Challenges in Data Acquisition and Management in Big Data Environments

Daniel Staegemann[1][a], Matthias Volk[1][b], Akanksha Saxena[1], Matthias Pohl[1][c],
Abdulrahman Nahhas[1][d], Robert Häusler[1], Mohammad Abdallah[2][e], Sascha Bosse[1][f],
Naoum Jamous[1] and Klaus Turowski[1]

[1]*Magdeburg Research and Competence Cluster Very Large Business Applications, Faculty of Computer Science, Otto-von-Guericke University Magdeburg, Magdeburg, Germany*

[2]*Department of Software Engineering, Al-Zaytoonah University of Jordan, Amman, Jordan*

{daniel.staegemann, matthias.volk, matthias.pohl, abdulrahman.nahhas, robert.haeusler, sascha.bosse, naoum.jamous,

Keywords: Big Data, Data Quality, Data Acquisition, Data Management, Literature Review.

Abstract: In the recent years, the term big data has attracted a lot of attention. It refers to the processing of data that is characterized mainly by 4Vs, namely volume, velocity, variety and veracity. The need for collecting and analysing big data has increased manifolds these days as organizations want to derive meaningful information out of any data that is available and create value for the business. A challenge that comes with big data is inferior data quality due to which a lot of time is spent on data cleaning. One prerequisite for solving data quality issues is to understand the reasons for their occurrence. In this paper, we discuss various issues that cause reduced quality of the data during the acquisition and management. Furthermore, we extend the research to categorize the quality of data with respect to the identified issues.

1 INTRODUCTION

In recent times, the amount of data produced and utilized has increased manifolds (Chen et al. 2014; Müller et al. 2018). A survey by Lehmann et al. (2016) revealed that 98.4 percent of IT decision makers accept that the data is going to increase in the coming years. Researches acknowledge that business owners and higher management are under duress to use more analytics to support their business decisions (Ebner et al. 2014; Khan et al. 2014; LaValle et al. 2011). Therefore, considerable investments are being made on big data applications (Lee 2017), resulting in a positive impact on the involved businesses (Bughin 2016; Müller et al. 2018), while at the same time posing considerable challenges (Al-Sai et al. 2020; Günther et al. 2017; Staegemann et al. 2020a). This development has led to several types of research (NIST 2019a) and a plethora of case studies (Volk et al. 2020a) in the field of big data. For example, big data is being used or discussed in the domains of media (Pentzold et al. 2019), political science (Couldry and Mejias 2019), education (Häusler et al. 2020) or the medical domain (Smith and Nichols 2018). While the focus is on analysing and extracting the information from the data to create value out of it, one major roadblock that every organization faces is the reduced quality of the data during the acquisition and management (Chen et al. 2014). Big data is used for data analysis and data exploration to derive hidden meaning from the data (Ward and Barker 2013). Therefore, it often leads to the discovery, evaluation or implementation of hypotheses on real systems (NIST 2019a). However, when the data quality is low, the acquired data needs to be pre-processed before it can be utilized to generate any meaningful insight.

[a] https://orcid.org/0000-0001-9957-1003
[b] https://orcid.org/0000-0002-4835-919X
[c] https://orcid.org/0000-0002-6241-7675
[d] https://orcid.org/0000-0002-1019-3569
[e] https://orcid.org/0000-0002-3643-0104
[f] https://orcid.org/0000-0002-2490-363X

Staegemann, D., Volk, M., Saxena, A., Pohl, M., Nahhas, A., Häusler, R., Abdallah, M., Bosse, S., Jamous, N. and Turowski, K.
Challenges in Data Acquisition and Management in Big Data Environments.
DOI: 10.5220/0010429001930204
In *Proceedings of the 6th International Conference on Internet of Things, Big Data and Security (IoTBDS 2021)*, pages 193-204
ISBN: 978-989-758-504-3

Analysis done on unprocessed data may lead to incorrect or even biased results (Ebner et al. 2014). In fact, 94 percent of respondents of the survey conducted by Lehmann et al. (2016) believe that reduced data quality causes loss to the business value. Hence, data preparation becomes an extremely critical part of the data analysis process (Géczy 2014) and is required to be done before the data is utilized further to derive any meaning or business value. There are mainly two possible issues with the data preparation. First, the data pre-processing tasks can take a lot of productive time and can cause unnecessary delays. Second, it is assumed that interpolating data is better than missing data (García et al. 2015; Li et al. 2007). However, imputed data values might not reflect the exact results. The assumption of interpolating data deduces that the result might not be accurate, but close to accurate. Nonetheless, it is the delay caused by data pre-processing that poses the biggest challenge in the big data environment, especially when there is a need to process data in real time (Zakir et al. 2015). Therefore, it becomes extremely important that the necessary focus is given to how data is collected and managed in order to reduce the consecutive effort spent on data preparation. As a result, more effort can be given on creating actual value out of the data (Chen et al. 2014; García et al. 2015). The issue of reduced data quality leads to the research question.

Which issues in data acquisition and management cause quality reduction in a big data context?

This requires evaluating the processes involved in big data acquisition and management to identify the reasons that might reduce the data quality. This helps businesses to be better equipped to directly address the issues that can affect the data quality (Abdallah et al. 2020). This would also reduce the amount of time spent on data pre-processing. Therefore, a literature review was performed to identify possible issues during data acquisition and management that might lead to reduced data quality in big data environments. The content of this paper is as follows. In the second section, the concept of big data is detailed, in order to help readers to understand it in the context of the publication at hand. Afterward, big data quality is discussed, the relevant dimensions identified and also the importance of each of those highlighted. In the following two sections, identified issues related to the data acquisition and its management are presented. Finally, a conclusion is drawn and possible directions for future research contemplated.

2 BIG DATA

Big data is mostly defined by researchers and organizations based on their own focus and requirements (Mauro et al. 2015), resulting in a plethora of slightly varying definitions (Press 2014; Volk et al. 2020b). However, in the context of this paper, big data is characterized by four key properties (Ebner et al. 2014; NIST 2019a). *Volume* signifies the massive amount of data that are produced, acquired and processed. This is reflected by the industry requirement of processing large data sets to infer better and more meaningful insights, which in turn provide more value to the businesses (Demchenko et al. 2013). *Velocity* describes two facets. On the one hand, the rate at which data are incoming, and on the other hand, the requirements regarding their timely processing, respectively the responsiveness to requests by the user. The large volumes of data are being acquired at an increasing frequency. In many cases, those data are collected in real-time resulting in a high velocity, and often, big data analytics also require real-time processing (Narasimhan and Bhuvaneshwar 2014). *Variety* refers to the multitude of types and formats of the processed data. Besides structured data, there can also be semi-structured and unstructured data. Furthermore, big data also caters to homogeneous and heterogeneous data sources as well as varying data types and measuring units (Kaisler et al. 2013). *Variability* means that the velocity, volume and variety of the data but also the content can vary from time to time and is often not constant (Katal et al. 2013). It is required that a big data solution implementation caters to all of these characteristics. To address these requirements, several paradigms have been suggested, such as scaling the conventional systems horizontally and vertically (NIST 2019a). There are several papers present on different paradigms of big data, such as (Cai and Zhu 2015; Gudivada et al. 2017; Staegemann et al. 2020b; UNECE Big Data Quality Task Team 2014) and it is beyond the scope of this paper to discuss these solutions. However, for the contribution at hand, the focus will be on evaluating the reasons for reduced data quality in big data environments. For this, it is important to delve deep into understanding how data quality is defined and why it is important.

3 BIG DATA QUALITY

Data quality is the characteristic that the data should possess in order to be harnessed by its user. The quality of data required could vary for different organizations (Pipino et al. 2002). However, it is necessary to determine a minimum of quality for data to be deemed usable (NIST 2019c). Data quality is measured on certain dimensions, which are slightly varying across different publications (Li et al. 2007; Merino et al. 2016; NIST 2019a; Pipino et al. 2002), as shown in Figure 1. While there is no fixed set of universally required properties, in the following, common factors will be presented.

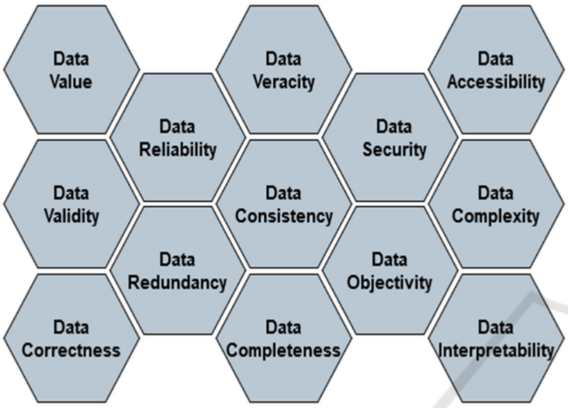


Figure 1: Dimensions of data quality.

3.1 Data Quality Dimensions

Even though the envisioned analytics endeavours might differ, the need for high quality input data is usually common. To specify this vague aspiration, it appears to be sensible to formulate a set of criteria that in conjunction constitute said quality. However, due to the sometimes fuzzy nature of the domain and its definitions, the presented dimensions are not always clearly distinct and might be similar or interwoven with each other. Yet, the following constitutes an attempt at describing the most important factors. While *data completeness* indicates to which degree there are missing values when collecting or storing data that may prove to be critical during the analysis phase (Cai and Zhu 2015; Ezzine and Benhlima 2018; Li et al. 2007), *data veracity* refers to the available data's accuracy. Therefore, a high veracity means that the data gives an truthful representation of the corresponding real world object (NIST 2019a). Another dimension is the *data consistency*, which means that the information conveyed by the data does not change throughout its processing and that there are no contradictions within the data (Cai and Zhu 2015; Li et al. 2007). A related topic is the *data redundancy*. It refers to the practice of storing the same data in more than one place, increasing the risk of discrepancies. For this reason, not considering backups, usually data with same value should be acquired and stored only once (Li et al. 2007). As a less technical property, the *data reliability* depends on the credibility of the source the data is collected from. While naturally, trusted and verified sources are to be preferred, especially the evaluation of the trustworthiness can be highly subjective (Li et al. 2007). In contrast, an objective, but at least partially hard to determine, dimension is the *data correctness*. It states, if data values are accurate, comply with the expected values ranges and are also in an error free state (Li et al. 2007). The correctness is also highly dependent on the *data objectivity* (Cai and Zhu 2015). The data being collected should be free of any biases or noise which may create possible outliers or falsifications during analysis. While data might have been once a perfect representation of an entity, time can eventually negatively affect this condition. For this reason, the *data validity* specifies, if the data acquired and stored are valid at the time of processing so that the results derived are meaningful in real time (NIST 2019a). Another sometimes hard to judge aspect is the *data value*. It refers to the gain, in terms of information or profit, which an organization can achieve by analysing the data (Kaisler et al. 2013; Pipino et al. 2002). While there could theoretically be a high potential value hidden in the data, its retrieval might be impeded by a high *data complexity*. The relationship between different data elements has to be at least somewhat comprehensible in order to extract corresponding information (UNECE Big Data Quality Task Team 2014). Another barrier might be the *data interpretability*. The processed data should be understandable to the analyst. It should also not have symbols or units that are ambiguous and not relevant (Cai and Zhu 2015; Pipino et al. 2002). A dimension not actually pertaining the data itself but their use is the *data accessibility*. It refers to the ease of access to the data (Cai and Zhu 2015; Pipino et al. 2002). Moreover, the *data security*, aiming at a state where the data are only accessible to authorized personnel, highly influences the trustworthiness of an organization (Pipino et al. 2002; UNECE Big Data Quality Task Team 2014). This dimension's importance especially emerges, when huge volumes of highly confidential or personal data are gathered, stored and processed.

3.2 Importance of Data Quality in Organizations

To measure the data quality, organizations require data quality parameters, which are typically determined on a case-by-case basis, depending on the case's specific requirements (Pipino et al. 2002). To our knowledge, no universally applied standard data quality measures, being applicable to businesses irrespective of the domain, exist. Yet, it is evident that data quality assessment directly affects the stakeholders trust in the data. Business decisions can be taken with more confidence if stakeholders have higher trust in the quality of data and the results will be better (Hazen et al. 2014; Loshin 2014; Staegemann et al. 2019b). Factors like productivity, business revenue and customer satisfaction are also directly affected by data analysis (Müller et al. 2018). Hence, better data quality would provide a way for valuable, faster and informed business decisions (Lehmann et al. 2016). In (Redman 2004)**,** the author provides several instances of real world disasters caused by using inferior quality data. According to the same publication, decisions taken based on data of inferior quality can cost a typical business a revenue loss of 10 percent or more. Ballou et al. (1998) proposed that using an information modeling approach would make it easier to identify the location of sources causing data quality issues. In the case of big data, enormous effort is required to maintain data quality by modeling the data flow of an organization. However, several researches have claimed that data quality will always remain an important issue (Côrte-Real et al. 2020; NIST 2019b; Redman 2004). Reduced data quality can be attributed to several factors, such as data acquisition, data management, data transfer, or the qualification of the analysts. For this paper, we will delve deep into possible issues caused while data are being acquired and in data management.

4 BIG DATA ACQUISITION

Acquiring big data poses a challenge to the organizations due to its characteristics. There are several technical as well as non-technical hurdles that make it difficult for organizations to invest in big data. Those comprise, inter alia, costly licenses, a lack of motivation from the stakeholders to adopt something new and the data integration with legacy systems (NIST 2019c). When an organization plans to invest in big data, a lot of effort goes into the data acquisition. Hence, it becomes crucial that those data provide value to the organization. For this paper, it is assumed to be true, and the obtained data lead to some gain in terms of information or revenue. To understand the issues that may cause a reduction in big data quality, it is important to understand how big data are acquired.

4.1 Sources of Big Data

An example for a data source whose content's quality is highly unpredictable, are social media. Large amounts of data are being collected from the likes of Facebook, Twitter, Instagram or LinkedIn. With the increase in the number of users as well as platforms, the amount of data collected from social media is going to increase even further in the future (Lee 2017). However, the highly unstructured data collected from social media is often full of noise, bias, errors and redundancy (UNECE Big Data Quality Task Team 2014). Besides those outlets for self-presentation, also everyday internet activities like browsing or posts and chat histories can be utilized for the purpose of analysis. Those data can provide enormous insights while predicting user's behaviour and finding patterns in it (Chen et al. 2013; Chen et al. 2014). It is an essential part of marketing strategies these days. Yet, this data may be filled with bias, it may be incomplete or even incorrect. An example where low data quality can have far worse consequences is the healthcare sector. Data collected in the medical domain range from demographic to morphological data. Illness statistics are being collected every day to develop better algorithms and medicine to cater to critical illnesses. Furthermore, medical data also relate to the insurance policy and the family history of the patient (Bhadani and Jothimani 2016). Users often do not feel comfortable filling up their personal details. While all the fields in this type of data are critical, it might lead to the issue of missing data. In addition to that, there are several other circumstances and occurrences where data quality can get affected, such as devices capturing incorrect data, transposed digits or medical staff missing out recording some data. Besides the medical area, there are also numerous other fields were real-time data sources are encountered and data are being collected from sensors and continuous data streams (Gudivada et al. 2015). Other data are created by audio and videos uploaded by users in real time, information generated by IoT devices or satellite imagery (Marjani et al. 2017; Shelestov et al. 2017). The data generated in real-time is not only vast but heterogeneous as well (Chen et al. 2014). Another common source are organization specific data. The amount of data generated by businesses is continuously increasing (Chen et al. 2014), since data are captured in all the phases of the product or service life cycle, such as production, manufacturing, marketing and sales (Zhang et al. 2017). Though most of the organizational data are inclined to be structured, a lot of unstructured and semi-structured data are also emerging, like emails, documents, chats and logs (Zakir et al. 2015). Usually, after their creation, data are acquired, processed and then stored (Curry et al. 2016, p. 18). Since the acquired data are often highly varying, the likelihood of data corruption during the data acquisition from those sources is quite high. As the intention of the data acquisition is to derive some value for the organization, any analysis done on substandard instead of high quality data might reduce or even nullify the value to the organization, in the worst case even leading to completely wrong results. Consequently, data processing becomes critical and might happen more than once. First, it may happen before storing the data efficiently. Second, processing based on the requirement may be done before analysing the data (UNECE Big Data Quality Task Team 2014). However, if reasons for reduced data quality are understood, we may be able to target those specifically and control data quality issues at the origin. This will also reduce the amount of time spent on data preparation and cleaning.

4.2 Reasons for Reduction in Data Quality during Data Acquisition

There are multiple aspects that facilitate a possible reduction in data quality or usability of said data during the phase of data acquisition (Anagnostopoulos et al. 2016). The connection between those factors and the data quality dimensions is depicted in Figure 2. One of those potential issues is the insertion of *noise*. Noise refers to a modification of correct values in a matter that the resulting values are no longer an actual representation of the corresponding object (García - Gil et al. 2019). A simple example could be quality reduction of a video while uploading or during data transfer. Another case could be data attached with incorrect labels. As a solution for some use cases, collecting more data when the data is noisy would yield better analysis. However, this does not apply to every situation and can therefore not be utilized as a panacea (Huberty 2015). In terms of data quality, noisy data might affect data veracity and correctness, since noisy data do not constitute an accurate representation of the object, as noise may introduce bias or outliers. This may affect data interpretability and data value. A typical cause for quality reduction can also be found when *data is manually inserted* (Li et al. 2007). Even when only one person enters the data in the system, inconsistencies can arise, let alone, when multiple users are involved, which might for example fill in the same information in different formats (Lehmann et al. 2016). While it is possible to train one's own *data entry staff* well before the task begins, reducing the number of errors made, a certain rate of errors is still inevitable. Nevertheless, a lot of data is also created by a service's users. Data *entries created by clients* usually happen online where clients enter information on a web page or on a mobile interface. Most of these interfaces have built-in mechanisms to validate the data entered by clients (Lehmann et al. 2016). Still, several issues can occur while a user is entering the data. For example, many users are not comfortable with or too lazy to share their complete details while

		Data Completeness	Data Veracity	Data Consistency	Data Redundancy	Data Reliability	Data Correctness	Data Objectivity	Data Validity	Data Value	Data Complexity	Data Interpretability	Data Accessibility	Data Security
1	Data Noise		X				X	X		X		X		
2	Data Filled by Staff			X			X			X				
3	Data Filled by Users	X	X	X		X	X	X	X	X		X		
4	Inappropriate Data Structures	X	X			X	X		X	X	X	X		
5	Missing Meta Data			X	X	X			X	X		X		
6	Incorporation of External Data			X				X	X	X				
7	Duplicate Data			X	X				X	X				
8	Data Silos	X		X	X					X			X	X
9	System Errors	X	X	X	X	X	X	X	X	X	X	X	X	X
10	Too Much Data									X				

Figure 2: Big data quality affected by data acquisition issues.

filling up a form. Hence, they end up not filling up all the required fields (Tan et al. 2014, p. 37). One can argue that careful implementation of validations on each field can prevent this issue. However, data quality can get affected if validations are missed out. If the data were critical, it would require imputation of the missing fields. This would affect data veracity, correctness, and reliability of the data. Another issue, stemming from the same cause, are so-called dirty data (Lehmann et al. 2016). This means, that instead of the required values, entirely wrong, incomplete or otherwise distorted data are entered. As a countermeasure, there may be validations in place to prevent user from entering dirty data. Nonetheless, instead of entering correct values, user may enter junk values that pass the validation criteria. This may give a false sense of data completeness. Furthermore, there may be cases, where some attributes are not applicable to the client. In such scenarios, either they will end up filling incorrect data or not fill the data entirely. While these concerns are usually taken care of by enabling applicable fields only when certain fields are filled in the form, this might still create issues during the analysis as all the fields are stored in the database for all the users. The above issues might lead to incomplete, inaccurate, non-reliable, incorrect, non-interpretable, and invalid data. As a result, all of these will affect the data objectivity and data value. Related to this occurrence is also the existence of *irrelevant or inappropriate data structures*, which do not reflect the corresponding object. During data acquisition, this may lead to inappropriate data representation (Chen et al. 2014). It implies that data is not complete, correct or valid, also meaning that the veracity of data cannot be trusted. This might also increase data complexity and decrease its interpretability and reliability. All of these will affect the value that data might have provided. Hence, in big data scenarios, it is extremely important to define data structures in terms of both types of data, those that an object accepts and the relationships or connections it possesses. One possibility for enhancing the meaningfulness of data is its storing with meta-data associated with it (Gudivada et al. 2015). When *meta-data are not present*, it can, depending on the use case, become difficult to understand and manipulate the data. This directly affects the value of the data since the significance of the data cannot be verified (Loshin 2014). For example, meta-data such as date, time and source of information could be required to understand the data for better analysis. Omitting that information might affect data interpretability, reliability and validity (Agrawal et al. 2012). Merging similar data from different sources would also be difficult with missing meta-data (Becker 2016, p. 158). This in turn might affect data consistency and increase redundancy. Another challenge arises when *acquiring data from third parties*. When data is being acquired from external sources, it is often collected without any usability context or purpose. Thus, it poses a big challenge in terms of consistency of the data when merged with the organization’s data for the analysis. Furthermore, data validity also becomes questionable and hence it directly affects data objectivity and value it creates for the business (Gudivada et al. 2015). An additional issue, often exacerbated by the plurality of sources, is *duplicate data*. There could be multiple places from where those could be acquired (Gudivada et al. 2017). For example, one person may have two different email ids and both email ids are present in the system (Tan et al. 2014, p. 34). While, in some cases, this may prove to be valuable to the businesses if both the accounts are analysed together. In other cases, it might not provide any value if one (or both) of the email ids are not correct. Another example could be if data acquired from different sources have the same content, it may not provide any additional value to the business. Data duplication leads to redundancy, reduced data consistency and data validity issues. One common cause for duplicate data also lies in the existence of *data silos*. Those emerge, when instead of acquiring all the data about an object together, they are acquired separately by different departments of an organization. For example, data in the health sector is usually siloed. Hence, analysis on those data is made difficult by diverse interfaces and standards of different departments (Lyko et al. 2016, 52f.). As a result, deriving meaning out of all the related data becomes challenging (Tekiner and Keane 2013). While this issue can be solved with the assignment of proper meta-data, data silos could still be an issue as different departments may not want to share the data due to legal reasons or because of the related expenditure. Thus, information silos are a critical issue that cause hindrances in creating value out of data. It can therefore be argued that data is incomplete when it is siloed. Data may also be inconsistent or redundant as the same information might be getting stored in different silos. Data silos also pose questions in terms of ease of access and security. Besides those potential challenges, there is further always the risk of *system errors*, since sources acquiring data might be faulty and record incorrect data. For example, a faulty sensor will record incorrect values. Therefore, the values might look like noise or outliers in the data. In cases such as these, it is difficult to identify if the unusual values of the object are due to errors at the source or values that need attention (Agrawal et al. 2012). A corrupt data source might affect all the dimensions of data quality. In addition, another issue that might be counterintuitive for many, considering the premise of big data, is the possibility of having *too much data*. Calude et al. (2017) revealed that an interesting correlation derived from large amounts of data might just be there by mere coincidence. Any analysis done on massive amount of data might lead to a wrong conclusion, if correlations are there due to coincidence and any business decisions taken based on those conclusions might not lead to the desired outcome. As a result, the data acquired will not prove to be as valuable as speculated. To address this issue, the authors suggested that any analysis done on big data should be a comprehensive scientific process rather than a simple analysis.

5 BIG DATA MANAGEMENT

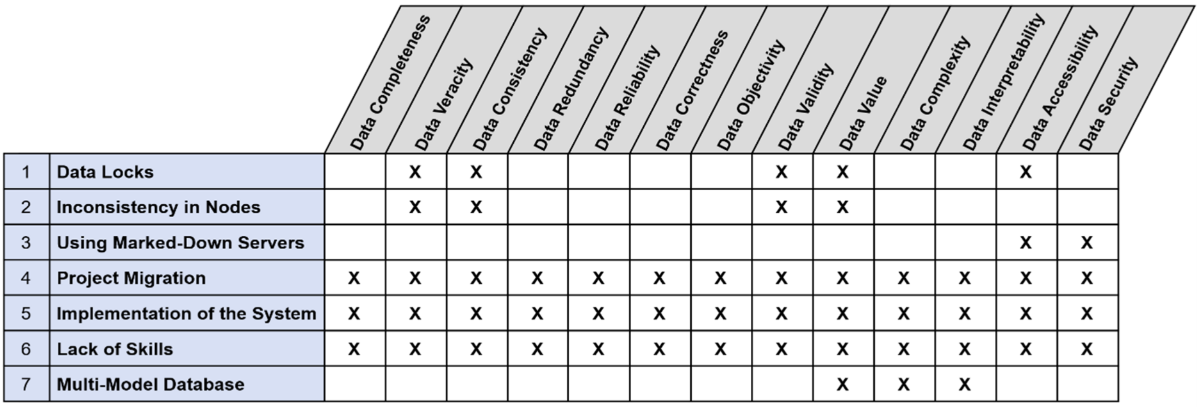
With the increase in velocity, variety and volume of data acquired, efficient data storage mechanisms are required that also support data cleaning and data transformation (Gudivada et al. 2017). Furthermore, it has to be assured that the used storage is reliable and consistent, which is a challenging task, considering the complexity of the data's characteristics (Chen et al. 2014; NIST 2019c).

5.1 Efficient Ways to Manage Big Data

Relational Database Management Systems (RDBMS) are a proven solution for managing structured and relational data. However, big data requires not only reliable but also fast databases (Bhadani and Jothimani 2016; Strohbach et al. 2016, p. 123). Moreover, there is a requirement to manage non-structured data and non-relational schemes. SQL (Sequential Query Language) and RDBMS cannot be relied upon with these new requirements as well as the characteristics of big data (Chen et al. 2014; Ebner et al. 2014). To cater to those extended requirements, several solutions such as MapReduce, cloud computing, distributed file systems or NoSQL are being employed as they are also efficient in managing heterogeneous data (Chen et al. 2014; Ebner et al. 2014; NIST 2019a) and varying types of databases are utilized. Key-Value database like Amazon's *DynamoDB* store data along with a key, *Cassandra* is a distributed storage system that provides fault tolerance as well as linear scalability and document based databases like *MongoDB* store documents in the JSON format. To cater to the increasing volume and the demand for a better performance, there is the necessity to scale data nodes, allowing for an efficient data management. Hence, systems like shared-disk and distributed file systems have come up (NIST 2019a). There are various cloud-based solutions that work on distributed nodes and manage big data effectively. However, the implementation of only one of these technologies is not sufficient to manage big data and commonly a combination of technologies is used to manage the variety of data that is being acquired (Ward and Barker 2013). Big data management solutions need to cater to the massive amount of data and the high velocity, variety and variability. Those characteristics are required to be addressed to maintain the data quality (Curry 2016, p. 30). Subsequently, there are several issues that can inversely affect data quality during the data management.

5.2 Data Management Related Reasons for Reduction in Data Quality

Big Data management requires transformation, cleaning and efficient storage of the acquired data. However, during this process, there are numerous potential issues (Staegemann et al. 2019c), which can contribute to a data quality degradation. This can for example occur because of so-called *data locks*. When the same data object is accessed from multiple user interfaces simultaneously, there are chances of contention. In scenarios when one data object is required to be updated from multiple channels, it needs to be locked by one of the interfaces in order to be updated. This is done to maintain the consistency of the data object. When a data object is locked, it cannot be updated by any other interface. However, in such scenarios, another interface can still read the data object and use it to process its results, potentially leading to inconsistencies. Locking in a large scale database also leads to reduced performance (NIST 2019a). In terms of data quality dimensions, data locks might affect data veracity as the read object is not an accurate representation of the data. Data validity is also affected as the read object might be invalidated if the update is revoked. It will also cause inconsistencies in such cases. Additionally, accessibility is affected as the object in question is not available to both the channels. This may create possible issues with data objectivity and affect data value. Another related threat are *inconsistencies amongst the utilized nodes*. The data are usually managed on distributed nodes consisting of primary and secondary nodes. Generally, consistency is ensured using proper locking mechanisms on the data object, as mentioned above. However, consistency issues arise when in a distributed data system or file system, an operation is being executed on a primary node as well as on a secondary node, which has not yet been updated. The results achieved might prove to be incorrect (NIST 2019a). It may lead to inaccurate and inconsistent data and inversely affect the value of data. Furthermore, also the choice of hardware can have an impact, when *marked-down servers* are used. Storing big data is an expensive process and some organizations use low quality commodity hardware and marked down servers to store their data. However, it will affect the overall performance of the system, especially in cases where faster access to data is required (NIST 2019a; Staegemann et al. 2019c). Low quality servers not just perform poorly but might even pose a threat to data security and accessibility. Sometimes it might also become necessary to *migrate* an up to now functioning project, which is a critical phase for the organization. It is required that proper standards and processes are in place so that data can be migrated successfully. Since there are several components involved in the migration process, there could be a possibility where the components are not compatible with each other (Lehmann et al. 2016; NIST 2019c). Incompatible components might prove critical to the business. To cater to this issue, it is important that a compatibility analysis is done while planning the migration. In the worst-case scenario, all the dimensions of data quality may be affected by project migration. Yet, not only the migration, but also the *implementation of big data systems* is a challenging endeavour. There are two ways to implement big data technologies. One is to replace the entire legacy system with an updated big data solution. The second is to slowly incorporate big data alternatives to the legacy system already in place (NIST 2019c). This corresponds to the well-known brownfield and greenfield consideration, which is prominent in the investment domain (Meyer and Estrin 2001; Qiu and Wang 2011). Just like project migration, the replacement of legacy systems or the integration into the legacy system requires a lot of planning and analysis. However, the same holds true for the creation of completely new systems. In each case, incorrect implementations may affect all the dimensions of data quality. For this reason, their thorough testing as an important part of the big data engineering (Volk et al. 2019) procedure is, despite its high complexity (Staegemann et al. 2019a), extremely important. This whole process of creating functioning and beneficial big data solutions puts high demands on the involved personnel's capabilities. Yet, this not only applies to the technical implementation but also the configuration and application of such systems, whose performance can be greatly reduced by a *lack of skills*. Big data management tools, such as Hadoop or Cassandra, require proficiently skilled people to manage data, since certain database management designs are more capable than others. To design an efficient system, data should be clearly understood by the responsible staff managing it (Agrawal et al. 2012). If the data is not thoroughly understood, erroneous assumptions can be made which might lead to unreliable system designs and implementations. Hence, it is important that big data management is taken care of by highly qualified experts. In terms of quality, a lack of skills can potentially directly affect all the dimensions. Furthermore, depending on the use case, the data's variety necessitates the *integration of multi-model databases*. Since big data comprises a collection of heterogeneous data and data sources, chances are that the data are managed in different databases conforming to different data models. To provide the users efficient access, these heterogeneous databases are integrated based on their respective data models



		Data Completeness	Data Veracity	Data Consistency	Data Redundancy	Data Reliability	Data Correctness	Data Objectivity	Data Validity	Data Value	Data Complexity	Data Interpretability	Data Accessibility	Data Security
1	Data Locks		X	X					X	X			X	
2	Inconsistency in Nodes		X	X					X	X				
3	Using Marked-Down Servers												X	X
4	Project Migration	X	X	X	X	X	X	X	X	X	X	X	X	X
5	Implementation of the System	X	X	X	X	X	X	X	X	X	X	X	X	X
6	Lack of Skills	X	X	X	X	X	X	X	X	X	X	X	X	X
7	Multi-Model Database									X	X	X		

Figure 3: Big data quality affected by data management issues.

and meta-data models are used to extract information out of them (Gudivada et al. 2015). However, when an accurate meta-data model is not available, the queries on the databases would not yield appropriate results. This can lead analysts to question the data sources. It might also substantially increase data complexity and reduce data interpretability, as the meta-data are not known. The creation of appropriate business value often requires the joint analysis of multi-model data. However, issues with heterogeneous data and the employment of multi model databases can be reduced by removing the need for integration, for example by implementing big data technologies that can deal with different types of databases. The connection of those identified challenges to the presented data quality dimensions is depicted in Figure 3.

5.3 Observations

It is evident that organizations are keen to invest in big data technologies, as there is a clear indication of retrieving business value out of it (Chen et al. 2013). In a survey conducted by Lehmann et al. (2016), big data cleaning tools and efficient big data management tools rank high in the most needed services by the questioned organizations. Even though, most of the organizations employ data cleaning mechanisms, nearly all of the tasks cannot be automated and must be done under human supervision. This implies that a lot of time is spent on data cleaning and pre-processing. Organizations need to critically analyse the implementation of big data infrastructure especially in terms of how data is being acquired and managed. Most of the reasons for reduced data quality discussed above will affect data value. Hence, if organizations evaluate the reasons for reduced data quality during data acquisition and data management, appropriate measures to improve data quality can be taken, subsequently saving the time spent on data pre-processing.

6 CONCLUSION

The conducted research identifies the dimensions that affect data quality as well as the specific reasons that cause reduced data quality during data acquisition and data management. Furthermore, matrices have been developed, relating the above-identified challenges with the data quality dimensions, providing scientists and practitioners a comprehensible overview of the matter. It can be concluded that all the issues that cause inferior data quality impact data value which can affect business decisions and revenue. In addition, data quality can be optimized if the reasons affecting it are known, which is facilitated by the publication at hand. In the future, this work can be extended by not only incorporating findings in the scientific literature, but also the experiences of practitioners, expanding the scope and therefore potentially also further increasing the significance. This could be done by including surveys and interviews of business experts based on the reasons identified in the paper.

REFERENCES

Abdallah, M., Muhairat, M., Althunibat, A., and Abdalla, A. 2020. "Big Data Quality: Factors, Frameworks, and Challenges," *Compusoft: An International Journal of Advanced Computer Technology* (9:8), pp. 3785-3790.

Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M., Gehrke, J., Haas, L., Halevy, A., Han, J., Jagadish, H. V., Labrinidis, A., Madden, S., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., Ross, K., Shahabi, C., Suciu, D., Vaithyanathan, S., and Widom, J. 2012. "Challenges and Opportunities with Big Data: A community white paper developed by leading researchers across the United States,"

Al-Sai, Z. A., Abdullah, R., and Husin, M. H. 2020. "Critical Success Factors for Big Data: A Systematic Literature Review," *IEEE Access* (8), pp. 118940-118956 (doi: 10.1109/ACCESS.2020.3005461).

Anagnostopoulos, I., Zeadally, S., and Exposito, E. 2016. "Handling big data: research challenges and future directions," *The Journal of Supercomputing* (72:4), pp. 1494-1516 (doi: 10.1007/s11227-016-1677-z).

Ballou, D., Wang, R., Pazer, H., and Tayi, G. K. 1998. "Modeling Information Manufacturing Systems to Determine Information Product Quality," *Management Science* (44:4), pp. 462-484 (doi: 10.1287/mnsc.44.4.462).

Becker, T. 2016. "Big Data Usage," in *New Horizons for a Data-Driven Economy*, J. M. Cavanillas, E. Curry and W. Wahlster (eds.), Cham: Springer International Publishing, pp. 143-165.

Bhadani, A. K., and Jothimani, D. 2016. "Big Data: Challenges, Opportunities and Realities," in *Effective Big Data Management and Opportunities for Implementation*, D. Taniar, M. K. Singh and D. K. G. (eds.), IGI Global, pp. 1-24.

Bughin, J. 2016. "Big data, Big bang?" *Journal of Big Data* (3:1) (doi: 10.1186/s40537-015-0014-3).

Cai, L., and Zhu, Y. 2015. "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era," *Data Science Journal* (14:2), pp. 1-10 (doi: 10.5334/dsj-2015-002).

Calude, C. S., and Longo, G. 2017. "The Deluge of Spurious Correlations in Big Data," *Foundations of Science* (22:3), pp. 595-612 (doi: 10.1007/s10699-016-9489-4).

Chen, J., Chen, Y., Du, X., Li, C., Lu, J., Zhao, S., and Zhou, X. 2013. "Big data challenge: a data management perspective," *Frontiers of Computer Science* (7:2), pp. 157-164 (doi: 10.1007/s11704-013-3903-7).

Chen, M., Mao, S., and Liu, Y. 2014. "Big Data: A Survey," *Mobile Networks and Applications* (19:2), pp. 171-209 (doi: 10.1007/s11036-013-0489-0).

Côrte-Real, N., Ruivo, P., and Oliveira, T. 2020. "Leveraging internet of things and big data analytics initiatives in European and American firms: Is data quality a way to extract business value?" *Information & Management* (57:1), p. 103141 (doi: 10.1016/j.im.2019.01.003).

Couldry, N., and Mejias, U. A. 2019. "Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject," *Television & New Media* (20:4), pp. 336-349 (doi: 10.1177/1527476418796632).

Curry, E. 2016. "The Big Data Value Chain: Definitions, Concepts, and Theoretical Approaches," in *New Horizons for a Data-Driven Economy*, J. M. Cavanillas, E. Curry and W. Wahlster (eds.), Cham: Springer International Publishing, pp. 29-37.

Curry, E., Becker, T., Munné, R., Lama, N. de, and Zillner, S. 2016. "The BIG Project," in *New Horizons for a Data-Driven Economy*, J. M. Cavanillas, E. Curry and W. Wahlster (eds.), Cham: Springer International Publishing, pp. 13-26.

Demchenko, Y., Grosso, P., Laat, C. de, and Membrey, P. 2013. "Addressing big data issues in Scientific Data Infrastructure," in *2013 International Conference on Collaboration Technologies and Systems (CTS),* San Diego, CA, USA. 20.05.2013 - 24.05.2013, IEEE, pp. 48-55.

Ebner, K., Buhnen, T., and Urbach, N. 2014. "Think Big with Big Data: Identifying Suitable Big Data Strategies in Corporate Environments," in *Proceedings of the 2014 HICSS,* Hawaii. 06.01.2014-09.01.2014, pp. 3748-3757.

Ezzine, I., and Benhlima, L. 2018. "A Study of Handling Missing Data Methods for Big Data," in *2018 IEEE 5th International Congress on Information Science and Technology (CiSt),* Marrakech. 21.10.2018 - 27.10.2018, IEEE, pp. 498-501.

García, S., Luengo, J., and Herrera, F. 2015. *Data Preprocessing in Data Mining*, Cham: Springer International Publishing.

García - Gil, D., Luque - Sánchez, F., Luengo, J., García, S., and Herrera, F. 2019. "From Big to Smart Data: Iterative ensemble filter for noise filtering in Big Data classification," *International Journal of Intelligent Systems* (34:12), pp. 3260-3274 (doi: 10.1002/int.22193).

Géczy, P. 2014. "Big Data Characteristics," *The Macrotheme Review* (3:6), pp. 94-104.

Gudivada, V. N., Apon, A., and Ding, J. 2017. "Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations," *International Journal on Advances in Software* (10:1).

Gudivada, V. N., Jothilakshmi, S., and Rao, D. 2015. "Data Management Issues in Big Data Applications," in *Proceedings of the ALLDATA 2015,* Barcelona, Spain. 19.04.2015 - 24.04.2015, pp. 16-21.

Günther, W. A., Rezazade Mehrizi, M. H., Huysman, M., and Feldberg, F. 2017. "Debating big data: A literature review on realizing value from big data," *The Journal of Strategic Information Systems* (26:3), pp. 191-209 (doi: 10.1016/j.jsis.2017.07.003).

Häusler, R., Staegemann, D., Volk, M., Bosse, S., Bekel, C., and Turowski, K. 2020. "Generating Content-Compliant Training Data in Big Data Education," in *Proceedings of the 12th International Conference on Computer Supported Education,* Prague, Czech Republic. 02.05.2020 - 04.05.2020, SCITEPRESS - Science and Technology Publications, pp. 104-110.

Hazen, B., Boone, C., Ezell, J., and Jones-Farmer, L. A. 2014. "Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications," *International Journal of Production Economics* (154), pp. 72-80 (doi: 10.1016/j.ijpe.2014.04.018).

Huberty, M. 2015. "Awaiting the Second Big Data Revolution: From Digital Noise to Value Creation," *Journal of Industry, Competition and Trade* (15:1), pp. 35-47 (doi: 10.1007/s10842-014-0190-4).

Kaisler, S., Armour, F., Espinosa, J. A., and Money, W. 2013. "Big Data: Issues and Challenges Moving Forward," in *Proceedings og the 2013 HICSS,* Wailea, HI, USA. 07.01.2013 - 10.01.2013, IEEE, pp. 995-1004.

Katal, A., Wazid, M., and Goudar, R. H. 2013. "Big data: Issues, challenges, tools and Good practices," in *Parashar (Hg.) – 2013 sixth International Conference*, pp. 404-409.

Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Ali, W. K. M., Alam, M., Shiraz, M., and Gani, A. 2014. "Big data: survey, technologies, opportunities, and challenges," *The Scientific World Journal* (2014), pp. 1-18 (doi: 10.1155/2014/712826).

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., and Kruschwitz, N. 2011. "Big Data, Analytics and the Path From Insights to Value," *MITSloan Management Review* (52:2), pp. 21-31.

Lee, I. 2017. "Big data: Dimensions, evolution, impacts, and challenges," *Business Horizons* (60:3), pp. 293-303 (doi: 10.1016/j.bushor.2017.01.004).

Lehmann, C., Roy, K., and Winter, B. 2016. "The State of Enterprise Data Quality: 2016: Perception, Reality and the Future of DQM," 451 Research.

Li, X., Shi, Y., Li, J., and Zhang, P. 2007. "Data Mining Consulting Improve Data Quality," *Data Science Journal* (6:17), S658-S666 (doi: 10.2481/dsj.6.S658).

Loshin, D. 2014. "Understanding Big Data Quality for Maximum Information Usability," SAS Institute.

Lyko, K., Nitzschke, M., and Ngonga Ngomo, A.-C. 2016. "Big Data Acquisition," in *New Horizons for a Data-Driven Economy*, J. M. Cavanillas, E. Curry and W. Wahlster (eds.), Cham: Springer International Publishing, pp. 39-61.

Marjani, M., Nasaruddin, F., Gani, A., Karim, A., Hashem, I. A. T., Siddiqa, A., and Yaqoob, I. 2017. "Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges," *IEEE Access* (5), pp. 5247-5261 (doi: 10.1109/ACCESS.2017.2689040).

Mauro, A. de, Greco, M., and Grimaldi, M. 2015. "What is big data? A consensual definition and a review of key research topics," in *Proceedings of the IC-ININFO 2014,* Madrid, Spain. 05.09.2014-08.09.2014, pp. 97-104.

Merino, J., Caballero, I., Rivas, B., Serrano, M., and Piattini, M. 2016. "A Data Quality in Use model for Big Data," *Future Generation Computer Systems* (63), pp. 123-130 (doi: 10.1016/j.future.2015.11.024).

Meyer, K. E., and Estrin, S. 2001. "Brownfield Entry in Emerging Markets," *Journal of International Business Studies* (32:3), pp. 575-584 (doi: 10.1057/palgrave.jibs.8490985).

Müller, O., Fay, M., and Vom Brocke, J. 2018. "The Effect of Big Data and Analytics on Firm Performance: An Econometric Analysis Considering Industry Characteristics," *Journal of management information systems* (35:2), pp. 488-509 (doi: 10.1080/07421222.2018.1451955).

Narasimhan, R., and Bhuvaneshwar, T. 2014. "Big Data – A Brief Study," *International Journal of Scientific & Engineering Research* (5:9).

NIST 2019a. "NIST Big Data Interoperability Framework: Volume 1, Definitions, Version 3," Gaithersburg, MD: National Institute of Standards and Technology.

NIST 2019b. "NIST Big Data Interoperability Framework: Volume 3, Use Cases and General Requirements, Version 3," Gaithersburg, MD: National Institute of Standards and Technology.

NIST 2019c. "NIST Big Data Interoperability Framework: Volume 9, Adoption and Modernization, Version 3," Gaithersburg, MD: National Institute of Standards and Technology.

Pentzold, C., Brantner, C., and Fölsche, L. 2019. "Imagining big data: Illustrations of "big data" in US news articles, 2010–2016," *New Media & Society* (21:1), pp. 139-167 (doi: 10.1177/1461444818791326).

Pipino, L. L., Lee, Y. W., and Wang, R. Y. 2002. "Data quality assessment," *Communications of the ACM* (45:4), p. 211 (doi: 10.1145/505248.506010).

Press, G. 2014. *12 Big Data Definitions: What's Yours?* https://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours. Accessed 29 November 2020.

Qiu, L. D., and Wang, S. 2011. "FDI Policy, Greenfield Investment and Cross-border Mergers," *Review of International Economics* (19:5), pp. 836-851 (doi: 10.1111/j.1467-9396.2011.00984.x).

Redman, T. C. 2004. "Data: An Unfolding Quality Disaster," *DM Review*.

Shelestov, A., Lavreniuk, M., Kussul, N., Novikov, A., and Skakun, S. 2017. "Exploring Google Earth Engine Platform for Big Data Processing: Classification of Multi-Temporal Satellite Imagery for Crop Mapping," *Frontiers in Earth Science* (5) (doi: 10.3389/feart.2017.00017).

Smith, S. M., and Nichols, T. E. 2018. "Statistical Challenges in "Big Data" Human Neuroimaging," *Neuron* (97:2), pp. 263-268 (doi: 10.1016/j.neuron.2017.12.018).

Staegemann, D., Hintsch, J., and Turowski, K. 2019a. "Testing in Big Data: An Architecture Pattern for a Development Environment for Innovative, Integrated and Robust Applications," in *Proceedings of the WI2019*, pp. 279-284.

Staegemann, D., Volk, M., Daase, C., and Turowski, K. 2020a. "Discussing Relations Between Dynamic Business Environments and Big Data Analytics," *Complex Systems Informatics and Modeling Quarterly* (23), pp. 58-82 (doi: 10.7250/csimq.2020-23.05).

Staegemann, D., Volk, M., Jamous, N., and Turowski, K. 2019b. "Understanding Issues in Big Data Applications - A Multidimensional Endeavor," in *Twenty-fifth Americas Conference on Information Systems,* Cancun, Mexico.

Staegemann, D., Volk, M., Jamous, N., and Turowski, K. 2020b. "Exploring the Applicability of Test Driven Development in the Big Data Domain," in *Proceedings of the ACIS 2020,* Wellington, New Zealand. 01.12.2020 - 04.12.2020.

Staegemann, D., Volk, M., Nahhas, A., Abdallah, M., and Turowski, K. 2019c. "Exploring the Specificities and Challenges of Testing Big Data Systems," in *Proceedings of the 15th International Conference on Signal Image Technology & Internet based Systems,* Sorrento, Italy.

Strohbach, M., Daubert, J., Ravkin, H., and Lischka, M. 2016. "Big Data Storage," in *New Horizons for a Data-Driven Economy*, J. M. Cavanillas, E. Curry and W. Wahlster (eds.), Cham: Springer International Publishing, pp. 119-141.

Tan, P.-N., Steinbach, M., and Kumar, V. 2014. *Introduction to data mining*, Harlow: Pearson.

Tekiner, F., and Keane, J. A. 2013. "Big Data Framework," in *Proceedings of the SMC 2013,* Manchester. 13.10.2013 - 16.10.2013, IEEE, pp. 1494-1499.

UNECE Big Data Quality Task Team 2014. "A Suggested Framework for the Quality of Big Data,"

Volk, M., Staegemann, D., Pohl, M., and Turowski, K. 2019. "Challenging Big Data Engineering: Positioning of Current and Future Development," in *Proceedings of the IoTBDS 2019*, pp. 351-358.

Volk, M., Staegemann, D., Trifonova, I., Bosse, S., and Turowski, K. 2020a. "Identifying Similarities of Big Data Projects–A Use Case Driven Approach," *IEEE Access* (8), pp. 186599-186619 (doi: 10.1109/ACCESS.2020.3028127).

Volk, M., Staegemann, D., and Turowski, K. 2020b. "Big Data," in *Handbuch Digitale Wirtschaft*, T. Kollmann (ed.), Wiesbaden: Springer Fachmedien Wiesbaden, pp. 1-18.

Ward, J. S., and Barker, A. 2013. *Undefined By Data: A Survey of Big Data Definitions*. https://arxiv.org/pdf/1309.5821.pdf. Accessed 16 January 2020.

Zakir, J., Seymour, T., and Berg, K. 2015. "Big Data

Analytics," *Issues in Information Systems* (16:2), pp. 81-90.

Zhang, Y., Ren, S., Liu, Y., Sakao, T., and Huisingh, D. 2017. "A framework for Big Data driven product lifecycle management," *Journal of Cleaner Production* (159), pp. 229-240 (doi: 10.1016/j.jclepro.2017.04.172).