, 2014, p. 35).

With 52% analytical solutions top the list of technologies for extracting value from business data cited by those surveyed and 36% of them plan to invest in predictive analytic products (IDG Enterprise, 2014). A more detailed picture can be drawn of the BARC survey and beside visualization and real-time reporting, which is not considered as a analysis technique, data mining for segmentation (23%) and predictive analysis (16%) are the top preferred techniques to analyze Big Data and companies want to invest heavily in those categories (55% and 62%). In terms of key technical fields of the Big Data analysis (see 2.2.4) the mobile data analysis is mostly used with 17% followed by text data analysis (11%), network data analysis (9%) and multimedia data analysis (8%). Most of the contributing firms plan to use network data analysis (54%), text data analysis (50%) and mobile data (46%). Companies that are best-in-class use in average more different Big Data techniques compared to the average. Significantly higher is the usage of visualization techniques like dash boards (55% vs. 33%) and network data analysis (23% vs. 6%) (BARC, 2014, p. 41).

According to the BARC Big Data Analytics Survey primarily standard RDBMS are used (62%). As already explained these tools are the basic toolkit of BI and data warehousing, which is why these tools are also frequently used for Big Data. Analytical databases (27%) and Individual solutions (26%) are frequently used. It has to be considered that individual solutions include self-made solutions and models based upon analytical tools as R, S, SAS or SPSS. Special Big Data technologies like other noSQL databases and event processing remain in the minority. Only 9% used the Hadoop framework . Companies plan to increase the usage of Big Data technologies. Analytical databases (plan 41%) and predictive analysis tools (32%) are top ranked for future use (BARC, 2014, p. 38). Best-in-class companies have a higher usage of explorative analytic databases (46% vs. 22%), Hadoop (20% vs. 6%), NoSQL databases (5% to 15%) and Big Data appliances (16% vs. 11%) – that is hard- and software from one company e.g. Oracle (BARC, 2014, p. 45). The NVP study among 40 very large companies revealed that for 30% a priority is providing greater agility by using for example the Hadoop framework (New Vantage Partners LLC, 2016, p. 13).

According to the SME survey in Germany in 2015 almost 20% use Big Data software on a regularly basis, were Big Data software was defined as, Hadoop MapR, Apache, Hortonworks, Cloudera, IBM InfoSphere BigInsights, Amazon Elastic MapReduce, Splunk or Palantir. 20% seem to be optimistic as only 10% of SMEs are investing in Big Data but the survey in Germany accepted bigger companies to participate and eventually they are responsible for the Big Data usage. Furthermore, the participants use traditional technologies like Microsoft Excel (92%), relational databases (77%) and DWH solutions (54%) (Vossen, et al., 2015) which correspond to the BARC findings.

For SMEs it can be said that Big Data technologies are only selectively used and traditional analysis methods prevail. In general the importance of DWH solutions is expected to reduce with a more focused usage of analytical databases and the Hadoop framework. Data mining is and will be the most important analysis technique and mobile data analysis is the top analytical key field.

4 Big Data Implementation

This chapter is dedicated to aspects of the surveys that relate to Big Data implementation. The hurdles and challenges that enterprises witnessed during implementation or expect to find when starting with a Big Data initiative are covered by most surveys. The positive side namely success factors are hardly subject of investigation. A frequently mentioned topic is the importance to define a strategy upfront a Big Data strategy. All three issues are detailed in the following chapters.

4.1 Big Data projects success factors

According to “Big Data project success – a meta analysis” the success factors can essentially be categorized along the dimensions of people, process and technology.

Along the people dimension Data scientists are the key professionals necessary for Big Data projects. They are difficult to find, expensive and these contributors need to have computational and analytical as well as domain-specific skills. In general Big Data projects require various skills where the bests result can be expected by creating a multidisciplinary analysis team that can handle issues like data security, data quality and trust as well as privacy and ethical issues (Koronios, et al., 2014, p. 4) (see 4.2).

The process for Big Data projects can be split into two stages the exploration of data and exploitation of data (Koronios, et al., 2014, p. 5). Both require different process approaches. The innovative character of the exploration phase requires team members to be capable of think in innovative ways and come up with creative ideas. This could also trigger a change in the corporate culture to a more innovative one. A Big Data initiative should be agile enough to help team members to act on ideas that evolve during this exploration phase (BITCOM, 2013, p. 39) and Big Data projects have the ability to be very agile due to iterative process by which they can load data, identify correlations and patterns and subsequently load more data (Bean, 2016). The exploitation phase should take advantage of information asymmetries for creating business value and competitive advantage. Traditional data management process models remain relevant to Big Data projects, although considering learning phases in the planning process is crucial (Koronios, et al., 2014, p. 5).

The technological dimension is the most significant difference to traditional IT projects, and one which has been used as its de facto definition (Koronios, et al., 2014, p. 5), referring to data volumes challenging normal database systems. This definition is somehow inflationary because whenever the traditional database systems become faster and more efficient this technological challenge increases. However the technology that is available along the Big Data value chain is certainly a major factor that decides the success of a Big Data initiative (see 2.2). Even though the technological challenges with Big Data has been critical, the various numbers of Big Data projects that have overcame technical hurdles show that it is feasible (Agrawal, et al., 2012, p. 35).

The following critical success factors (SCF) were identified with an ABC analysis out of 60 case studies of Big Data projects (Koronios, et al., 2014):

1) Information Strategy for Big Data: For leveraging the opportunities of Big Data it is critical to develop a strategy that aligns IT with the line of business (LOB) to ensure that they are focused on the same business needs.
2) Identifiable Business Value: The project must have a clear vision of how it may contribute to the business value.
3) Top Management Support: this success factor was identified by all cases of the study.
4) Skills for Big Data projects: multidisciplinary teams are required involving business and domain experts, technical experts and analytics professionals.
5) Information Quality, Security and Integrity as well as ensuring the privacy is a very important activity of Big Data projects.
6) Technological capability: Big Data projects are dependent to a large extent on new technologies to handle the huge amounts of complex data at speed.

The most important success factor is frequently cited by the surveys and also by several articles and that is why a separate chapter deals with this topic (see 4.3). Most of the other success factors are highlighted as challenges in the surveys (see 4.2). Explicitly success factors were only posted by NVP and the IDG survey. According to NVP a clear plurality of 40% named business and technology partnership followed by a strong business leadership and sponsorship (23%) and recognition that data is a shared corporate asset (11%) as most relevant (New Vantage Partners LLC, 2016, p. 10). The first two findings are best integrated in the SCF 3 - top management support and the last finding fits to the SCF 1 – information strategy. According to IDG survey 2014 IT and business leaders agree that Big Data projects succeed best when implemented as a jointly owned project designed to solve specific business challenges. 91% of LOB managers ranked most important for the success to identify the business areas and processes where Big Data can have the greatest impact. 72% of respondents cited the importance of having professionals with the right skill sets in house as a critical or very important success factor to their projects. 62% of respondents also said that identifying Big Data evangelists to act as corporate sponsors was either critical or very important (IDG Enterprise, 2014). IDGs top finding corresponds with the SCF 2 – identifying business value and the importance of having professionals as well as the data evangelists fit to SCF 4 – skills for Big Data projects. These SCF can also be compared with a BITCOM success factor list of and the top three are more or less identical (BITCOM, 2013, p. 36).

A common best practice advice is to include a prototyping phase in the project plan. For example Roland Berger advices to launch a successful Big Data project a company has to plan a strategy -, enabling -, experimentation - and a rollout phase. It is recommended to start with a prototype with Big Data on an experimental basis and subsequently expand very promising use cases (Roland Berger, 2013). Very pronounced is the phrase "Start small, fail fast" and especially for SMEs it is crucial due to limited resources to build a low-cost, scalable system that allows agile systems and applications (Rising, et al., 2014, p. 20). During the prototyping phase a few intuitive examples shall allow the organization to see what the data can do and fast results that are easy to test should show what type of lift the analytics provide (Franks, 2012).

These success factors are relevant for all business sizes and represent the positive aspects of how to succeed with a Big Data initiative.

4.2 Challenges and hurdles in Big Data

Every IT project faces challenges and hurdles during implementation and the problems Big Data projects have are not much different to those DWH projects experienced. Issues of the conceptual, technical or project management sort should be addressed in well known project management style to finish the project in time and in budget. However, doubts and disbelief in the Big Data paradigm and questions of global scale like privacy concerns are issues that have to be outlined in terms of challenges and hurdles that hinder Big Data initiatives. Such challenges were summarized in an article in 2012 to the correlation problem, the human understanding problem, the privacy problem and the provenance problem (Pentland, 2012). These problem categories incorporate most of the challenges that were found in the literature that this thesis is based upon.

The correlation problem scrutinizes correlations that can be found within data especially in a huge amount of data. However, who can answer whether such findings are causal and not accidently? This simple question is a question of vital importance for the entire Big Data paradigm. Imagine the outcome of a Big Data analysis that on Monday's people who drive to work rather than take public transportation are more likely to get the flu. Is it real or is it simply a problem of data quality and even if it is not a quality problem do we trust this finding? What we have to come up with are new ways to test the causality of connections in the real world much more frequently and earlier than before (Pentland, 2012). Indeed, Big Data is still only based upon samples. While it is the ambition of Big Data to dispense with the need for random sampling techniques by collecting 'everything there is', be aware that especially the online world is only a sub-sample of everything there is (Hilbert, 2016). Even nowadays, not the entire population is using the internet, Facebook or Twitter.

The human understanding problem refers to the notion that while finding correlations in data is one thing, understanding them in a way that allows you to build a new, better system is another. The use of Big Data for research purposes is generally questionable as it is arguably difficult to verify the correctness of the data and prove the correlations (Müller, et al., 2016). The author of the paper in which the basic problem categories are defined states "that 70 to 80 percent of the results that are found in the machine learning literature — which is a key Big Data scientific field — are probably wrong, because the researchers did not understand that they were overfitting the data" (Pentland, 2012). There is no final answer to this discussion, although using Big Data incorporates the responsibility to prove the results; for example, based upon different sources. The use of Big Data can have some interesting outcome seeing the example that online ad delivery could in some cases discriminate. A black-identifying name was 25% more likely to get an ad suggestive of an arrest record (i.e. ' …., arrested? '), thus raising questions concerning whether Google's advertising technology exposes racial bias in society and how ad and search technology can develop to assure racial fairness (Sweeney, 2013). This example shows that the use of Big Data incorporates the risk of using a black box and the user has to be fairly vigilant in respect to possible outcomes.

The provenance problem tackles the notion that the necessary data for Big Data applications is typically not owned by one company and that new types of collaboration are necessary to overcome such hurdles (Pentland, 2012). Analysis of Big Data is an interdisciplinary research where experts in different fields work together. The Big Data architecture has to serve the cooperation of different functions by gaining access to different kinds to complete the analytical objectives as fast and frictionless as possible. Often, it is a managerial problem to obtain the available data integrated within a company, given that different functions and lines of business have different interests and the big picture is lost. Extrapolating these challenges to integrate data across different enterprises emphasizes the importance of top management support and a sound use case; for example, this could be the extension of the supply chain from at the present supplier to customer to a new chain from supplier´s supplier to the customer´s customer. Expecting sharing of data between companies is awkward due to the need to gain an edge in business. Sharing data about their clients and operations threatens the culture of secrecy and competitiveness (Bajaj & Ramteke, 2014, p. 1882).

The privacy problem means that consumer has a right to prevent the collection and use of every bit of data that he leaves behind. Big Data privacy strategies and the governance of Big Data privacy have to be considered very closely. It is in the responsibility of the board and senior executives to ensure correct policies, processes and procedures and an appropriate skill set for the IT. Only the proper governance prevents data to be used in an intrusive and damaging way. An offensive strategy involves anonymizing data, it still supports statistical trending and analysis and supports privacy, enforcing disclosure and reminding the consumer of the exchange of services and data (ISACA, 2013). The EU plans to enforce a Privacy Impact Assessment by law if the processing of data entails outstanding risks for the individual. This could impact analysis concerning economic situation an outlook, residences, the state of health or personal preferences of individuals. In the case of Big Data initiative, it is necessary either to secure an agreement of the individual to store and use its data or it is necessary to make data anonymous. Therefore, it is necessary to delete identification attributes or aggregate particular characteristics but it has to be verified that the department is unable to rewind the data anonymization with justifiable effort (BITCOM, 2013).

Very often related to the privacy issue is the data security. For Big Data it is a very important topic. Even if state-of-the-art security mechanisms are applied there are concerns that open-source tools are not sufficiently secure, this includes the increasingly popular NoSQL open-source tools. “These open source tools have increased security risk because some fail to maintain a minimal level of security”. Technical security problems and data breaches are widely and publicly discussed especially when prominent companies or individuals are the victims. Still internal security is statistically the greatest risk like demonstrated by Edward Snowden, the individual who leaked classified information from the National Security Agency (Cole, et al., 2015). The data confidentiality hast to be watched especially in case services are outsourced. By relying on professionals to analyze such data the potential safety risks increase. For example, the transactional dataset generally includes a set of complete operating data to drive key business processes. Therefore, the Payment Card Industry (PCI) data security standard was developed and such data must not include sensitive information such as credit card numbers. The analysis of Big Data may be delivered to a third party for processing only when proper preventive measures are taken to protect such sensitive data (Chen, et al., 2014, p. 172).

A more technical but still very important issue is data quality and automatically generate the right metadata to describe what data is recorded and how it is recorded and measured (see 19, page 1878). Basic data quality factors can be and have to be checked at the time when the data is collected. This is not so easy because unlike the manually data collection process, where several checks are implemented in the graphical user interface to validate and the user helps to verify data, it is hardly possible to check unstructured data in a fast moving data stream. And the quality of the information that is included is not assessable until the data are used and put into context (Clarke, 2016, p. 78). And commonly Big Data analytics features the use of data for purposes independent of its original purpose. This problem is somewhat imminent. Data quality issues are exacerbated by the loss of context, which increases the likelihood of misinterpretation. The information that is stored relates to more than only the data and this is called metadata. The metadata and data life cycle management is essential to keep pace with the unprecedented rates and scales of data creation. Individuals contribute digital data in mediums comfortable to them like documents, drawings, pictures, etc., and with or without adequate metadata. However, especially unstructured data have to include some descriptive information, otherwise the interpretation is almost impossible (Bajaj & Ramteke, 2014, p. 1883).

Another hurdle that is often posted in the literature is finding data scientists or more generally the human resources problem also has to be named. Since Big Data is an emerging technology, it needs to attract organizations and youth with diverse new skill sets. Qualified workers not only need technical skills but also research, analytical, interpretive and creative ones. These skills need to be acquired or developed in individuals and this requires training programs. This challenge has direct impacts on the costs of Big Data projects as a shortage of qualified manpower result in higher salaries (Köhler & Meir-Huber, 2015, p. 127) and higher project costs. Furthermore, this problem cannot be solved easily as it will take time to educate the employees. Moreover, universities and undergraduate education need to focus more on analytics and data processing to produce skilled employees in this expertise. European countries like Germany or Austria are comparatively ill-equipped to satisfy domestic demand with internal source and the world's large developing economies (Brazil, Russia, India and China) produce 40% of global professionals with deep analytical skills (Hilbert, 2016).

Analyzing the surveys regarding Big Data challenges the Fortune 1000 survey is referring to the Provenance problem. When asked to name the most critical factor to ensuring successful business adoption, a clear plurality of 33.9% named business and technology partnership, up from 23.4% in 2015, strong business leadership and sponsorship was cited by 23.2% of firms and recognition that data is a shared corporate asset was cited by 10.7% of firms. Firms were clear that partnership and cooperation with business leadership are the keys to Big Data success (New Vantage Partners LLC, 2016). Questioned for the organization's three largest impediments to using Big Data for effective decision-making the respondents with 56% cited 'organizational silos' as their major problem in making better use of Big Data followed by 'shortage of skilled people' 51% and 'unstructured content in big data is too difficult to interpret' 42%, 'Big data is not viewed sufficiently strategically by senior management' 35% and 'the high cost of storing and manipulating large data sets' 17% (CapGemini, 2012)[1]. These findings refer to the Provenance and the human understanding Problem.

[1] Categories related to data processing are not listed

A slightly different conclusion can be drawn out of the Deloitte survey which is more focused on the decision-making process and the interviewees identified the missing clarity of available data as the main problem (40%) and that the relevant data is either not available (23%) or the decision-makers have no or limited access to the data (26%). However, in general the participating companies are satisfied with the data available concerning validity (77%), relevance (76%), reliability (76%) and actuality (67%) (Reker & Andersen, 2014). These are challenges that generally should be solved with a Big Data initiative.

Other surveys had the challenges more focused on the technical and ordinary organizational problems beside the privacy and security issue. It is difficult to tell whether this expresses the real concern of the participants or it emerged caused by the type of interview. The information concerning whether the problem categories were offered for the interviewee to choose from or free text was entered and the grouping was conducted afterwards by the author of the survey is missing.

60% of IT executives interviewed during the 2012 IDG Enterprise Big Data research (2012) believe that Big Data integration will be very/extremely challenging. The average respondent cited more than five challenges in 2012 and in 2014 the primary challenge was limited budget with 45%. The legacy issues second biggest challenge in 2012 with 36% and third security issues (34%) was supplanted in 2013 by the lack of skilled data scientists with almost 40 % of respondents. In 2012, the list was followed by development time (34%) and growing demand on storage capacity/infrastructure (32%). IT executives (42% vs. 28% LOB) care more about identifying business areas and processes where Big Data can have the greatest impact LOB, managers (47% vs. 32%) are concerned about having sufficient human capital. While most organizations employed database programmers (62%), business analysts (56%), engineers (51%) and data analysts (47%) in 2012 and planned to invest in skill sets necessary for Big Data deployments in the next 12-18 months, including data architects (30%), data analysts (29%), database programmers (26%), directors or managers of analytics (26%) and research analysts (26%) the survey of 2014 obtained that only 18% employed data scientists which explains the shift in the challenge ranking (IDG Enterprise, 2012) (IDG Enterprise, 2014).

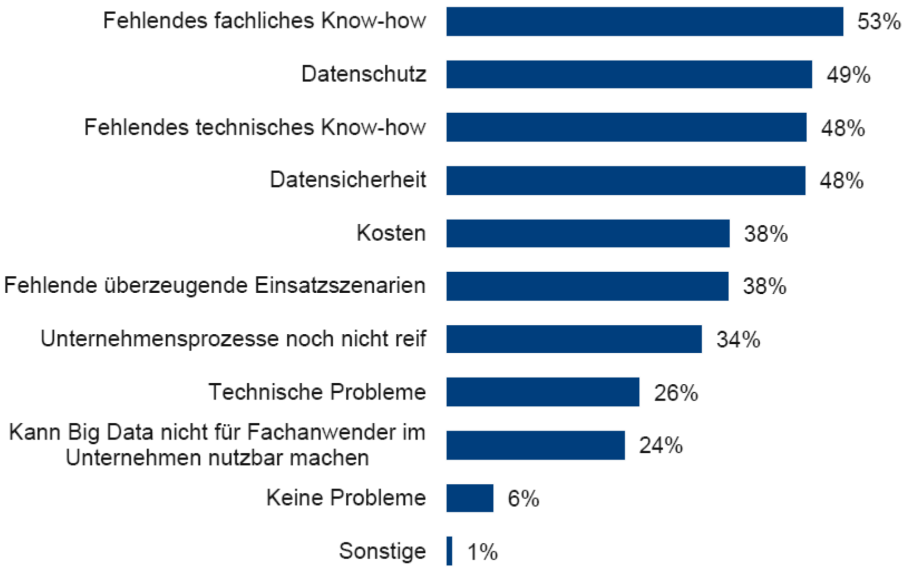


Figure 8 challenges for Big Data implementation according to BARC

Since 2012 missing technical and functional know-how is the top challenge (53% in 2014) according to BARC surveys. This can be linked to the IDG results that skilled workers are missing. The profession of data scientist is relatively new and it is obviously difficult to find talents and this is not a regional problem as companies of North America and Europe cite this issue with more than 50% as the top challenge. Data privacy and data security is very challenging for the organizations in terms of the functional embodiment, the technical implementation and the legal aspect. It should be highlighted that these concerns are not only relevant in Europe but also in North America where in fact it is slightly higher. The figures of 2013 doubled compared to the assessment of 2012 which most probably can be connected to the NSA scandal and this confirms the IDG findings. Nonetheless, only 10% consider their company's reputation at risk if their analytical practices would be published. A decline from 2012 (40%) to 2014 (38%) shows that company´s creativity rises in terms of the usage of Big Data or lack of sound use cases but it remains on a high level. The Big Data activity level of the LOB within the organizations will help to further reduce this challenge. The factor costs as a challenge of Big Data solutions is 2013 coming back to the level of 2011 to 38% whereas in 2012 it reached a peak of 45% (BARC, 2015, p. 36). This can be related to the IDG top finding limited budget because with rising costs at some point you always will surpass your budget. However, the term is too general to identify whether the problems are software/hardware costs or increasing expenditures for skilled employees.

Looking at the responds for challenges that hinder Big Data implementation of a survey among German SMEs it is obvious that a different mindset predominates. Top ranked is the category ‘no time’ followed by ‘needed competence is on the market not available’, ‘Big Data has a statistical focus’ and ‘missing competences within the company’ (Vossen, et al., 2015). Obviously the categories differ significantly from the IDG and BARC grouping and it is not realistic to predict whether respondents to these surveys wouldn´t have chosen ‘not time’ if it was available or if it is due to the size of the interviewed companies. In most reports a diversification of the results for SMEs versus big companies is missing. Only one paper related to the management of Big Data projects has the challenges in handling corporate data split over the enterprise size (BITCOM, 2013). The distribution among the different sizes of companies can be neglected. A variance can be ascertained in terms of disposition of and inconsistency of data, whereby it is arguable to assume that the size is not the only reason. On the other hand, ‘no time’ is somehow not so different to ‘limited budget’ and the aspects of the competence can be linked to the previous results. 70% of participants of an Austrian-based SME survey argued that it has to be shown that the implementation is necessary and ecologically worthwhile as a basic condition for them to invest into Big Data (Digital Networked Data, 2015). The study #big Data in Austrian names complexity as the main challenge for Big Data solutions (26%) followed by the challenge of data interpretation (24%), while high costs of a Big Data solution (19%), the lack of internal know-how (18%) and the security of data are ranked behind. Cloud-based solutions could help to reduce the challenge of scalability (Köhler & Meir-Huber, 2015, p. 78 f.). Furthermore, the second Austrian SME survey reports that a lack of know-how is the major reason not to start a Big Data initiative, followed by the expected high costs. It was stated that the term Big Data is too commonly used to say inflationary and that not even the “No Big Data” potentials have been lifted yet (Digital Networked Data, 2015, p. 7).

To summarize the top challenges are

- limited resources in terms of budget, time or high costs;
- lack of skilled workforce;
- data privacy concerns;
- data security issues; and
- lack of sound use cases.

4.3 (Big) Data strategy

The key to success in a Big Data project is strategy, not infrastructure (Roland Berger, 2013) – this was stated in 2013 in a column of Roland Berger and expressed that Big Data is not a product you can purchase, install and your Big Data project is implemented. For sure there are frameworks and procedures that help to successfully implement a Big Data project but the core message is that such a project has to be more a management driven strategy than an IT infrastructure project. A (Big) Data strategy should be a prerequisite for all Big Data initiatives. This strategy should be derived from the corporate strategy and should integrate the LOB and the IT. As for every strategy it is necessary to define the starting point, the time horizon, the vision, the aims and goals and the road map how to achieve these goals. The development of a Big Data strategy should include steps like (BITCOM, 2013, p. 38):

- Evaluating the processes in place regarding the IT and data architecture
- Identify the available data, define the relevance for the corporate strategy and look for blind spots
- Align the data architecture with the business objectives
- Develop a resource based view of the available analytical skills in relation to the available data and IT architecture
- Determine appropriate tools and technologies

It is necessary to remember that the Big Data initiative is not an end in itself. Like the corporate strategy the data strategy should help to gain competitive advantage and sharpen the unique selling proposition. One part of the Big Data structure should be an infrastructure strategy with following main pillars (Roland Berger, Sept. 2012):

- Software infrastructure: Companies have to determine which methods will be used for the data analyses. This software infrastructure comprises a Big Data platform such as Hadoop, connectors to the relevant data sourced in the data architecture, the analysis tools - for example Tableau or Hive - a visualization tool and if necessary a data mining or machine learning software. Some organization will have to define whether they will use the tools that are provided by their main software developer like SAP or they try to take advantage of a best-of-breed approach.
- Technical infrastructure: It is essentially a classic make-or-buy decision. The provided Big Data platforms can be rented to a very competitive price, although in case of very high

usage and a well established IT department it could still be cheaper to operate the Big Data platform in-house, especially considering that open-source technologies can be used. The business case developed as part of the data due diligence can shed light on that aspect.

- Data infrastructure/architecture: Companies must determine what the leading systems for each of their data sources will be.

As a part of the strategy, the organizational structure should be defined. Big Data certainly needs special technical know-how and this can be focused in one organizational entity like a competence center. For example, data mining and predictive analysis requires in-depth specialist knowledge of mathematical techniques and significant hands-on experience in applying them (Roland Berger, 2013). Furthermore, to gain the benefit of real value generation, it is certainly necessary to have deep functional and professional insights. This favors a decentralized organizational form like a business analyst supporting the line of business in their departments. Every company has to find the best fit for its culture and mostly it is somewhere in between.

A survey from 2013 showed that only 14% of the organizations have a data strategy and that 23% are planning to develop one. 50% of them have a dedicated BI competence centers (BICC) or a corresponding organizational entity. Surveys underpin the basic differentiation between small versus Big Data, whereby the user of Big Data should not be in charge of the data processing and querying. Companies that are very successful with Big Data implementation have their own BICC and at those enterprises that are not successful the single departments have to deal with BI on their own. This confirms the surveillance that a cross-functional position can maximize the exploited value out of data (BARC, 2013, p. 16). Once again, a significant difference between medium-sized companies and global players can be illustrated. Indeed, the NVP survey obtains that the Chief Data Officer (CDO) role is now well established and is an important voice in Fortune 1000 firms (New Vantage Partners LLC, 2016). In 2012, only 12% of firms reported naming a CDO. In the 2016 survey, this number jumped to 54%. In addition, 20% of firms now report that the CDO is the executive with primary ownership responsibility for Big Data initiatives for the firm. Presumably the medium-sized companies will follow this trend and also concentrate the competences and responsibilities within the organization.

Especially for SMEs a Big Data strategy is essentially to invest the budget as focused as possible. It has to be considered carefully that all these different applications need lot of expensive resources and outsource these tasks is a valid option considering the possibilities a platform as a service offers (see 5). Still the organization needs a unit who is in charge for the data strategy and who constantly improves the value generation.

5 Big Data market

So far the Big Data paradigm, the findings related to Big Data in general and aspects for the implementation have been discussed. This chapter addresses findings related to the Big Data vendors that were covered in the surveys. Thereafter a Big Data solution is described that according to literature is appropriate for SMEs and based on this solution three vendor solutions are compared. As this thesis is dedicated to SMEs in Austria the market size for this segment is at the beginning of this chapter outlined.

5.1 Austria's Big Data market size

According to IDC, the global Big Data market was worth about 12.6 billion USD in 2013 and should grow to 32.4 billion USD in 2017. This would reflect a 27% compound annual growth rate (CAGR) and takes the entire value chain of Big Data into account. Hardware for cloud infrastructure has the largest share with a growth rate of almost 49% from 2012 to 2017, followed from storage solutions with 38%, Big Data services 27% and the Big Data software solutions with 21% (Köhler & Meir-Huber, 2015, p. 73). It is expected that the European market is more conservative due to privacy regulations and concerns about data security. According to Experton Group the market size for Germany will grow from 1.3 billion Euro in 2015 to 2.2 in 2017 and 3.2 in 2019 that equates to 24% CAGR. They expect that services will contribute with 56%, software with 23% and hardware only with 21% (Experton Group AG, 2016). Based on the available documents the significant divergence in the market segments cannot be explained. For Austria a market size from almost 19 MM Euro for 2012 was calculated by IDC and the CAGR is 31%. This would mean 73 MM in 2017 (Köhler & Meir-Huber, 2015, p. 74). This estimation can be considered as conservative as in general the German market is approximately 10x the Austrian market. Experton Group also estimated the market size for Switzerland in 2017 with roughly 300 MM Euro. The market size for SMEs is approximately 40% (Experton Group AG, 2016, p. 18). Using the Experton Group´s market distribution, as it is the more recent one, the market for Big Data for SMEs in Austria is about

30MM with services contributing 17 MM Euro, for software about 7 MM and hardware 6 MM.

5.2 Big Data vendors in Austria

In terms of the vendor landscape technology giants such as IBM (21%), Oracle (12%) followed by HP and Microsoft are most often cited by respondents as thought leaders. The remaining cited vendors were all under ten percent, including next-generation vendors such as Hadoop (4 percent), Amazon (3 percent) and Google (3 percent) (IDG Enterprise, 2014). This is confirmed by a BARC study where participants were asked to cite vendors that are relate to Big Data and IBM and Oracle are top ranked (see Figure 9). The authors disclose that the survey cannot be seen as a representative market survey and it has to be disclosed that some vendors have sponsored the survey (BARC, 2013, p. 37). The list of Big Data consultants is also led by IBM (48%) followed by Microsoft (29%), Capgemini (16%), Accenture (15%), etc.

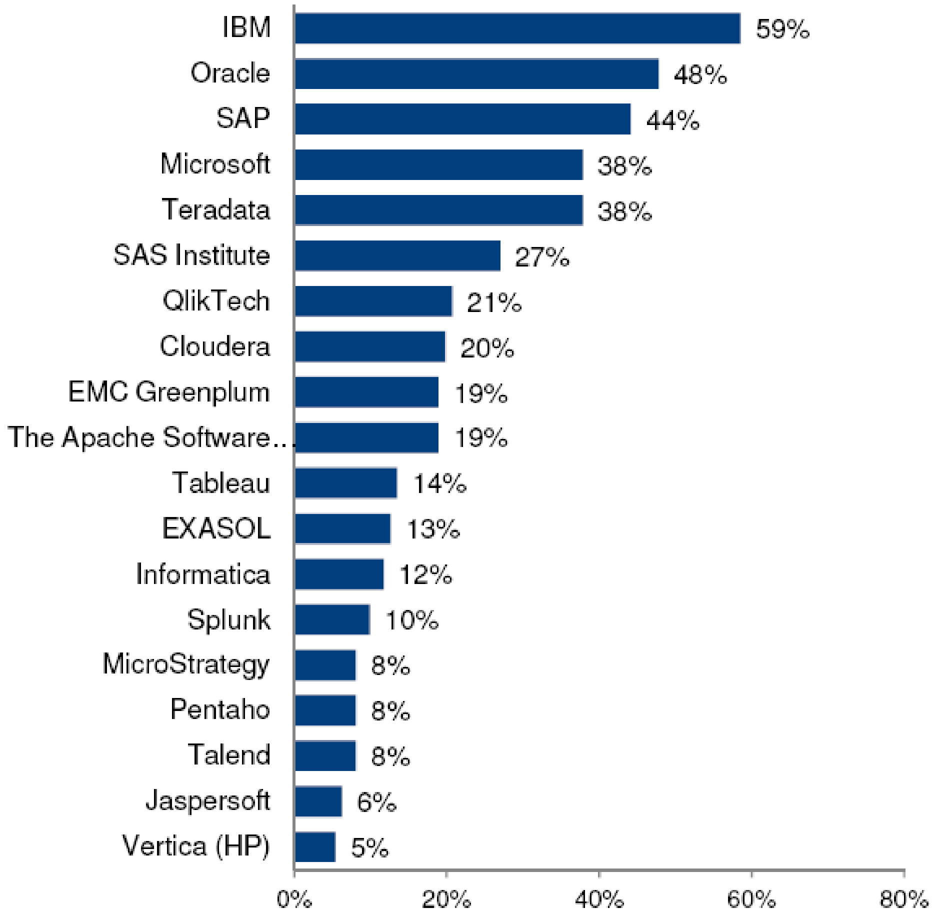


Figure 9 Vendors that are related to Big Data according to BARC 2013

To identify vendors that provide solutions for SMEs in the Big Data field it is necessary to define a basic solution that has the potential to fulfill most of the requirements for SMEs. In "Is Big Data too Big for SMEs?" such a solution is described. Three distinct technical categories were identified that form a pivotal role in creating a low-cost data utilization framework for SMEs (Rising, et al., 2014, p. 8):

- a scalable cloud network of servers - allowing for developing a small scale operation, that can then be expanded upon.
- a framework for data management, processing and storage, such as Hadoop and MapReduce.
- Analytics software, such as Tableau, SAS, SPSS, etc.

The survey "#big data in Austria" expects SMEs to primarily use cloud based solutions and that industry solutions will prevail (Köhler & Meir-Huber, 2015, p. 75). In general cloud services have been defined as a multi-layered infrastructure and are classified as Software as a Service (SaaS), Platform as a Service (PaaS) or Infrastructure as a Service (Iaas). IaaS means that the cloud service provider provides storage and computing capacity and the customers deploy and run their own applications on their software platform and own operational responsibility. Using PaaS means that the provider also takes care of the software platform for systems. SaaS is the next stage where applications are developed and hosted by the provider and delivered online via a web browser offering traditional desktop functionality (Delibašic, et al., 2014, p. 39). Big Data as service is based on this SaaS idea and means that the provider supports the customer with consulting service, customized application development and application management. Figure 10 explains how the responsibilities are split between the customer and the vendor (IBM, 2015, p. 61).

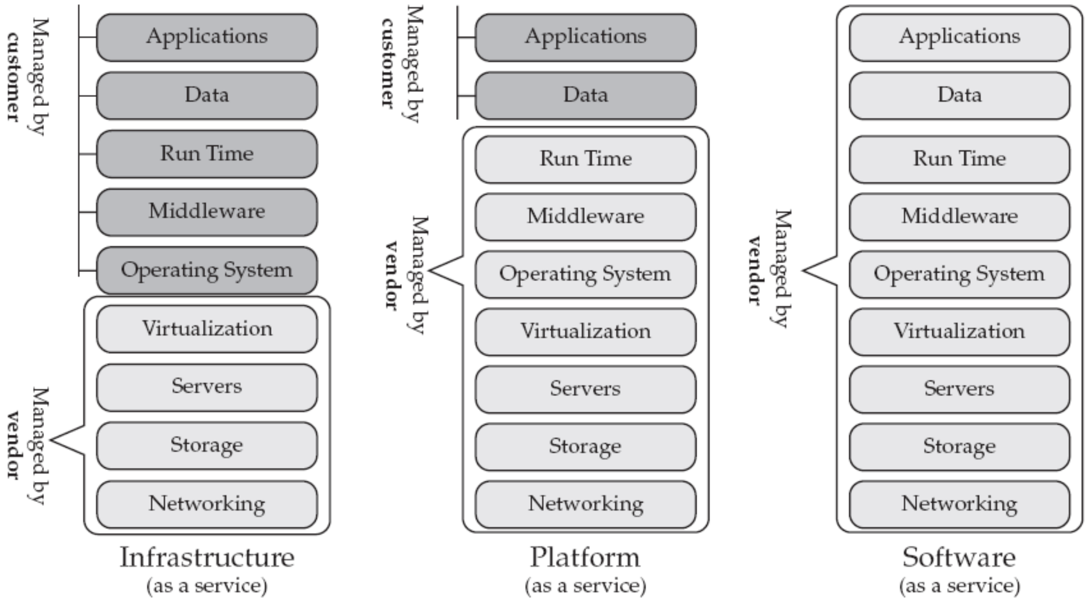


Figure 10 shared responsibilities at each as-a-service level according to IBM

The Big Data Vendor Landscape study separates the suppliers into 10 groups from data storage supplier to Big Data consultants and evaluates 250 vendors (Experton Group AG, 2016, p. 19). An Austrian based market research indentified 60 participants for the Austrian market (Köhler & Meir-Huber, 2015, p. 40 f.). Based upon these comparisons and the basic assumption that SMEs will prefer a cloud based solution a shortlist with IBM, Microsoft Azur and Oracle Big Data Cloud Services was composed. These companies provide at least a PaaS. The basic idea of this short list is to evaluate based upon publicly available whitepapers whether the basic Big Data challenges for SMEs are addressed with these services. All of these vendors provide tools that fit to the proposed framework for SMEs (Rising, et al., 2014, p. 8). All these companies are trustworthy enough to assume that they are able to provide a cloud network of servers like they promote in their whitepapers and on their homepages. For the Microsoft Azure platform for example the user can update the performance level online. This enables organizations to change the server landscape as the needs of the application change. This is especially a huge advantage for SMEs as it is possible to use very powerful platforms without huge basic investments.

	IBM	Microsoft	Oracle
Scalable cloud network of servers	OK	OK	OK
Data management framework	Watson Foundation	Azur	Oracle Big Data Cloud Services
Analytical software			Oracle BI suite

Table 5 vendors PaaS matrix for SMEs

IBM provides with the Watson Foundations a toolkit that enables the user to use a DWH, stream computing or a Hadoop System and provides BI and analytical tools as well as content analysis, decision management, planning and forecasting software (IBM, 2015, p. 121 f.). Microsoft provides a data management framework within Microsoft Azur that allows the user to store data on an SQL Server or on no SQL databases like DocumentDB ad document database, Azur Tables a key/value database or Azure HDInsight which is a managed service that supports MapReduce and HBase. Azure´s toolkits on the analytical side include SQL Server Reporting Services, Power BI, SharePoint Server and tools for data mining and machine learning. Big Data Cloud Service: Hadoop and Spark are delivered as an automated Cloud Service (Chappell, 2014). Oracle PaaS Cloud Services includes the Cloudera Data Hub Edition, Oracle Big Data Connectors, Oracle Spatial and Graph, Oracle Data Integrator with Advanced Big Data Option and Database Cloud Service. Big Data Discovery Cloud Service is used to explore data residing in Hadoop. Business Intelligence Cloud Service provide ad-hoc query and data visualization and Big Data SQL Cloud Service an optimal solution for using Oracle Database SQL to query data residing in SQL and noSQL databases. The Big Data Cloud Service from Oracle can be extended to use R as an NLP engine for machine learning and an IoT cloud service (Oracle, 2016, p. 26).

According to these whitepapers, it is fair to say that the services are very flexible and powerful to satisfy the needs of most SMEs, although the complexity of the Big Data philosophy is not and presumably cannot be reduced. The users of these services still have to be very IT-affine or the companies need consulting to implement the solution on the PaaS. Nevertheless the possibilities that SMEs have with these Big Data PaaSs seem to be compelling enough to refuse arguments that the technology is hindering SMEs to use Big Data.

6 Interpretation, Discussion, Future prospects

Some information technology vendors tend to over-promote technology opportunities, which indeed has happened with Big Data. Big Data is entering the Gartner Hype Cycle (Gartner, 2014) "Trough of Disillusionment", although there are generally few specifics about how to use the new data streams to gain value (Delibašic, et al., 2014). Beside surveys among large companies – which tend to result in a positive picture about the possibilities of Big Data – other publications paint a more skeptical picture. It is very difficult to compare all of these different surveys based upon the published and already-aggregated data and many questions remain open when interpreting the findings.

The surveys confirm the general notion that large multinationals mostly based in North America use Big Data and those companies plan to further invest in this topic, while smaller and European-based companies lag behind. In terms of SMEs, the results of the surveys are not very significant. The participation rate of surveys focusing on SMEs is very low and this could be interpreted as a lack of interest in the Big Data topic. Furthermore, the results of such studies show that only 10% of SMEs invest in Big Data. The definition of Big Data along the criteria volume, velocity and variety implies a certain but non-defined data volume and hence organization size. With a steady increase in performance of traditional database management systems, SMEs can stick to traditional analytical platforms as long as the data volume can be processed and queried with an acceptable performance. The surveys suggest that only 10% of all participants plan to analyze data near real time and only a small portion of these companies are SMEs. A common finding among all surveys is that business data – e.g. operation and transaction data – represents the most important source of data analytics, followed by log data and human interaction data, although only 10% of companies use all three categories. These results somehow explain why Big Data is not the first priority for SMEs as a real need along the 3Vs could not be shown.

When asked about the most important steps that need to be taken to start Big Data in their company 46% of the respondents among Austrian SMEs answered to increase Big Data know-how and identify processes where the Big Data impact is most severe followed by steps related to hard- and software tasks (Köhler & Meir-Huber, 2015). The reasons behind the lack of interest are:

1) Lack of understanding and clarity: The knowledge of big data analytics by SME representatives is not very high and this hinders implementation of Big Data. It is obvious that SMEs will not embark on a Big Data paradigm which they are not familiar with. However, the term Big Data itself is far too general to really help SMEs to focus on its benefits even worth it is inflationary used and there is a risk of stigmatizing Big Data as management hype.
2) Lack of use cases and success stories: SMEs are aware of the possibilities data analysis could bring but they are not convinced that Big Data could also be a use case for their business. It is questionable whether the documented use cases and success stories are sufficiently sound to convince them otherwise.
3) Lack of knowledge and skilled labor: For new technologies to develop within an organization it is necessary to set the managerial conditions accordingly. This means to educate employees to learn how to handle Big Data technologies and over time the usage will increase due to the advantages these tools bring. This education will take some time and the lack of affordable skilled workforce hits in the first place SMEs who are unable to compete for talents with global players.
4) Lack of necessity: The problem with Big Data is that the characteristics are focused on large global companies. SMEs do not really fit into this framework and they do not need to. SMEs can for sure benefit from new techniques and technologies that are embedded in the Big Data universe but the intrinsic conservatism, the workload of the daily business and the attitude to lean management hinders SMEs to further embark on the Big Data paradigm.
5) Lack of time and money: Like every project a Big Data initiative competes with available resources within the organization. As long as Big Data is not seen as strategic investment the surveys reveal that the business cases are not good enough to start Big Data projects particularly for SMEs but the notion remains that when they would start it would be beneficiary. This couldn´t be confirmed or proven wrong.

The Big Data goals (see 3.3) revealed by the surveys – such as improving the quality of decision-making, developing greater business insights, planning and steering operational processes as well as more detailed and faster analysis – do not necessarily only relate to Big Data and could also be achieved with Small Data. According to the surveys, Big Data technologies like noSQL databases or the Hadoop framework are in the minority in terms of data analytics, although companies plan to increase their usage. They also plan to use more data sources, especially social media data and thus they need to use more techniques such as network and text data analysis. Documented use cases highlight that predictive analysis is a very important technique and that use cases of the manufacturing field are very focused on improving quality and services based upon these new algorithms. To benefit from these use cases, one prerequisite is that the facilities are equipped with sensors and ideas such as Industry 4.0 and the Internet of Things become mainstream. This could bring Big Data to companies without the need for these organizations to become involved with the technical aspects of Big Data, given that the vendors of the equipment provide a service where the customer – thus the manufacturer of goods – can use predictive analysis and the advantages of Big Data. Indeed, this would particularly help SMEs.

The surveys show that the existence of a data strategy is a very important success factor for Big Data implementation. Such a strategy can be seen as a prerequisite for a Big Data initiative and it has to be aligned with the business strategy. Top management support is very crucial and a dedicated department to concentrate the necessary skills is pivotal, or at least a dedicated organizational role responsible for this data strategy within smaller organizations. In terms of Big Data implementation, the explorative innovative character has to be considered and it needs a multidisciplinary team equipped with sufficient time and resources to find ways to gain value out of the data; for example, during a prototyping phase.

According to the studies, a very prominent hurdle related to Big Data is the security challenge. In recent years, several data leaks have raised IT managers' awareness of data security, whereas the privacy concerns are not so well focused yet but no less important. Wherever possible, anonymized data should be used as the privacy regulations will be hardened, especially in Europe. Hurdles like the correlation problem and the human understanding problem come along with the usage of Big Data. Users should not be too sanguine when it comes to the interpretation of the data glut: just because a huge amount of data is available, this does not mean that the right data is available. The results of Big Data analysis have to be double-checked and critically questioned. The IT managers responsible for Big Data have to take care of meta-data and data management. It is essential that the acquired data is reasonably described and that the data is consistent and correct.

IT companies like IBM, Microsoft, Oracle, etc. provide tool kits of Big Data technologies hosted on their servers that enable other companies to start immediately with Big Data. This is a big advantage especially for SMEs who normally don´t employee huge IT departments. These services provided as platform as a service helps to keep the initial costs for a Big Data initiative low. The Big Data project team can concentrate on idea generation and can easily use SQL databases combined with noSQL data storage and deploy a machine learning module on a very small and therefore cheap environment. Of course the team has to have the necessary skills to do the technical implementation but during a prototyping phase the possibility to start immediately is very powerful. For the production phase it is still possible to operate the system within your IT department.

Nonetheless, it could be shown that business opportunities arise along the Big Data value chain. It is more about possible values that can be generated out of Big Data rather than how to solve Big Data issue technically. Especially SMEs lack analytical skills to identify business opportunities that are hidden within their processes and in their data. However, Big Data is not necessarily the key to identifying and unlocking these possibilities. Accordingly, SMEs should not rush to jump on the Big Data train; rather, they should start to intensify their Small Data abilities. Accordingly, such companies will slowly engage with new techniques and methods to think about data and data analytics, which will ultimately lead them to more data acquisition, a greater variety of data and finally the need to enhance the speed of this process. A company equal of size that has performance issues with RDBMS would wish change to an analytical platform if this solves the velocity problem, although whether this will ultimately be of a magnitude where it is possible or necessary to entitle it as Big Data or not is debatable. The Big Data universe includes several tools that supplement and enhance the actual Small Data toolset and the provider of Big Data platforms enable a very cheap, secure and easy-to-access service whereby organizations can primarily concentrate on the business aspects in the long run and do not have to care about technical issues. For SMEs, the challenge is to change the habits of decision-making as 52% of decisions are made based upon insufficient information, even if companies are already fairly comfortable with the validity, reliability, relevance and actuality of the available data (Reker & Andersen, 2014). The active involvement in the hunt for data-driven value generation will help companies to make their processes more efficient and personalize their services. In this respect, data is somehow the new oil of the economy as it enables organizations to work more efficiently, productively and smoothly. Companies have to care about their data to further increase productivity and if they do not care about data the organization is at risk. Nowadays, SMEs have the vital opportunity to outperform established companies by focusing on and building the organization around data, which enables small ventures to disrupt entire industries. *"Data-enabled disruption may represent an anomaly to the existing theory, but it's here — and it's here to stay"* (Wessel, 2016).

7 Bibliography

Agrawal, D. et al., 2012. *Challenges and Opportunities with Big Data: A white paper prepared for the Computing Community Consortium committee of the Computing Research Association..* [Online]
Available at: http://cra.org/ccc/resources/ccc-led-whitepapers/
[Accessed 05 2016].

Bajaj, R. H. & Ramteke, P., 2014. Big Data – The New Era of Data. *International Journal of Computer Science and Information Technologies*.

BARC, 2013. *Big Data survey europe,* s.l.: s.n.

BARC, 2014. *Big Data Analytics,* s.l.: s.n.

BARC, 2014. *Datenmanagement im Wandel,* s.l.: s.n.

BARC, 2015. *Big Data Use Cases; Getting real on data monetization.* [Online]
Available at: http://barc.de/docs/big-data-use-cases
[Zugriff am 10 05 2016].

Bean, R., 2016. Just Using Big Data Isn’t Enough Anymore. *Harvard Business Review*.

BITCOM, 2013. *Management von Big Data Projekten,* s.l.: s.n.

BITCOM, 2015. *Big Data und Geschäftsmodell-Innovationen in der Praxis: 40+ Beispiele,* s.l.: s.n.

CapGemini, 2012. *The Deciding Factor: Big Data & Decision Making,* s.l.: s.n.

Cecere, L., 2012. *Big Data Report Supply chain,* s.l.: s.n.

Chappell, D., 2014. *Microsoft Azure Data Technologies: An Overview.* [Online]
Available at: http://www.davidchappell.com/writing/white_papers/Azure_Data_Overview-Chappell_v2.0.pdf
[Accessed May 2016].

Chen, M., Shiwen, M. & Yunhao, L., 2014. Big Data: A Survey. *Mobile Netw Appl*, pp. 171 - 209.

Clarke, R., 2016. Big data, big risks. *Info Systems J, 26: 77–90. doi: 10.1111/isj.12088*.

Cole, D., Nelson, J. & McDaniel, B., 2015. *Benefits and Risks of Big Data.* [Online]
Available at: http://aisel.aisnet.org/sais2015/26

Coleman, S. et al., 2016. How Can SMEs Benefit from Big Data?. *Challenges and a Path Forward. Qual. Reliab. Engng. Int., doi: 10.1002/qre.2008.*

Delibašic, B. et al., 2014. *Decision Support Systems V – Big Data Analytics for Decision Making.* s.l.:s.n.

Digital Networked Data, 2015. *Big Data und Data-driven business für KMU,* s.l.: s.n.

Experton Group AG, 2016. *Big Data Vendor Benchmark 2016.* [Online]
Available at: https://www.t-systems.com/blob/252534/33b3ad7f00c41f9c4c0164111a5f2d23/dlmf-wp-bigdata-big-data-vendor-benchmark-2016-data.pdf
[Accessed 2016].

Finlay, S., 2014. *Predictive analytics, data mining and big data Mythos, Misconceptions and Methods.* s.l.:Palgrave Macmillan UK.

Franks, B., 2012. TO SUCCEED WITH BIG DATA. *Harvard Business Review.*

Frauenhofer IAIS, 2012. *BIG DATA – Vorsprung durch Wissen Innovationspotentiale,* s.l.: s.n.

Gartner, 2014. *Gartner's 2014 Hype Cycle for Emerging Technologies Maps the Journey to Digital Business.* [Online]
Available at: http://www.gartner.com/newsroom/id/2819918
[Accessed 2016].

Hilbert, M., 2016. Big Data for Development: A Review of Promises and Challenges. *Development Policy Review*, pp. 135-174.

IBM 4Vs, n.d. *www.ibmbigdatahub.com.* [Online]
Available at: http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg

IBM, 2015. *Big Data beyonde the Hype.* [Online]
Available at: http://www-01.ibm.com/software/data/bigdata/
[Accessed April 2016].

IDC, 2012. *Worldwid Big Data technonolgy and Service 2012-2015 forecast.* [Online]
Available at: www.idc.com
[Accessed May 2016].

IDG Enterprise, 2012. *Big Data Initiatives High Priority for Enterprises but Majority Will Face Implementation Challenges.* [Online]
Available at: http://www.idgenterprise.com/news/press-release/big-data-initiatives-high-priority-for-enterprises-but-majority-will-face-implementation-challenges/
[Accessed May 2016].

IDG Enterprise, 2014. *Big Data Survey,* s.l.: s.n.

Inmon, W., 2005. *Building the Data Warehouse.* 4th ed. Indianapolis: s.n.

ISACA, 2013. *Privacy & Big Data, an ISACA whitepaper.* [Online]
Available at: an ISACA white paper; http://www.isaca.org/Knowledge-Center/Research/Documents/Privacy-and-Big-Data_whp_Eng_0813.pdf
[Accessed May 2016].

Köhler, M. & Meir-Huber, M., 2015. *#Big Data in Austria,* s.l.: s.n.

Koronios, A., Gao, J. & Selle, S., 2014. *Big Data Project Success - a meta analysis.* [Online]
Available at: http://aisel.aisnet.org/pacis2014/376
[Accessed 05 2016].

Kune, R. et al., 2016. *The anatomy of big data computing Softw. Pract. Exper.,* s.l.: s.n.

Laney D, 2001. *3-d data manamement: controlling data volume, velocity and variety.* [Online]
Available at: https://blogs.gartner.com

McKinsey Global Institute, 2011. *Big data: The next frontier for innovation, competition and productivity,* s.l.: s.n.

Müller, O., Junglas, I., Brocke, J. & Debortoli, S., 2016. Utilizing big data analytics for information systems research: challenges, promises and guidelines. *European Journal of Information Systems advance*, 9 February.

NESSI, 2012. *Big Data - A New World of Opportunities.* [Online]
Available at: http://www.nessi-europe.eu/
[Accessed 07 03 2106].

New Vantage Partners LLC, 2016. Big Data Executive Survey. *Harvard Business Review*.

Oracle, 2016. *An Enterprise Architect's Guide to Big Data.* [Online]
Available at: http://www.oracle.com/technetwork/topics/entarch/articles/oea-big-data-guide-1522052.pdf?ssSourceSiteId=ocomen
[Accessed April 2016].

Pentland, A., 2012. The promise and challenge of Big Data. *Harvard Business Review*.

Power, D. J., 2014. *What are some use cases with expanded data sources?.* [Online]
Available at: http://dssresources.com/faq/index.php?action=artikel&id=303
[Accessed 05 2016].

Reker, J. & Andersen, N., 2014. *Data Analytics im Mittelstand. Die Evolution der Entscheidungsfindung,* s.l.: Deloitte & Touche GmbH Wirtschaftsprüfungsgesellschaft.

Rising, C. J., Kristensen, M. & Tjerrild-Hansen, S., 2014. *Is Big Data too Big for SMEs?,* s.l.: Stanford University.

Roland Berger, 2013. *Experiment with Big Data.* [Online]
Available at:
http://www.rolandberger.com/expertise/functional_issues/information_management
[Accessed 04 2016].

Roland Berger, Sept. 2012. *Big Data – from buzzword to strategy,* s.l.: s.n.

Supply Chain Insights LLC, 2012. *Big Data Go Big or Go Home?,* s.l.: s.n.

Sweeney, L., 2013. *Discrimination in Online Ad Delivery,* s.l.: Harvard University.

Vossen, G., Lechtenbörger, J. & Fekete, D., 2015. *Big Data in kleinen und mittleren Unternehmen — eine empirische Bestandsaufnahme,* Münster, Germany: Westfälische Wilhelms-Universität.

Wessel, M., 2016. How Big Data Is Changing Disruptive Innovation. *Harvard Business Review*, 27 January.

Wikipedia, 2016. *Big Data.* [Online]
Available at: https://en.wikipedia.org/wiki/Big_data
[Accessed 15 03 2016].

WKÖ, 2015. *Klein- und Mittelbetriebe in Österreich, Wirtschaftskammer Österreich.* [Online]
Available at:
https://www.wko.at/Content.Node/Interessenvertretung/ZahlenDatenFakten/KMU_Definition.html
[Accessed 05 2016].