IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 14, 2020, accepted September 24, 2020, date of publication October 1, 2020, date of current version October 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3028127

Identifying Similarities of Big Data Projects–A Use Case Driven Approach

MATTHIAS VOLK, (Graduate Student Member, IEEE), DANIEL STAEGEMANN, IVAYLA TRIFONOVA, SASCHA BOSSE, AND KLAUS TUROWSKI

Faculty of Computer Science, Otto von Guericke University Magdeburg, 39106 Magdeburg, Germany

Corresponding author: Matthias Volk (matthias.volk@ovgu.de)

ABSTRACT Big data is considered as one of the most promising technological advancements in the last decades. Today it is used for a multitude of data intensive projects in various domains and also serves as the technical foundation for other recent trends in the computer science domain. However, the complexity of its implementation and utilization renders its adoption a sophisticated endeavor. For this reason, it is not surprising that potential users are often overwhelmed and tend to rely on existing guidelines and best practices to successfully realize and monitor their projects. A valuable source of knowledge are use case descriptions, of which a multitude exists, each of them with a varying information density. In this design science research endeavor, 43 use cases are identified by conducting a thorough literature review in combination with the application and adaption of a corresponding template for big data projects. By a subsequent categorization, which is performed by identifying and employing a hierarchical clustering algorithm, nine different standard use cases emerge, as the contribution's artifact. This provides decision-makers with an initial entry point, which can be utilized to shape their project ideas, not only by identifying the general meaningfulness of their potential big data project but also in terms of concrete implementation details.

INDEX TERMS Big data, use case analysis, clustering, categorization, literature review, design science research.

I. INTRODUCTION

Due to the ever-growing amount of data produced and captured by humanity [1], [2], the ability to analyze and subsequently use the contained information has gained a widely acknowledged significance in today's society [3]. While the usage of data, in general, is no new concept, the prevailing ''data deluge'' poses new challenges that overstrain traditional technologies and demand new solutions [4], leading to the term ''big data''. Even though there is no unified definition of the term, the approach by the National Institute of Standards and Technology (NIST) belongs to the most common ones. It states that ''Big Data consists of extensive datasets – primarily in the characteristics of volume, velocity, variety, and/or variability – that require a scalable architecture for efficient storage, manipulation, and analysis'' [5]. These data characteristics describe the nature of the data to be processed and, thus, are often recognized as crucial influence factors when it comes to the planning and realization of big data projects. The volume, for instance, describes the amount of data to be stored, managed, and processed [5]. In turn, the variety of the data either represents the heterogeneity of the structure [6], [7], but sometimes also the origin [5]. The same differentiated view exists for the velocity that either addresses the speed with which the data is incoming or the time for its processing [5], [7]. The variability refers to changes of the dataset, for instance, in terms of the structure, rate of the data flow or the size of the data [5]. Together, those characteristics are covered under the umbrella term of the fours *Vs* of big data, with each depicting the abbreviation of one characteristic. Besides those known *core* characteristics, many others emerged in recent years, such as the value of the data or the veracity [5], [8]. To handle the additional challenges, posed by those *Vs*, compared to traditional workloads, numerous new strategies, tools, and systems have been introduced in recent years. Nowadays, those technologies and concepts are applied to a wide array of domains, comprising, but not being limited to, mobility [9], smart cities [10], media distribution [11], healthcare [12], sports [13], education [14] and business [15].

The associate editor coordinating the review of this manuscript and approving it for publication was Fanbiao Li.

Furthermore, also the positive effects of the utilization of big data technologies have been shown [16], [17], emphasizing the importance of such efforts.

However, implementing and operating those systems is a sophisticated endeavor with many possible pitfalls [18], [19], which could diminish the benefits or even result in negative outcomes [20]. By considering the shortage of qualified experts in this domain and the concurrent demand [21], [22], independent from the actual size of the enterprise, it appears to be reasonable to support concerned decision-makers and technicians. This applies especially to fundamental tasks like the design of the underlying technical architecture [23]. Hence, a thorough description of corresponding use cases could facilitate the realization of those kinds of complex projects by providing a suitable source of information. However, the biggest deficiency of most of these is the level of detail. Although every year, numerous contributions are published, describing the general endeavor, they often omit information regarding the actual project, the occurred problems, and the specific implementation. Furthermore, in some cases, similar methods and paradigms are applied or even technological considerations made, leading sometimes to a sole distinction in terms of the main scope. This reduces transparency for those who want to utilize such cases to obtain information for similar application scenarios, in terms of the general meaningfulness, best practices, concrete technologies, or even specific architecture details. Consequently, today a multitude of case studies exists, potentially offering valuable guidance for the realization of big data projects, but their exploitation still remains a challenge. To uncover inherent relations between those and thus overcome the outlined barriers, the following research question (RQ) shall be answered in the course of this work:

Which standard big data use cases are revealed, applying cluster analysis to the corresponding case studies, found in the literature?

Answering this question and structuring a collection of successful big data projects, accordingly, could constitute a valuable resource for the instantiation of future projects providing a general orientation and guidance but also concrete implementation details for specific scenarios. Due to the presumed similarity of the cases in parts, the formulation of standard use cases for the structuring of a potential collection appears to be a suitable solution. As a result, decision-makers will obtain the opportunity to quickly identify cases, which are similar to the challenges they are facing and thus, lean on the existing knowledge base. It also enables them to classify their projects and identify the needed expertise in a more systematic and therefore more meaningful way. Furthermore, those categories could be used as a foundation for a big data decision support system. This allows for a comparison of the expected usage scenario with familiar exemplary cases and the incorporation of gained experiences in a technology-supported way [24].

Since the first introduction of the most famous big data technology – Hadoop – in 2005 [25] the maturity tremendously increased. To take this development into account, only use cases published between 2015 and 2018 were incorperated, as cases from the period before have already been sufficiently investigated by other authors, such as in [26]. Due to the start of this research in 2019 and the decision to cover only literature of completed years, the end date 2018 was chosen. Furthermore, to reduce the complexity, misinterpretations, and effort for further modifications and extensions, additional document analysis techniques were applied. While the general approach resembles the work of Ylijoki and Porras [26], using case study analysis and clustering, the specific methodology, the analyzed time frame, and the respective objectives differ. Apart from that, the publication at hand especially addresses researchers and scientists concerned with the project feasibility, technology selection as well as implementation and application details in a big data context.

A. METHODOLOGY

To find an answer for the initially formulated RQ, multiple methodologies need to be applied in a combined way. As a general *foundation* for the realization of this endeavor, the design science research methodology according to Hevner *et al.* [27] was used, providing an artifact as a solution to the formulated problem. In particular, similarities of successful big data projects are investigated and standard use cases are derived as the main artifact of this work. To further improve the reproducibility and clarity of the conducted measures, the six-stepped procedure according to Peffers *et al.* [28] was followed to ensure, that the development of the intended solution is systematically approached. The first step of this workflow focuses on the brief motivation and description of the problem. Subsequently, the main objectives are highlighted, which is directly followed by design and development. As a transition, the theoretical foundation needs to be investigated and relevant material collected. Since the main artifact of the research, in the form of standard use cases, builds upon existent use cases, a structured literature review [29], [30], as well as a use case analysis were conducted, examining the content of each of the cases in-depth. Often companies write case studies for advertisement aims, for example, to win new customers or to present themselves in social media. On the other hand, they prepare the case studies as a documentation of their "best practices" [31]. Those describe the decisions made by companies, the reasons for those decisions, their implementation, and the following results [32]. Case studies can also be used as a guideline for subsequent users, having a particular problem, or striving for a concrete solution. Either way, to provide maxiumum value, case studies must be written according to pre-defined standards [33]. Hence, it was presumed that those case studies describe the usage of big data technologies and the related processes in their context. To ensure this, additionally, the comprehensiveness of each of them was checked by using a modified version of an existing use case template [34]. At this stage, also important features for a later clustering

approach, implemented during the subsequent step, are identified. The following three steps concern the demonstration, evaluation, and presentation of the artifact.

B. STRUCTURE

Based on the used methodology, the structure is as follows. Within the first section, an initial overview of the current situation, as well as the derived research question and main objectives, are presented. Along with this, the used methodology as well as the structure is introduced. In the subsequent second section, the conducted structured literature review, resulting in the collection of the regarded use cases, is described. The third section thoroughly describes the performed clustering, each of the clusters and their particularities, as well as the actual development of the standard use cases. An evaluation of the obtained results is presented in the fourth section. In here, the standard use cases are tested on the base of previously unseen data, at which a categorization of those is pursued. Within the fifth section, concluding remarks are given. Apart from a summary, this comprises a discussion and an outlook on future research.

II. THE STRUCTURED LITERATURE REVIEW

For the identification of relevant big data use cases that have been published between 2015 and 2018, a structured literature review is performed. In particular, the methodologies according to Levy and Ellis [30] as well as Webster and Watson [29] were used. To verify the comprehensiveness of the found out contributions, additionally, an existing use case template was adopted, which is provided by the NIST [34]. Within the following section, the review protocol, the used template as well as the results are meticulously described [35].

A. REVIEW PROTOCOL

To obtain a broad overview of the entire domain, the focus of the search was not set on a single database. Instead to ''exhaust all sources that contain IS research publications'' [30, p. 183] the scientific literature database Scopus was used for the initial keyword search. Although it was noted by different authors [29], [30] that multiple sources should be queried, to receive an extensive overview of all of the relevant articles, due to its comprehensiveness, only the mentioned database was used, since most of the widely accepted literature databases and their relevant articles are listed here, referring to the source. For these reasons, Scopus serves more as a kind of a meta-database indexing relevant contributions. Secondly, it was not required to perform alterations on the queried terms and used operators. According to the targeted domain of interest, the terms ''*case study*'', ''*use case*'' and ''*case description*'' were used in combination with ''*big data*''. Further, to reduce the number of irrelevant search results, additional inclusion and exclusion criteria were formulated. Some widely accepted ones are, for instance, the used language, the publication in a conference, journal or book, and relevance for answering the formulated research question. Only when all of the aforementioned inclusion criteria were met, the paper was accepted. However, sometimes it was noted that some of the contributions did not encompass as much information as needed. Due to this, various criteria were formulated, like, if a use case was not presented very well and did not contain the required key information suitably, a paper was rejected. In turn, in a few of the found out use cases, the information density was very high but only focusing on the introduction, development, or evaluation of new technologies. An additional exclusion criterion was formulated for those cases. The complete collection of all of the used inclusion and exclusion criteria is summarized in Table 1. The initial material collection was performed by applying the described keywords and some of the mentioned inclusion criteria directly through the advanced search mechanics of the literature database. As a result, 2,379 non-redundant publications were found. Following that, it was required to check the papers on their actual usability. The refinement of the material was performed in a two-stepped procedure. In the first step, the title, abstract, and structure were checked. This resulted in having 108 relevant contributions. Within the second step, the actual content of the remaining case studies was investigated. It was noticed that in most of the cases the content differed strongly, in terms of the information density.

TABLE 1. Inclusion and exclusion criteria of the structured literature review.

Inclusion Criteria	Exclusion Criteria
The paper focuses on the presentation of a big data use case.	The paper focuses mainly on the introduction, development, or evaluation of new technologies.
The paper must be written in English.	The paper does not provide any information about the data to be used.
The paper was published between 2015 and 2018.	The paper does not provide any information about the previous approach for data processing and analysis.
The paper must be published in either conference proceedings, a book or a journal.	The paper does not highlight the main objective and expectations for the adoption of big data.
	The papers do not describe any requirements for the planned project.
	The papers do not mention the data source.

Only a rough presentation was performed for most of the use cases, neglecting important descriptive information, such as about the data or the situation before. To simplify the mapping of the needed information, the qualitative analysis, and the evaluation of the comprehensiveness for each use case, the use of a corresponding template was deemed appropiate. As a basis, the very extensive template provided by the NIST, covering eight different parts and 57 big data project-related questions, was modified and used [34]. The original template was designed by the NIST Big Data Public Working Group (NBD-PWG) to collect existing use cases. Due to its overarching purpose, the template as well as the

respective categories were continuously validated regarding their applicability and, thus, compared to the actual content of the found out contributions.

B. USE CASE TEMPLATE

The basic template comprises a multitude of different aspects. Apart from the general project description and the situation before, the relevant big data characteristics, applied techniques, and multiple other information are requested. Due to the reason that not all of the template's fields were from major interest, modifications to the original version were performed. After an initial scan, during the second step of the refinement, fields which were not related to the formulated criteria (cf. Table 1) and not required for the general applicability of big data-related projects were removed (R). This includes, for instance, the last two parts and questions like "*do you foresee any potential risks from public or private open data projects?*" or "*under what conditions do you give people access to your data?*" [34]. At the same time, additional points, like the veracity of the data or privacy-related information, were newly added (N). All other fields, meaningful for the found out contributions, were adapted. An overview of the general content of each template category is depicted in Table 2, whereas a complete depiction of all of the made considerations is shown within the appendix in Table 8. While the first column describes the targeted content of the field, the latter is focusing on the performed changes. Either the field of the original use case template was adapted (A), removed (R) or a new one was added (N).

TABLE 2. Overview of the use case template categories and their general description (cf. [34]).

No.	Part	Content
1	Overall project description	This part aims to create an overall project description, covering general information about the use case, relevant stakeholders, current, and future approach.
2	Big data characteristics	This part deals with the (raw) data to be processed and their related characteristics.
3	Big data science	This part handles different questions that are related to the management of the data, in terms of the performed analytics, curation and governance, metadata, and data types.
4	General security and privacy	This part tackles general questions about the security and privacy of the data.
5	Classify use cases with tags	This part intents to classify the use case with predefined tags.
6	Overall big data issues	This part covers issues that cannot be assigned to any of the other parts, like user interface and mobile access issues.
7	Workflow processes	This part deals with projects that face multiple stages, at which different characteristics and attributes may alter. To record and present those changes, in here the information can be provided.
8	Detailed security and privacy	This part contains questions that „*are designed to gather a comprehensive image of security and privacy aspects*".

In total, 23 of the questions were removed, 27 adapted without any changes, 7 modified, and 11 newly added. For instance, within the first part, only minor changes were made, to prevent misleading interpretations during the investigation of the use cases. This includes the renaming of the parts, the deletion of one question, and the addition of the part *advantages of harnessing big data*, which was covered by almost all cases except for one [36]. In the subsequent second part, additional characteristics were introduced or moved from the adhering third part, due to the high coverage by the use case description. In the following, every part was evaluated regarding the usefulness for the intended comprehensiveness check. Consequently, smaller changes were conducted, as they are depicted in the referred Table 8. Noteworthy are the last two parts, which were removed completely, since almost none of the use cases hold information related to those. While part seven deals with various workflow steps and respective changes in the data characteristics, the last one goes into more detail, when it comes to privacy and security concerns. Although some papers, such as [33], [37], [38], discussed various issues in detail, the relatively small number of related cases made those parts not universally applicable. Eventually, the final template was used to check the comprehensiveness of each found out use case.

C. RESULTS OF THE LITERATURE REVIEW

After the two-stepped refinement procedure and the comprehensiveness check through the used template were finished, 40 different case studies from a keyword-based search in the academic area remained. To broaden the overview, towards the practitioner's perspective, the same criteria, keywords, and competency questions were used for a search procedure on industrial case studies. After the keyword search procedure, querying the Google Search Engine, 208 additional cases were identified.

Through the subsequent observation of the mentioned criteria and the use of the modified template, only three cases remained. Those are from Lufthansa [39], Dell [33] and Fujitsu [40]. One of the prevailing factors was the lack of specific information. Instead, mostly advertisements were presented to showcase the company's competence. One possible reason for this might be the missing motivation to present critical information, fearing the loss of competitive advantages. An overview of all of the described steps is depicted in Figure 1. In total, the material collection resulted in 43 different cases, which were further used in the course of this work. Although it was expected that an additional step in the form of a forward-backward search [29] will increase the number of promising cases, no new cases were identified. Most of the cases were published in 2016 (seventeen case studies), followed by 2017 (twelve case studies), 2018 (nine case studies) and finally 2015 (five case studies). A general investigation of the respective application domain reveals that the chosen case studies are coming from various areas.

Almost 25 percent of all contributions originate from the healthcare area. Another important area, with 38 percent,

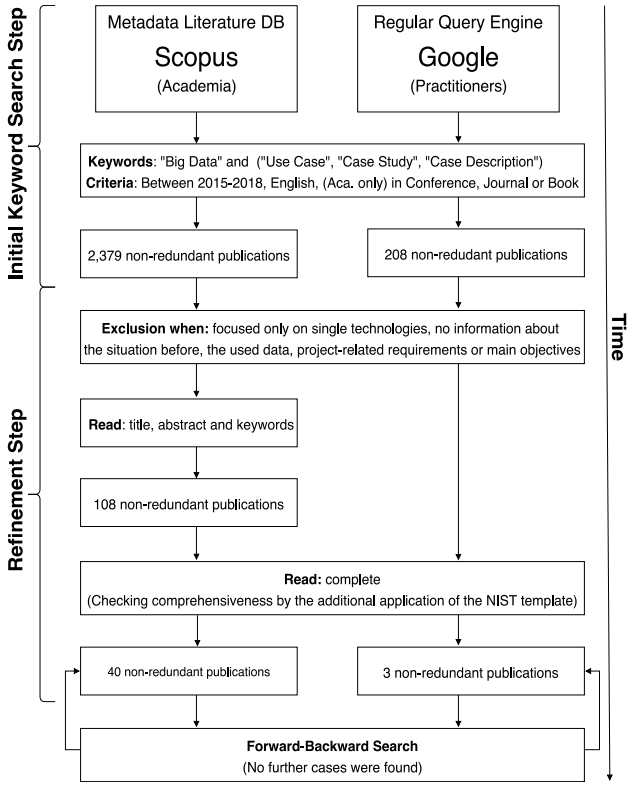

FIGURE 1. Conducted steps of the literature review, including the initial keyword search and the subsequent refinement steps.

is the internet of things (IoT), aiming to realize concepts like smart city, smart transportation, smart buildings, and more. Considering the various application areas, the time scope and the different databases from which the case studies originate, it can be concluded that those are a representative sample of the existing projects, regarding the usage of big data technologies. A complete listing of all of them is depicted in Table 3. While most of the related databases are directly addressed, *others* comprises the databases *Taylor and Francis*, *Gesellschaft fuer Informatik*, *ACM*, *IADIS Portal*, *Scitepress,* and the use cases originating from company resources. Furthermore, for each paper, a distinct number is assigned that will be used in the further context of this work.

TABLE 3. The Results of the Literature Review Mapped to the Respective Literature Databases.

Database	Paper References
Science Direct	**1.**[36], **2.**[41], **3.**[42], **13.**[43], **19.**[44], **27.**[45], **29.**[46]
IEEE Xplore	**4.**[47], **5.**[38], **8.**[48], **9.**[49], **10.**[37], **12.**[50], **14.**[51], **15.**[52], **16.**[53], **17.**[54], **18.**[55], **23.**[56], **24.**[57], **25.**[58], **30.**[59], **35.**[60], **37.**[61], **38.**[62], **39.**[63]
Springer	**7.**[64], **11.**[65], **21.**[66], **22.**[67], **31.**[68], **32.**[69], **33.**[70]
Other	**6.**[71], **20.**[72], **26.**[73], **28.**[74], **34.**[75], **36.**[76], **40.**[77], **41.**[39], **42.**[40], **43.**[33],

III. USE CASE ANALYSIS

Prior, a quantitative overview was given and a mapping of the selected case studies was performed. In the following, those will be analyzed in a detailed way. Manual approaches often result in great effort if the analysis has to be repeated or extended in possible future work. Because of the high number of found out results, the subjective observation that may come with this kind of investigation on extensive documents, and to increase the comprehensibility of this research, a more objective analytical approach was chosen. In particular, document clustering was selected, not only to find relevant information but also as described before, to identify standard use cases. Those shall faciliate decision support for practitioners and researchers, willing to perform a big data-related project. The creation of the intended solution, performed here, is equivalent to the general design and development of the implicitly followed design science research methodology.

A. SELECTION OF THE CLUSTERING ALGORITHM

To reduce the effort of analyzing the case studies, different methods that are typically used for document clustering have been compared with each other in terms of their applicability, as they are intensively investigated in various contributions [78], [79]. In particular, partitioned, density-based, and hierarchical types were investigated. While the first intends to put similar objects to the same cluster while maintaining the space between the different clusters as high as possible [80], density-based approaches follow the idea to identify regions where a high density (cluster) exists and separate them from those regions with a lower density (noise) [81], [82]. The last approach, on the other hand, is the building of a tree structure where each node, except for the leaf nodes, is a cluster that contains its children as sub-clusters (dendrogram) [78], [79], [82], [83]. Although all of the mentioned algorithms come with multiple potentials and benefits, not all of them are applicable to the current problem. Partitioned clustering, such as the K-means algorithm, for instance, allows the use of multiple calculation methods for cluster building [78]. However, some authors such as in [79] critically observed this kind of algorithms, stating issues such as the choice of the initial centroid, the strict predefined number of clusters, and further highlight the importance of other approaches [83].

A density-based approach, such as DBSCAN, tends to work best with noisy data and outliers, but it is not robust against high-dimensional data [83]. In this particular approach, high-dimensional data must be processed and no outliers are present. This results predominantly out of the meticulous qualitative assessment of the contributions during the literature review. Eventually, the hierarchical clustering algorithm was chosen as a suitable alternative, avoiding the need for starting parameters, specifying the strict number or size of the clusters [81]. Furthermore, also high-dimensional data can be recognized. Typically this approach *structures* a given dataset and "provide[s] a view of the data at different levels of abstraction" [79]. One of the most

frequently used approaches is constituted by the agglomerative clustering that assigns each object to one cluster and merges them until a whole tree is formed. The first step requires the calculation of a proximity matrix between the objects. Following that, the two closest clusters, with the lowest distance, are merged and the proximity matrix is updated for the new cluster. The procedure is repeated until only one cluster remains [79], [81], [83]. Compared to the K-means and the DBSCAN algorithm, the hierarchical clustering does not divide the points into end clusters, instead, a hierarchical structure based on the unit distance is provided. In terms of this, different results can be obtained through the alteration of the used distance metrics and linkage functions [81]. Due to the reason that all elements are connected to each other, outliers cannot be efficiently handled, but this was not required for the given task. However, at the same time, qualitative assessments can be realized through the manual definition of a desired distance. In summary, the basic steps that are needed for the clustering are the definition of the feature set for all use cases, the creation of the input matrix, the examination of the hierarchical clustering, the definition of the cluster structure as well as the determination of the intercluster distance. After everything is defined, the clusters need to be reviewed and modified in case they are not correctly assigned. By the end, those will be defined as standard use cases and thoroughly described. To summarize all of the aforementioned steps described before, an overview is depicted in Figure 2.

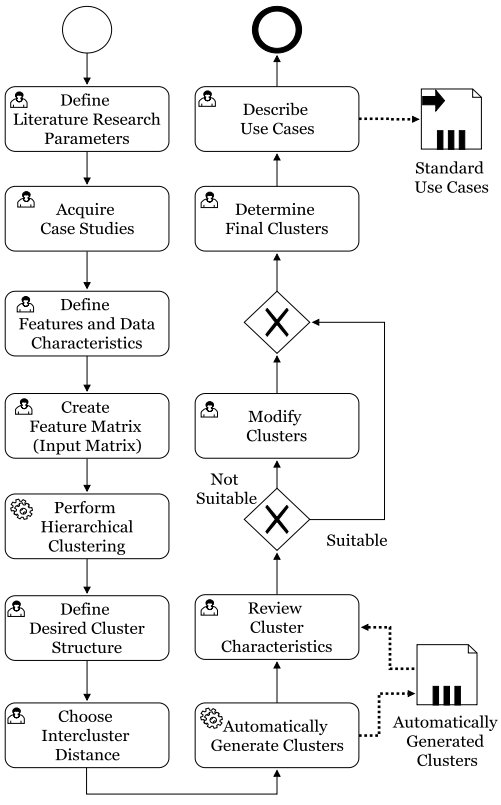

FIGURE 2. The use case analysis as a BPMN model.

B. CLUSTER ANALYSIS

As previously described, the standard use cases shall be deduced from the found out clusters of the hierarchical clustering, especially from a higher level of abstraction. In the beginning, an input matrix needs to be defined. Initially, the cases were collected and qualitatively checked by using a modified version of the NIST template, describing the current situation (e.g. represented by the aim and data characteristics) as well as the obtained solution (used methods and technologies). Although the template formed a promising starting point for the description of the feature matrix, it was not possible to use it as a direct input for the clustering algorithm. This is not only due to some unnecessary descriptive fields, such as title, author, or the rough description of the use case, but also for needed information that can be manifoldly expressed like the variety of the data or the used algorithms. After an additional examination of the filled templates, a total of 30 binary features were identified. For the construction of the input matrix, it was required to check each use case on the base of a formulated feature set. Since the number of use cases did not differentiate too much, compared to the number of attributes, this task was manually performed. An overview of all features together with the respective numbers of the index and the occurrences is depicted in Table 4.

As one may note, some of the listed features were more frequently identified compared to the others. In a descending order this includes: *Dynamic Data, Data Fusion, Unstructured Data, Heterogeneous Data, Statistical Calculations, Multiple Sources, Big Data Analysis, Real-time Data, Hadoop, and Batch Processing*. The complete mapping of all features and the respective use cases is given in Table 5. While one column represents one feature, each row stands for one use case. All of the features are related to the relevant data characteristics, used methods, fields of application, and also applied technologies. In terms of the data characteristics, for instance, it was needed to clarify whether the data is coming from various types of sources and if it should be shared between different users or applications during the operation phase. Following that, the type of file system was checked, in particular, if a distributed file system (e.g. HDFS), wide-area file system (e.g. Lustre) or parallel file system was used. If a particular use case (row) fulfilled one of the formulated features (column), a filled dot (•) was noted, whereas for not related features, an empty dot (○) was used.

Since most of the required functionalities of the intended algorithm are already included within Matlab, this computational framework was used for the cluster analysis. The created table was transformed into a binary matrix, filled with ones and zeros, transposed, and eventually used for the input. Besides the actual data, only a few *inputs* were required in Matlab. This includes the used distance measure and linkage function [84]. While the first computes the distance between each pair of observations, the second uses the calculated distance between the observations and links them according to the type of function that is chosen [81]. In particular, the Euclidian distance and Ward's

TABLE 4. A list of all features and their occurrences.

No.	Feature	Number of Occurrences
1.	Unstructured Data	40
2.	Semi-structured Data	4
3.	Structured Data	16
4.	Heterogeneous Data	40
5.	Historical Data	11
6.	Dynamic Data	43
7.	Real-time Data	32
8.	Event Processing	4
9.	Stream Processing	6
10.	Batch Processing	19
11.	High Velocity	20
12.	Multiple Sources	39
13.	Data Cleaning	5
14.	Data Pre-Processing	13
15.	Data Fusion	41
16.	Parallel Processing	7
17.	Parallel File System	1
18.	Distributed Computing	7
19.	NoSQL	21
20.	Internet of Things	15
21.	GIS	11
22.	Statistical Calculations	40
23.	Machine Learning	16
24.	Data Mining	16
25.	Structured and Unstructured Data	12
26.	High Performance Computing	4
27.	Big Data Analysis	34
28.	Map Reduce	20
29.	Hadoop	23
30.	Spark	8

linkage function were used, resulting in the dendrogram depicted in Figure 3.

On the x-axis, it shows each of the previously listed use cases. The y-axis describes the distance between the various use cases and aggregated clusters. While most of the existing clustering methods require a strict number of clusters or the elements contained in them, the used approach requires only minimum and maximum values for both, ranging from two to n (as the number of cases). During the investigation and formulation of the features, huge differences between the cases were noticed in parts, which directly influenced this range selection. By having too many clusters that differ only slightly from each other, a decrease in the usability for later standard cases can be expected. This would diminish the general idea of standard use cases, especially when it comes to the classification of a planned project for a potential user. Apart from the time-consuming planning steps needed in beforehand, also detailed knowledge about specific features would be required to make further distinctions. This, in turn, would neglect the general sensibility and applicability of the targeted outcome. For that reason, the inter-cluster distance together with the number of agglomerated clusters was examined to understand at which point all cases were assigned. As one can note in the depicted diagram, at a distance of two, only six clusters consisting out of multiple cases are built, whereas 31 cases remain as one separate cluster. At the level of 3.5, only one case remained unassigned and in total 13 clusters were formed. A distance of 4 resulted in seven distinct clusters, which comprise all of the 43 use cases. At a level of 4.5, only five agglomerated clusters exist. By having the aforementioned disadvantage of too few cases in mind, the achieved seven clusters at a distance of 4 were chosen. A further qualitative assessment and cross-checking of each of these seven clusters, however, revealed the disadvantage of the non-weighed Euclidian distance function.

C. EXAMINATION OF THE BUILT CLUSTERS

Due to the reason that only binary decisions on the feature set are recognized, no in-depth information extraction and connections were made yet. As one can note, some of the use cases revealed a rather high inter-cluster distance. For those reasons, and to obtain a better understanding of the similarities of those cases, further examinations were required. First and foremost this includes the overall aim of the use case and the interplay of the features. In the following, each of those and their specifics are described in detail. Table 6 provides an overview of the automatically built clusters and their assigned use cases at a distance of 4.

1) DESCRIPTION OF CLUSTER 1

The first automatically built cluster comprises the seven use cases no. ***2*** [41], ***3*** [42], ***5*** [38], ***6*** [71], ***8*** [48], ***37*** [61] and ***39*** [63]. One particularity of the cases located in this cluster that was noticed at the very beginning of the examination and comparison, was the aim to improve already existing analysis processes. On one side this comprises the automation of currently manually performed actions, like the financial rumor detection [71], the monitoring of patients at their homes, instead of using the hospital capacities [63], or the analysis of traffic sensors data to recognize congestions and traffic patterns [48]. On the other side, the improvement of existing processes can be defined as achieving better quality of the performed analysis. This can be realized by using all of the available information, like unstructured medical and genome data on the way to personalized medicine [63], investigating the meaning of social media data to improve crisis mapping systems [41] or the broadening of the scope [61].

All of those cases make use of unstructured data, coming from various sources in different formats. Especially if those have to be realized in real-time, sophisticated approaches are required. This is for example the case for the automatized financial rumor detection, considering more than 300 trades per day [71], or the simulation of wind turbine

TABLE 5. Resulting matrix showing highlighting the occurrence of a feature (F) in a specific use case (U).

U/F	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	●	○	○	●	●	●	○	○	○	●	●	●	○	●	●	○	○	○	○	●	○	●	○	●	○	○	●	○	○	○
2	●	○	○	●	●	●	○	○	●	○	●	●	○	●	●	○	○	○	●	○	○	●	○	●	○	○	●	●	○	●
3	●	○	○	●	○	●	●	○	○	○	●	●	○	●	●	○	○	○	○	●	○	●	●	●	○	○	●	●	○	●
4	●	○	○	●	●	●	●	○	○	○	●	●	○	○	●	○	○	○	●	●	○	○	○	○	○	○	○	○	○	○
5	●	○	○	●	●	●	●	○	●	○	●	●	○	○	●	●	○	○	○	●	○	●	●	○	○	○	●	○	●	○
6	●	○	○	●	●	●	●	○	●	○	●	●	●	●	●	○	○	○	○	○	○	●	●	○	○	○	●	○	●	○
7	●	○	○	●	○	●	○	●	○	●	○	●	○	○	●	●	○	○	○	○	○	●	●	●	○	○	●	●	●	○
8	●	○	○	●	●	●	●	○	○	●	●	●	○	○	●	○	○	○	○	●	○	●	●	●	○	○	●	●	●	●
9	●	○	○	●	○	●	○	○	○	●	○	●	○	○	●	○	○	○	●	●	○	●	●	●	○	○	●	○	○	○
10	●	○	○	○	○	●	○	○	○	●	○	○	●	●	○	○	○	○	●	○	○	●	●	●	○	○	●	●	●	○
11	●	○	●	●	○	●	○	○	○	●	○	●	●	●	●	○	○	○	●	○	○	●	○	○	○	○	○	○	●	○
12	●	○	●	●	○	●	●	○	○	○	○	●	○	○	●	○	○	○	●	○	○	●	○	●	○	○	●	○	●	○
13	●	○	○	●	○	●	●	○	○	●	●	●	○	○	●	○	○	○	○	●	●	●	○	○	○	○	○	●	○	○
14	●	○	○	●	○	●	●	○	○	○	○	●	○	○	●	○	○	○	●	○	○	○	○	○	○	○	○	○	○	○
15	●	○	●	●	●	●	●	○	●	●	○	●	○	○	●	○	○	●	●	○	○	●	●	○	○	○	●	○	●	○
16	○	○	●	○	○	●	●	○	○	○	○	○	○	○	●	○	○	○	●	○	○	●	○	○	○	○	○	○	○	○
17	●	○	●	●	○	●	●	○	○	●	●	●	○	●	●	○	○	○	●	●	○	●	●	○	●	○	●	○	○	○
18	●	○	●	●	○	●	●	○	○	●	○	●	○	●	●	○	○	●	○	○	○	●	○	●	●	○	●	○	○	○
19	●	○	○	●	○	●	●	○	○	○	○	●	○	○	●	○	○	○	○	●	●	●	○	○	○	●	●	●	●	●
20	●	●	●	●	○	●	●	○	○	○	●	●	○	○	●	●	○	○	○	○	●	●	○	○	●	○	●	●	●	●
21	●	●	○	●	○	●	●	○	○	○	○	●	○	○	●	○	○	○	○	●	●	●	○	○	○	○	●	○	○	○
22	○	○	○	○	●	●	●	○	○	●	●	●	○	●	●	●	○	○	○	○	●	●	○	○	○	●	●	○	○	○
23	○	○	○	●	○	●	●	○	○	○	●	●	○	○	●	○	○	○	●	●	●	●	○	○	○	○	○	○	○	○
24	●	○	●	●	○	●	●	○	○	○	●	○	○	○	●	○	○	○	●	○	○	●	●	●	●	○	●	●	●	●
25	●	○	●	●	○	●	●	●	○	●	○	●	○	○	●	○	○	○	○	●	●	●	○	●	●	○	●	○	○	○
26	●	○	○	●	○	●	●	○	○	●	○	●	○	○	●	○	○	●	○	○	○	●	○	○	○	○	●	●	●	○
27	●	○	●	●	○	●	○	○	○	●	●	●	○	○	●	○	○	○	○	○	○	●	○	○	●	○	○	●	●	○
28	●	●	●	●	○	●	○	○	○	○	○	●	○	○	●	○	○	○	●	○	○	●	●	○	●	○	●	○	●	○
29	●	●	●	●	○	●	●	○	○	○	●	●	○	○	●	○	○	○	●	●	○	●	○	○	●	○	●	○	○	○
30	●	○	○	●	○	●	●	●	○	●	○	●	○	○	●	○	○	○	●	●	●	●	○	○	○	○	●	●	○	○
31	●	○	○	●	○	●	●	○	○	○	●	●	○	●	●	●	●	○	○	○	○	●	○	●	○	○	●	○	●	○
32	●	○	●	●	○	●	●	○	●	●	○	●	○	○	●	○	○	○	●	○	●	●	●	●	●	○	●	●	●	●
33	●	○	○	●	○	●	○	●	○	○	○	●	○	○	●	○	○	○	●	○	○	●	○	○	○	○	○	○	○	○
34	●	○	○	●	○	●	●	○	○	○	●	●	○	○	●	○	○	●	●	○	○	●	○	○	○	○	●	○	●	○
35	●	○	○	●	○	●	●	○	○	○	○	●	○	○	●	○	○	○	●	●	●	●	○	○	○	○	●	○	○	○
36	●	○	○	●	○	●	●	○	○	○	○	●	○	○	●	○	○	●	●	○	●	●	●	●	○	○	●	●	●	○
37	●	○	○	●	●	●	●	○	○	○	●	●	●	●	●	○	○	○	○	○	○	●	●	○	○	●	●	●	○	○
38	●	○	●	●	○	●	●	○	○	○	○	●	○	●	●	○	○	○	○	○	○	●	●	●	●	○	●	●	●	○
39	●	○	○	●	○	●	●	○	○	○	●	●	●	●	●	○	○	○	○	○	○	●	○	○	○	○	●	●	●	○
40	●	○	●	●	●	●	●	○	○	○	○	●	○	○	●	●	○	●	○	○	○	●	○	●	●	○	●	●	●	○
41	●	○	●	●	●	●	●	○	●	●	●	●	○	○	●	○	○	●	●	○	○	●	○	○	●	○	●	●	●	○
42	●	○	○	●	○	●	○	○	○	●	○	○	○	○	○	●	○	○	○	○	○	●	○	○	○	●	●	○	●	○
43	●	○	○	●	○	●	○	○	○	●	○	●	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○	●	○	○

U/F	31	32	33	34	35	36	37	38	39	40	41	42	43
1	●	●	●	●	●	●	●	●	●	●	●	●	●
2	○	○	○	○	○	○	○	○	○	○	○	○	○
3	○	●	○	○	○	●	○	●	○	●	●	○	○
4	○	●	○	○	○	●	○	○	○	●	●	○	○
5	○	○	○	○	○	○	○	○	○	○	○	○	○
6	○	○	○	○	○	○	○	○	○	○	○	○	○
7	○	○	○	○	○	○	○	○	○	○	○	○	○
8	○	●	○	○	○	●	○	●	○	●	●	○	○
9	○	○	○	○	○	○	○	○	○	○	○	○	○
10	○	○	○	○	○	○	○	○	○	○	○	○	○
11	○	○	○	○	○	○	○	○	○	○	○	○	○
12	○	○	○	○	○	○	○	○	○	○	○	○	○
13	○	○	○	○	○	○	○	○	○	○	○	○	○
14	○	○	○	○	○	○	○	○	○	○	○	○	○
15	○	○	○	○	○	○	○	○	○	○	○	○	○
16	○	○	○	○	○	○	○	○	○	○	○	○	○
17	○	○	○	○	○	○	○	○	○	○	○	○	○
18	○	○	○	○	○	○	○	○	○	○	○	○	○
19	○	○	○	○	○	○	○	○	○	○	○	○	○
20	○	○	○	○	○	○	○	○	○	○	○	○	○
21	○	○	○	○	○	○	○	○	○	○	○	○	○
22	○	○	○	○	○	○	○	○	○	○	○	○	○
23	○	○	○	○	○	○	○	○	○	○	○	○	○
24	○	○	○	○	○	○	○	○	○	○	○	○	○
25	○	○	○	○	○	○	○	○	○	○	○	○	○
26	○	○	○	○	○	○	○	○	○	○	○	○	○
27	○	○	○	○	○	○	○	○	○	○	○	○	○
28	○	○	○	○	○	○	○	○	○	○	○	○	○
29	○	○	○	○	○	○	○	○	○	○	○	○	○
30	○	○	○	○	○	○	○	○	○	○	○	○	○
31	○	○	○	○	○	○	○	○	○	○	○	○	○
32	○	○	○	○	○	○	○	○	○	○	○	○	○
33	○	○	○	○	○	○	○	○	○	○	○	○	○
34	○	○	○	○	○	○	○	○	○	○	○	○	○
35	○	○	○	○	○	○	○	○	○	○	○	○	○
36	○	○	○	○	○	○	○	○	○	○	○	○	○
37	○	○	○	○	○	○	○	○	○	○	○	○	○
38	○	○	○	○	○	○	○	○	○	○	○	○	○
39	○	○	○	○	○	○	○	○	○	○	○	○	○
40	○	○	○	○	○	○	○	○	○	○	○	○	○
41	○	○	○	○	○	○	○	○	○	○	○	○	○
42	○	○	○	○	○	○	○	○	○	○	○	○	○
43	○	○	○	○	○	○	○	○	○	○	○	○	○

●: was documented; ○: was not found

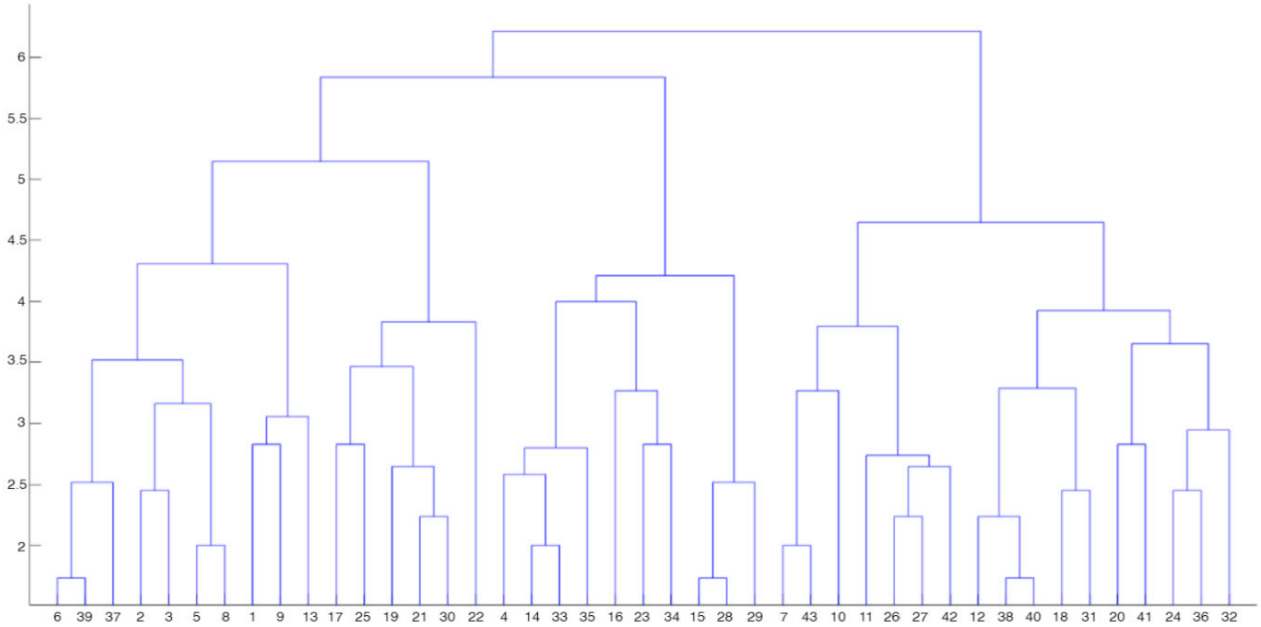

FIGURE 3. Dendrogram of the cluster analysis, showing the cluster distance (y-axis) and use case number (x-axis).

TABLE 6. Automatically built clusters of the hierarchical clustering.

Cluster No.	Mapped use cases and their respective publication reference
1	**2.**[41], **3.**[42], **5.**[38], **6.**[71], **8.**[48], **37.**[61], **39.**[63]
2	**1.**[36], **9.**[49], **13.**[43]
3	**17.**[54], **19.**[44], **21.**[66], **22.**[67], **25.**[58], **30.**[59]
4	**4.**[47], **14.**[51], **16.**[53], **23.**[56], **33.**[70], **34.**[75], **35.**[60]
5	**15.**[52], **28.**[74], **29.**[46]
6	**7.**[64], **10.**[37], **11.**[65], **26.**[73], **27.**[45], **42.**[40], **43.**[33]
7	**12.**[50], **18.**[55], **20.**[72], **24.**[57], **31.**[68], **32.**[69], **36.**[76], **38.**[62], **40.**[77], **41.**[39]

configurations resulting in a data stream of 100MBytes per hour [61]. Furthermore, dynamic and permanent (historic) data are used in all of the cases. The same hold true to the application of statistical methods, like market activity statistics [71] or the possibility of a medical condition to occur in different situations [42], [63]. Moreover, the search, query, and indexing of the data, used in the cases, aiming to improve existing processes, should be enabled. The searching process of the available data is an important part of the analysis. To make the analysis of big data easier, the data can be classified in different categories.

The consensus between all the cases is that they rely on deep learning techniques to reach their goals. Those techniques are for example needed to map financial news to a trained set to decide the authenticity of the news [71]. In the medical sphere, machine learning is used to uncover similarities between patients and thus to faciliate the proper therapy prediction, or to enable the remote patient monitoring by making sense of the medical records and pre-defined rules [42], [63]. In [61] deep learning is used to analyze different configurations of wind turbines, in order to decide about their optimal location and design.

The technologies used on the way for the optimization of existing processes differ from one another, as all of those cases have their specific subtasks. Some of the cases make use of the HDFS to deal with the volume of the data [38], [48], [63], [71]. Those analyze some of the data in batch mode to create a trained set, which is later used as a basis for the real-time analysis and the decision-making. Furthermore, in one of the cases studies, for the purpose of reducing the dimensionality of the data, the parallel file system Lustre is used [61], allowing the simulation of different turbine configurations. For the implementation of the crisis and mapping system [41], the ElasticSearch database is used, as this one enables the near real-time processing of the data. To enable remote patient monitoring and sensemaking of all the collected sensor data, in [42], the analytics engine Spark is used to speed up the query performance.

2) DESCRIPTION OF CLUSTER 2

The second cluster comprises the three use cases no. ***1*** [36], ***9*** [49], and ***13*** [43]. Each of them focuses on value creation through the analysis of IoT sensor data. The data itself originates in each case from various sources. In [36], for instance, smart meters, different appliances, and smart home devices are used. Diverse weather sensors that measure temperature, wind speed, and humidity are sourced in [49]. In the case of [43], gas or imaging sensors that can typically be found in recycling systems are focused. Apart from the behavior of the occupants of smart buildings [36], or anomalies and failure

detection of the weather sensors [49], also statistics over the recycled goods and their usage to improve the recycling system are targeted [43].

In each of the cases, different kinds of sensors are used. Big data technologies are needed in all of them to cope with the unstructured, large amounts of data. In particular, comprehensive analysis and clustering algorithms are applied to uncover yet unknown patterns, as they were described before. For this reason, special pre-processing measurements like data cleaning, which removes outliers and irrelevant data, and measurements to structure the data have to be applied [36], [49]. To receive the desired insights, detecting the previously hidden patterns, a real-time data processing is not necessarily needed. For example, for the detection of failures in the weather sensors in [49], the data is firstly collected and bulk-loaded into the IoT framework. Afterward, the data is analyzed to identify similar values and potential failures. The analysis of the data in the recycling system is realized after the data from the different sensors were collected and pre-processed [43]. The main aim of all use cases of this cluster is to uncover some relevant patterns out of the huge amount of IoT data. To reach this target, the gathered data needs to be classified and subsequently analyzed with the help of suitable algorithms. In the case of [36], the K-means algorithm is used to discover the hourly usage of appliances and their usage on different weekdays by the occupants of a building. This algorithm is also used to discover patterns out of the values, delivered from different weather sensors [49].

Another important part of the analysis process is the usage of basic statistics. When the analysis is performed, the visualization of the results is mandatory. For instance, this is used to represent the hourly usage of different household devices from the occupants of a building [36] or to show clusters, built on the basis of the different weather sensor values [49]. As this group tries to realize the concept of IoT in different areas of life, the sharing of information between users or devices should be enabled. For example, the occupant of a smart home should be able to exchange information with his household devices like dishwasher or oven [36]. Another example of such information exchange can be observed in the smart recycling system, where retailers and consumers should be also able to communicate [43]. The various uncovered patterns can be used, for instance, to generate energy reduction recommendations, to avoid failures in weather sensors, and to increase the efficiency of the current recycling systems [36], [43], [49].

3) DESCRIPTION OF CLUSTER 3

Based on the set of given features and their occurrences in the respective use cases, no. ***17*** [54], ***19*** [44], ***21*** [66], ***22*** [67], ***25*** [58] and ***30*** [59] were automatically assigned to the third cluster. Again, a thorough investigation and comparison was made to identify conspicuous similarities. Three of the cases are aiming to realize smart city concepts [44], [59], [66]. The strong connection of those can solely be observed according to the low inter-cluster distance in Figure 3, but also regarding their overall scope. For example, in [66] the concept of an itinerary planning platform for tourists, which suggests activities according to pre-defined criteria, such as location, time, and period preferences is proposed. In the case of [44], unstructured data is used to improve transportation. In doing so, information about incidents on the highway shared by tweets on Twitter, videos of a disaster, or pictures of a traffic jam are utilized. A touristic recommender system is shown in [59]. The personalized recommendations are not only based on permanent data, such as information about city infrastructure, existing restaurants, or hotels, but also dynamic data that is constantly changing. The latter is not only related to the velocity of the data but also its structure. These data include, for instance, information coming from wearable bracelets, social networks, and used sensor data applied for the traffic and weather tracking [44], [59], [66]. Eventually, all of those cases consider data coming from different sources. Hence, the data fusion plays a dominant role in all of those cases.

In this context, geographic information systems (GIS) are used to gather, store and analyze the whole geographic information such as the location of users, traffic jams, incidents, disasters, hotels, restaurants or attractions [44], [59], [66]. For the actual provisioning of personalized recommendations, real-time data analyzes are required. In [44] those are realized through the use of Spark streaming. Furthermore, in each of those, sophisticated statistical methods are applied. As the steps of data gathering, processing, and analysis are conducted, the results need to be represented understandably and appealingly for the user. Therefore, the usage of various visualization techniques is crucial for this group. For instance, the personal touristic recommendations for activities, accommodations, and restaurants in [59] are shown. Another example of visual techniques is the representation of the current transportation situation for a selected region in video or image format [44].

However, this cluster contains three further cases that have not been considered yet [54], [58], [67]. One of them represents a system for the remote 24/7 patient monitoring, which can be also used to determine the future medication procedure [54]. The second one represents a framework that utilizes the data originating from the financial sector and various IoT devices to improve the user experience [58]. Although connections to the previously described cluster were identified, distinctions in terms of the data processing and the format were observed. In particular, real-time and batch processing on differently structured data is performed. The last case study makes use of high-performance computing (HPC) to analyze and optimize the operation of wind turbines [67]. This one also deviates from the smart city concept, represented in the first three cases. Despite the fact that those are serving in here as outliners, all of them share other characteristics, such as the processing speed, used data formats, and the overall aim to improve existing methods. Due to those similarities, they could serve as a *cluster* or rather a *group* themselves.

4) DESCRIPTION OF CLUSTER 4

Cluster four contains the cases no. ***4*** [47], ***14*** [51], ***16*** [53], ***23*** [56], ***33*** [70], ***34*** [75] and ***35*** [60]. During the observation and further examination of each of those, a similar distinction as in the previous cluster was observed. In particular, two *sub-groups* were identified, since they do not share one sub-cluster but instead multiple characteristics.

The first sub-group comprises the cases originating from [47], [56], [60]. They have in common that the main focus is put on the realization of smart city concepts. The first case in this group attempts to integrate information from sensors and IoT devices used in a building, weather information sensors as well as data from environmental sensors in a cognitive building framework [47]. This framework shall improve energy consumption by learning from the behavior of the habitant and adjusting the functionality of the devices accordingly to the users' behavior [47]. The second case introduces a smart traffic pilot, making use of the traffic light data, weather and disaster information, as well as GPS data about the positions of the vehicles. This information can be used in different applications like route optimization or a driving coach, suggesting fuel-saving driving patterns [60]. The last case in this group proposes a platform that integrates data from IoT devices, GIS, and energy-related information to improve the energy consumption and to reduce the CO_2 emissions [56]. Each of those three concepts corresponds to the general aim of realizing the smart city concept.

Notwithstanding that, all of them make use of data, coming from various devices like household appliances, environmental sensors, smart meters, or GPS devices [47], [56], [60]. This heterogeneous data is mostly unstructured and can have different content formats, in the form of texts, images, or videos. Furthermore, personally identifiable information about the habits of the buildings' occupants, GPS, and passengers in a car are used [47], [56], [60]. The personally identifiable information together with the input about the city and buildings' infrastructure make up the permanent data used in this group. Moreover, the real-time processing or at least the near-real-time processing of the data should be enabled by the proposed smart city concept. For example, GPS data and environmental information should be processed on the fly so that a plausible route recommendation can be delivered [60]. Further, the near-real-time integration of data, coming from IoT devices and various networks (e.g. electrical and heating networks), is a requirement for the energy management systems, proposed in the last case in this group [56].

Additionally to those aspects, it was found out that all cases use GISs to locate the user, particular vehicles, or relevant objects. To represent the delivered smart city solution, the cases deploy different visualization techniques. For example, the analyzed driving behavior of the user and the derived fuel economy recommendations can be represented in a mobile application [60]. This solution was also used in [56] to visualize the energy consumption data and the possible improvements that can be conducted. To derive the above-mentioned recommendations, various statistical methods are used. Those methods are needed to calculate values such as the energy consumption in particular rooms, average fuel consumption on a road segment, or corresponding indicators for certain time frames [47], [60].

The remaining cases are no. ***14*** [51], ***16*** [53], ***33*** [70] and ***34*** [75]. Those have different application areas and neither fit in the smart city concept nor can build up a separate group. The first one presents a smart clinical workflow implementation that should automate some parts of the patient care [51]. This one could fit to the smart city concept, but the developed solution does neither make use of visualization techniques nor a GIS and thus does not fit in the already built subgroup. The second one [70] differs strongly regarding the aim to integrate several bioinformatics databases and, thus, improve the scalability of the cancer analytical system. The next case study has a similar aim to the previous one. Here, the query performance of a library information system needed to be improved [53]. However, it has a different pattern of the fulfilled feature set, including only structured data from only one source. The last case study considers event-manufacturing data and aims to improve existing processes [75]. This one does not maintain personally identifiable information and does not use GIS, which is a crucial requirement of the above-defined smart city subgroup.

5) DESCRIPTION OF CLUSTER 5

The next cluster, derived from the hierarchical clustering results, consists of three cases, namely no. ***15*** [52], ***28*** [74] and ***29*** [46]. The general aim of those cases can be described as the integration of data from different sources, improving the scalability and leading to better analysis results. For example, the first case study proposes a system for real-time traffic control [52]. In comparison to the existing traffic control systems, the proposed solution should be able to consider more than one aspect by involving more data resources. The second case study aims to improve the quality of the user experience in the communications area by involving more resources in the analysis. In the past, data mining methods were used in the telecom area to figure out problems only in an isolated way, for example fraud detection based on call detail records. In order to consider different telecommunication aspects, nowadays various information, coming from mobile networks, GPS devices and social media has to be considered [74]. The target of the last case study in this category is to turn regular factories into smart factories, where resources and machines communicate and deliver smart products that are aware of their production history. To realize those factories, the integration of data from different machines and operators needs to be pursued [46].

The data used in this cluster can be both, structured and unstructured. Examples for structured data can be records of the log and machine operating times [46], [74]. In contrast, social media data, camera pictures, and sensor data are examples of unstructured data [46], [74]. Furthermore, all three case studies maintain personally identifiable information, which can require special processing techniques.

This information can be the location of a driver, smart card data of public transport users, or the call history log of a telecom company customer [52], [74]. Permanent as well as transient data (data that is deleted by the end of a session) play an important role in the performed analysis in this cluster.

Regarding the processing of the data, real-time analytics are required. For instance, to enable real-time traffic control, incident detection has to be performed on the fly [52]. The same applies to the communication between products, resources, and machines to realize smart factory concepts [46]. Besides that, in all of the cases, NoSQL databases are used to enable the querying of the diverse information. In the presented solutions, basic statistics are used to calculate values such as average speed, travel time, subscriber churn likelihood, and operational time of a machine [46], [52], [74]. Deep learning algorithms are used to reach the goals of the analysis. Those enable, for example, the consideration of user ufeedback in the analysis of the users experiences's quality [74].

6) DESCRIPTION OF CLUSTER 6

The sixth cluster includes the seven case no. ***7*** [64], ***10*** [37], ***11*** [65], ***26*** [73], ***27*** [45], ***42*** [40] and ***43*** [33]. When inspecting the cases, two different groups were found out.

The first one consists of the four use cases no. ***7*** [64], ***27*** [45], ***42*** [40], and ***43*** [33], aiming to deal with the growing amount of medical data and to improve the analysis quality in the healthcare area through the integration of additional data. In the first use case of this group [64], a Hadoop ecosystem to deal with the volume, variety, and velocity of medical data, coming from various applications and devices is proposed. The use case no. 42 [40] deals with the analysis of human genome data, which is continuously growing and even HPC clusters cannot deal with the challenge to process this amount of data. As a solution to the problem, a Hadoop system is proposed. In [45] a framework that allows querying both structured and unstructured medical data is introduced. The main goal is to improve the decision basis for medical experts. The last case [33] in this group presents a cloud-based analytics solution that should turn the massive amounts of medical data into value. In almost all cases, structured and unstructured data, located in the medical area, are used. While the first mainly contains structured documents like patient records like the name, age, or previous diagnosis, the latter comprises images, clinical notes, unstructured documents, and genome data [33], [40], [45], [64]. Due to the handling of personally identifiable information, which falls under special regulations, sophisticated security measurements and storage solutions are required in the final system [33]. Again it was noticed that data fusion is of major importance, because of the use of data originating from different (healthcare) institutions, devices, and other sources. For example, in [64] a system is proposed that is intended to improve the healthcare situation in Algeria by an efficient distribution of medical resources and staff. To achieve this, information from one university hospital, five public hospitals, one medical school, 51 polyclinics and some laboratories needs to be integrated. For the technical implementation, all of the use cases rely on the HDFS. Furthermore, it was found out that in none of the use cases a real-time processing was needed [40], [64]. For the analysis itself, all use cases utilize basic statistics and data mining.

Compared to the cases described before, all of the remaining, including ***10*** [37], ***11*** [65], and ***26*** [73], follow a different aim. In here, the linkage of data from different sources is of major interest. The first case [37] of those presents a framework for the detection of insurance frauds. At the moment, insurance fraud detection methods are solely being used in separate fields like healthcare, financial services, and others. To find a suitable solution and to achieve a broad coverage, data from 34 different fields, including sources such as customer information, contracts and insurance claims, are integrated [37]. Case no. 11 [65] presents an architecture for the processing of data coming from different social media channeld and in [73] a data management system environment that is supposed to deal with unstructured data from various sources is introduced. All of those case studies require the processing of unstructured data [37], [65], [73]. Most of the considered data such as insurance contracts and claims are in an unstructured textual form [37]. As the cases in this subgroup propose frameworks and architectures for the integration of the vast amount of data from different sources, the cleaning of the original data constituted an important step. This includes auxiliary activities, like outlier detection or fixing missing values [37], [73]. Similar to the previous group, real-time data processing is not required.

7) DESCRIPTION OF CLUSTER 7

The last cluster, resulting from the hierarchical clustering, contains the ten use cases no. ***12*** [50], ***18*** [55], ***20*** [72], ***24*** [57], ***31*** [68], ***32*** [69], ***36*** [76], ***38*** [62], ***40*** [77] and ***41*** [39]. Similar to the sixth cluster, the use cases can be split in two separate groups, regarding their reasons for using big data technologies.

The first group consists of the three cases [55], [62], [68] that are focused on supporting informed decision making, mainly in the healthcare area. The first case study attempts to provide a basis for precision medicine, by facilitating the data analysis of various molecular profiles [55]. The second proposes a general framework that should enable the integration of healthcare data from various resources and thus allow researchers to conduct innovative types of analysis [62]. The last case in this group aims to improve the daily practices in a hospital by utilizing transactional data [84]. All of the three cases make use of heterogeneous, unstructured data of different content formats, such as text, images (e.g. diagnostic tests), signals, and phenotypes [62]. Besides that, personally identifiable information, in the form of patient records, are used [55], [62], [68]. A further important feature that is shared in all of the case studies, is the initial data pre-processing. Here, it constitutes a crucial step as it increases the quality of the information used for the decision-making process. It is

for example used to deal with spelling and grammar mistakes or anonymization [55], [62], [68]. The analysis, used for the decision making, is performed by the help of basic statistics and data mining techniques. Different associations, classification, and clustering methods are incorporated, in order to discover relevant patterns, as well as to figure out inter-attribute correlations and important relations [62], [68].

The second group contains the remaining six documents, namely the cases no. ***12*** [50], ***20*** [72], ***24*** [57], ***32*** [69], ***36*** [76] and ***41*** [39]. All of them share the aim to enable the real-time analysis of data, incoming with high-speed. In the first of the case studies, a framework for the online analysis of high-speed physiological data is proposed, which should improve the neonatal intensive care [50]. The second one uses real-time analysis to forecast the power output of solar plants [72]. Within the third case study, a framework for the analysis of patent information and its usage for research and development is introduced [57]. The subsequent case no. 32 [69] proposes an analytic platform for smart transportation, which analyzes data from heterogeneous data sources such as sensors and cameras in real-time. In [76] real-time analysis is used to process social media data for a disaster management system. The last use case uses social media data to improve the quality of passenger's experience [39]. As one may note, all of the described use cases in this group focus on real-time data analysis. The origin for this circumstance arises presumably out of the high velocity.

A further important feature of this group is the processing of heterogeneous data that is coming from different devices or source. For example, cameras and traffic sensors are used as an input for the realization of a smart transportation system in [69]. Apart from that, social media data, for example provided by Twitter, Google+, or YouTube, can be harnessed for disaster management or the improvement of the quality of the customer's experience [39], [76]. For case no. [72] structured, semi-structured (e.g. weather forecast data) and unstructured data (e.g. customer behavior, video files) are used. Because of the flexibility and scalability, the HDFS is incorperated as the foundation for each use case. The final results are eventually visualized and presented for instance by dashboards [72], bar charts [57], or a decision map [76].

D. DERIVED STANDARD USE CASES

From the first explanation and qualitative examination of the seven found out clusters it was noted that some of the use cases did not match properly to each other, even though they were assigned to one agglomerated cluster. Hence, modifications were required in different ways, such as insertion, deletion, and consolidation, to better highlight the data characteristics, used methods, and aim of the use case. Above all, this was required to ensure that not only key indicators, such as the distance, are used for the creation of the standard use cases, but also qualitative assessments are realized. Subsequentelly, in total, nine different clusters were derived from the qualitative examination. Those are depicted in Table 7.

In the first and second column, the number of the derived cluster as well as the mapped use cases are stated. The general aim, all relevant features, and the needed modifications are described in the remaining columns.For instance, it was noted that the first identified groups of clusters three and four, containing cases no. 19, 21, 30, and 4, 23, 35, share similar interests. Besides the general focus on smart cities, also the same characteristics are shared, except for one case (no. 23) that uses near-real-time processing instead of real-time-processing [56]. Hence, both of them were merged into one new cluster (cf. Table 7 cluster no. 3). All remaining cases of the initially calculated third cluster, focusing on sensor analysis, became the new second cluster.

Furthermore, the cases 16, 33, 34, 40 appeared to be as outliers, not only for the respective fourth cluster but in parts also for the entire dataset. This does not represent an error in the qualitative analysis but rather how heterogeneous the individual use cases can be. For instance, in case no. 16 [53], the main goal was to improve the query performance of a library information system. Within the given collection, this was the only case that solely handled structured data. Case no. 33 [70] exclusively used graphical-processing of the data to efficiently handle queries on multiple integrated bioinformatics databases. To prevent misleading information, those use cases were removed or assigned to another cluster. Case no. 14 [51] proposes a smart clinical workflow that aims to increase the volume of medical data that can be processed. In doing so, health data from different sources are integrated and used to facilitate predictive therapy and to improve the wellbeing of the patient. Due to the similarities to the first cluster and the goal to generally improve the quality of the performed analysis, this case was assigned to said cluster.

As already highlighted during the description of the original third cluster, the cases no. 17 [54], 22 [67] and 25 [58] revealed no real interconnection to the overall features presented in this cluster. By comparing those cases, it was found out that all of them aim to optimize existing processes by using big data. Eventually, a new cluster was manually built (cf. Table 7 cluster no. 9). The separate groups, which were identified within the initially calculated sixth and seventh cluster were extracted and declared as a separate one. Hence, out of both clusters, two additional ones emerged (cf. Table 7 cluster no. 5-8).

Overall, for most of the initially formed clusters, only minor modifications were required. Researchers, as well as practitioners, can utilize these to obtain an idea not only about the general meaningfulness of their own project but also possible implementation details from specific use case descriptions. This is especially the case if a similar approach is pursued. For an increased understandability, within the following paragraphs, each of the found out standard use cases are briefly described, comprising the common features in a narrative way.

TABLE 7. Standard Use Cases (UC) derived from the clusters.

No	UC	Aim	Relevant Features	Mod.
1	2, 3, 5, 6, 8, 14, 37, 39	Improve the analysis algorithms	high velocity; data used for analysis; real-time (near-real-time) processing; data fusion; unstructured data; permanent data; dynamic data; basic statistics; search/query/indexing; classification	Case 14 was added
2	1, 9, 13	Analyze the data from (IoT) sensors	data from different devices/sensors; visualization; data used for analysis; unstructured data; batch processing; data shared between users/applications; classification; clustering algorithms	Cases left from the third initial third cluster
3	4, 19, 21, 23 30, 35	Realize smart city concepts	data fusion; visualization; unstructured data; real-time (near-real-time) processing; personally identifiable information; permanent data; data shared between users/applications; basic statistics	Merged out of the third and fourth cluster
4	15, 28, 29	Integrate heterogeneously structured data for multi-leveled problems	structured and unstructured data; real-time processing; data fusion; personally identifiable information; permanent data; transient data; data shared between different users/devices; NoSQL; dynamic data; deep learning; basic statistics; search/query/indexing; data classification	Former fifth cluster
5	7, 27, 42, 43,	Improve the analysis quality through additional data	data coming from different institutions; unstructured data; valuable data; batch-processing; data mining; personally identifiable information; HDFS; permanent repository; basic statistics; search/query/indexing	First group of the former sixth cluster
6	10, 11, 26	Link data from different sources	data cleaning; batch-processing; unstructured data; permanent data; dynamic data; HDFS; NoSQL; basic statistics; classification;	Second group of the former sixth cluster
7	18, 31, 38	Enable decision-making	real-time processing; data fusion; unstructured data; pre-processing; text; images; personally identifiable information; Permanent data; basic statistics; search/query/indexing; data mining; classification	First group of the former seventh cluster
8	12, 20, 24, 32, 36, 41	Enable real-time analysis for data, incoming with high-speed	real-time processing; data fusion; visualization; structured and unstructured data; personally identifiable information; data shared between users; permanent data; invaluable data; search/query/indexing; classification; basic statistics	Second group of the former seventh cluster
9	17, 22, 25	Optimize existing processes	high-speed data; both real-time and batch processing; structured and unstructured data; data mining (recommenders); visualization techniques; data fusion	Manually built cluster
Outliers	16, 33, 34, 40			

1) STANDARD USE CASE 1 – DATA ANALYSIS IMPROVEMENT

By adopting big data technologies, an improvement in the quality of the data analysis is pursued. A significant step to achieve this aim is to make sense of massive amounts of unstructured data that are coming with high-speed, and the exploitation of sophisticated methods, such as deep learning. Additionally to that, statistics and classification methods are often used to increase the quality of the analysis. The described characteristics of this general case and the used methods can be mapped to different cases, coming from healthcare, transportation, manufacturing areas, and social media. Details of the particular cases can be viewed in [38], [41], [42], [48], [61], [63], [71], [85].

2) STANDARD USE CASE 2 – BATCH MODE SENSOR DATA ANALYSIS

One of the reasons for harnessing big data technologies is to enable the processing of large amounts of (IoT) sensor data to obtain new insights. Key factors in this use case are the integration of different data sources, such as sensors and devices, as well as enabling the data exchange between users and applications. The data commonly does not exist in a structured format, thus processing unstructured data plays an important role. Real-time processing is not required, as the data is firstly gathered and then processed in batch mode. To uncover different types of patterns, clustering approaches are used for the analysis. The visualization of the processed data is crucial to represent the findings. Based on those, strategies, for instance, to improve the user experience, resource allocation, process costs, and others, can be developed. Concrete specifications for this standard use case are explained in [36], [43], [49].

3) STANDARD USE CASE 3 – SMART CITY

This category deals with the challenges of smart cities by involving various resources in real-time data analysis. The concept itself utilizes data from various devices, sensors, and human actors to improve the quality of life for citizens. For this purpose, structured, unstructured as well as transient and permanent data can be used as analysis input. In order to turn a large amount of heterogeneous data into value, deep learning algorithms are used. In this case, a robust storage solution for massive amounts of differently structured

data should be used, such as a NoSQL database. Due to the nature of this domain, personal information have to be recognized and privacy-preserving techniques are applied. All related cases are comprehensively described in [44], [47], [56], [59], [60], [66].

4) STANDARD USE CASE 4 – MULTI-LEVEL PROBLEMS

In this standard use case, sophisticated multi-level problems are stated, which require thorough planning from different perspectives, covering not only the needed system but also the data being processed. Organizations facing those problems are confronted, in particular, with the growing amount of data coming from various institutions, such as in the healthcare sector. Apart from the required high reliability of the targeted solution and the needed ability to efficiently search, query and store the data, also privacy-preserving techniques have to be considered. Moreover, processing unstructured data, such as handwritten documents or images, needs to be enabled. For the analysis of the data, different data mining approaches, which analyze stored data (e.g. on an HDFS) in batch mode, can be considered. This standard use case originates from the following contributions [46], [52], [74].

5) STANDARD USE CASE 5 – EXPAND DATA SOURCING

In this case, data coming from various resources needs to be combined into one functioning system. As the considered data originates from different sources or instances, not only the structure but also the data itself can be highly volatile. Due to this reason, not only sophisticated storage solutions for those various types of data (e.g. NoSQL), but also sophisticated pre-processing techniques are needed. After the initial collection and cleaning, various statistical methods can be used. The data is usally processed in batch-mode. Concrete details of all relevant use cases can be found in [33], [40], [45], [64].

6) STANDARD USE CASE 6 – DATA CONNECTION

Adopting big data technologies in areas with widespread collections of information can improve decision-making by incorporating a larger information basis. As wrong decisions, especially in domains like healthcare, can have enormous consequences, guaranteeing the correctness of the analyzed data is a significant step, necessitating extensive pre-processing. Depending on the application area, this can additionally require special processing steps like anonymization or classification. For the analysis, data mining techniques can be used and also efficient querying and searching over the data in real-time should be enabled. Further information are provided in [37], [65], [73].

7) STANDARD USE CASE 7 – DECISION SUPPORT

Real-time analytics on differently structured data are used in those use cases, to facilitate decision support for data-driven problems. Through basic statistics, classifications and other analytical methods, previously unused ata are converted into valuable information. For a better presentation of the obtained results, visualization techniques are highly important. This use case can be characterized by the phrase *turn volume into value*. Details can be observed in [55], [62], [68].

8) STANDARD USE CASE 8 – HIGH-SPEED ANALYSIS

Within this use case, the input data comes in a structured and unstructured format and needs to be processed in (near-) real-time, to ensure that all functionalities and results can be immediately provided. In addition, to maintain, search, query, index and analyze all data, complex solutions are required. For a comprehensible representation of the results and the performed calculations, visualization techniques are paramount. For particular insights, the following contributions can be used [39], [50], [57], [69], [72], [76].

9) STANDARD USE CASE 9 – PROCESS OPTIMIZATION

Big data technologies turned out to be an enabler for the general optimization of existing processes. Usually, the data is incoming with high velocity and needs to be processed in real-time. However, also batch-processing mode should be available either as a backup solution or for specific analytical tasks. In this case, both, structured and unstructured data are considered. Clustering techniques support the identification of recommendations with which existing processes can be optimized. Various visualization techniques allow for a better presentation in an appealing way. Further details are described in [54], [58], [67].

IV. EVALUATION

In order to check the validity of the artifact and, thus, the proposed standard use cases, a thorough evaluation is required at which multiple aspects are verified [27], [28]. On the one hand, it is necessary to assure a sufficient coverage of the regarded application domains, methods as well as data characteristics and on the other hand, the undertaken trade-off between possible degrees of fragmentation has to be looked at (cluster building). To assess the coverage with a practical orientation, an approach that is inspired by machine learning's division into training and test data is utilized. For this purpose, the steps of the literature review are replicated while applying the same criteria (cf. Table 1), but this time only for selected case studies published in 2019.

Apart from the search procedure, also the comprehensiveness check through the use of the altered template was performed. In total three additional use cases were found and used for the evaluation of the found out results. Since those cases were not involved in the creation of the standard use cases, they function as the equivalent of a test data set. The first case study used for the evaluation comes from the area of online retail [86]. It provides an approach for a recommendation system that can be realized in an online store, requiring a user to sign up. Besides harnessing historical and transactional data, it also makes use of the customer's browsing history. Taking a look at Table 7 and considering that the case study involves both structured as well as unstructured data and requires real-time processing, regarding those

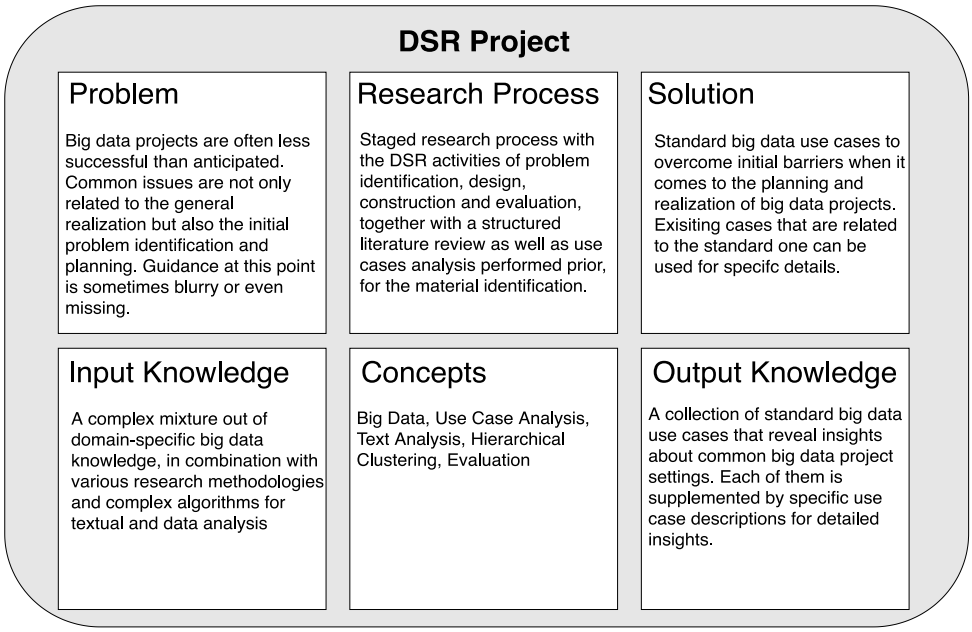

FIGURE 4. The DSR Grid according to (Vom Brocke and Maedche, 2019).

features, this one can be placed in the fourth, eighth, and ninth cluster. However, the used recommender engine in this case study has a key role in the data analysis, which reduces the categorization possibilities only to the ninth cluster.

In conclusion, the analyzed case study aims to improve the online retail by introducing personal recommendations, based on real-time processed browsing data, fitting the ninth general use case that has the target to optimize existing processes with the deployment of big data technologies. The second case study presents a system for the incorperation of real-time social media data in the analysis in the area of tourism [87]. The analysis of the data comprises the following main steps – gathering of the data, cleaning and storage, querying and filtering as well as the visualization of the results. The data is collected from different social media sources in this case – Instagram, Flickr, Foursquare and Twitter, resulting in an unstructured content format that can manifest in the form of posts, reviews, images or videos. Utilizing Table 7 and considering the case study's aim to involve real-time social media data in the analysis, this one can be placed in the eighth cluster, which targets real-time analysis of data, incoming with high-speed. The last case study, harnessed to evaluate the identified use cases, originates from the area of smart transportation [88]. Compared to the already existing approaches, which deal with single issues like congestion avoidance or environmental-friendly driving, the considered case study shows a system that proposes a solution to multiple problems. It aims to track vehicles, suggest optimal routes and realize a smart parking concept, utilizing predominantly unstructured data from various sources like sensors, cars or navigation systems. With regard to Table 7, this example fits into the third general use case.

In conclusion, the successful categorization of the three evaluation case studies in one of the defined general use cases suggests that adequate coverage was achieved. The degree of fragmentation, in turn, is based on the intended application scenario. While a more general approach might increase the coverage even further, it offers no clear orientation in the selection of potentially similar case studies. Vice versa, every case as its own category would effectively negate the idea of a categorization. For that reason, the current number constitutes a trade-off that allows for a choice of relevant properties, while still providing several example cases as a knowledge base for the aspired endeavor.

V. CONCLUSION

In recent years, big data has been one of the most prominent topics in the IT-sector. However, there is still a lot of unawareness and uncertainty when it comes to the execution of such projects, especially right at the beginning of the planning phase. Hence, in the contribution at hand, an in-depth investigation of successfully conducted projects was performed, to provide future practitioners as well as other researchers, inter alia, with decision support concerning the realization of their potential big data projects. As a result of a literature review, 43 cases published between 2015 and 2018 were identified. Those cover detailed information about the presented big data projects. To achieve a categorization for the obtained results, all use case descriptions were thoroughly examined using a textual analysis technique. At this stage, the hierarchical clustering proofed to be a promising solution, revealing various clusters with a similar feature set. Based on the gathered information and further modification, a total of nine distinct clusters were identified.

TABLE 8. Amended and adapted use case template.

No	Description	A	R	N	Made Changes	Type
1.1	Use Case Title	X			-	Open-Ended
1.2	Use Case Description	X			-	Open-Ended
1.3	Use Case Contacts	X			-	Open-Ended
1.4	Domain	X			Renamed to „Application Area“	Open-Ended
1.5	Application		X		-	-
1.5	Advantages of Harnessing Big Data			X	-	Open-Ended
1.6	Current Data Analysis Approach	X			Renamed to „Current Approach“	Open-Ended
1.7	Future of Application and Approach	X			Renamed to „Future Approach“	Open-Ended
1.8	Actors / Stakeholders	X			-	Open-Ended
1.9	Project Goals or Objectives	X			-	Open-Ended
1.10	Use Case URL	X			-	Open-Ended
1.11	Pictures and Diagrams	X			-	Open-Ended
2.1	Data Source	X			-	Open-Ended
2.2	Data Destination	X			-	Open-Ended
2.3	Volume	X			-	Open-Ended
2.4	Velocity	X			-	Open-Ended
2.5	Variety	X			-	Open-Ended
2.6	Variability	X			-	Open-Ended
2.7	Veracity			X	-	Open-Ended
2.8	Variability			X	-	Open-Ended
2.9	Value			X	-	Open-Ended
3.1	Veracity and Data Quality		X		-	-
3.2	Visualization		X		-	-
3.1	Content Format			X	-	Open-Ended
3.3	Data Types	X			Became „3.2. Types of Data“	Open-Ended
3.3	Data Processing			X		Open-Ended
3.4	Metadata		X			-
3.5	Curation and Governance		X		-	-
3.6	Data Analytics	X			-	Open-Ended
4.1	Classified Data, Code or Protocols	X			Became „3.4 Data Analytics“	Closed-Ended
4.2	Does the system maintain personally identifiable information?	X			-	Closed-Ended
4.3	Does the system maintain commercial secrets?			X	-	Closed-Ended
4.3	Publication rights	X			-	Closed-Ended
4.4	Is there an explicit governance plan or framework for the effort?	X			Became “4.4 Publication rights”	Closed-Ended
4.5	Dou you foresee any potential risks from public or private open data projects?		X		Became “4.5 Is there an explicit governance plan or framework for the effort?”	-
4.6	Current Audit Needs		X		-	-
4.7	Under what conditions do you give people access to your data?		X		-	-
4.8	Under what conditions do you give people access to your software?		X		-	-
4.6	List Security, Compliance, Regulatory Requirements			X	-	Open-Ended

TABLE 8. *(Continued.)* **Amended and adapted use case template.**

4.7	Data Governance Issues (List Responsibilities, Steps to be Conducted etc.)		X	-	Open-Ended
4.8	Privacy-preserving Practices		X	-	Open-Ended
4.9	Other Data Privacy and Security Issues		X	-	Closed-Ended
5.1	Data: Application Style and Data Sharing and Acquisition	X		"IoT concept realized" was added as point in this subpart	Multiple Choice
5.2	Data: Management and Storage	X		Further specification of NoSQL databases - document-oriented, columnar stores, graph database or key-values stores, required	Multiple Choice
5.3	Data: Describe other Data Acquisition, Access, Sharing, Management, Storage Issues	X			Open-Ended
5.4	Analytics: Data Format and Nature of Algorithm Used in Analytics	X		"Data mining techniques" tag added	Multiple Choice
5.5	Data Analytics: Describe other Data Analytics Used	X			Open-Ended
5.6	Programming Model	X		"Apache Hive", "Apache Mahout" and "Apache Kafka" added as further tags	Multiple Choice
5.7	Other Programming Model Tags	X			Open-Ended
5.8	Estimate Ratio I/O bytes/flops	X			Open-Ended
5.9	Describe Memory Size or Access Issues	X			Open-Ended
6.1	Other Big Data Issues	X		-	Open-Ended
6.2	User Interface and Mobile Access Issues	X		-	Open-Ended
6.3	List Key Features and Related Use Cases	X		-	Open-Ended
7	Completely		X		-
8	Completely		X		-

Subsequently, those standard use cases constitute the artifact of the conducted DSR endeavor as well as the answer to the research question. To summarize and highlight the main pillars, implications and key aspects of this research, in the following, the corresponding DSR grid according to vom Brocke and Maedche [89] is depicted in Figure 4. One part of the contribution is constituted by the collection, structuring and presentation of comprehensively described use cases published in recent years, in the academic area. Additionally, a template was used and modified for the analysis of the identified cases. Through the use of this template, the general comprehensiveness of one's endeavor can be validated and possible shortcomings or unrecognized gaps identified. Beyond that, a presentation of standard use cases, derived from the investigated publications, is made that can serve as an orientation and initial starting point for the realization of related projects. Consequently, researchers as well as practitioners may greatly benefit from the results discussed in this work.

A. LIMITATIONS AND FUTURE RESEARCH

Although a suitable answer to the initially formulated RQ was achieved, certain aspects have to be mentioned, which may call for future optimization or new research directions. This refers not only to the results as such but also to the methods used to achieve them. For instance, this includes the recognition of additional weightings during the analysis of the input matrix, since sometimes a particular feature appears to be more important than another one. An example for this are features that are directly related to the data characteristics or methods used to analyze them.

During the qualitative analysis of the use cases it was noticed that many of the project descriptions also contained concrete specifications. However, most of the decisions are tailor-made. Hence, an additional investigation of them may, in turn, result in an even more complicated analysis of the data. For now, implementation details can be viewed in each of the referred cases within the standard use cases. Additionally, the chosen algorithm represents only one suitable way for the creation of the found out clusters. Apart from the discussion of the available algorithms and their potential usability, also other algorithms were alternatively tested, especially with a view on future enlargements of the dataset, for which the manual processing of each case would require too much effort. Hence, for the creation of the input matrix, a computer-supported solution was tested that automatically processes the data and identifies important phrases on the base of the term frequency. Despite a thorough pre-processing procedure, which focused on the cleaning of unnecessary stop words, endings, and inconsistent descriptions, no promising results were found. Even after an additional filtering step of the found out phrases, the clusters had too many dissimilarities. This was not only assessed on the qualitative level but also due to multiple irrelevant phrases, such

as "the authors" or "it has". Nevertheless, as already shown before, this has not achieved the desired effect.

Another tested method was the use of natural language processing to uncover existing topics in a collection of unprocessed textual documents [90]. In particular, the topic modeling approach LDA was examined, which is a probabilistic model that considers each topic as a combination of keywords and each document as a combination of multiple topics. Even though comprehensive pre-processing steps, such as lemmatization, stop words and punctuation removal, were repeatedly performed [90], no satisfying results were found. Frequently, it was noticed that buzzwords, especially used in the introduction and conclusion of the papers, have been identified.

The evaluation of the coverage of the formulated cases had a positive result, however, the sample size was rather small and especially future big data projects might potentially necessitate adjustments. Consequently, the sample size of the found big data projects could be enlarged. At this point, an extension of the actual dataset could be realized through the investigation of additional years, further literature databases and also by interviewing larger companies, that conduct big data-related projects. Beyond that, also a long-term evaluation is planned, which shall be realized through the application at the very beginning of a project. Here, not only the general meaningfulness but also the possibility to derive concrete information and implications from the individual use cases could be tested.

By referring to this, an implementation of the derived standard use cases within a concrete decision support system for big data projects, for instance conceptually described in [23], appears to be a promising direction for future research.

APPENDIX

See Table 8.

REFERENCES

[1] C. Dobre and F. Xhafa, "Intelligent services for big data science," *Future Gener. Comput. Syst.*, vol. 37, pp. 267–281, Jul. 2014, doi: 10.1016/j.future.2013.07.014.

[2] S. Yin and O. Kaynak, "Big data for modern industry: Challenges and trends [point of view]," *Proc. IEEE*, vol. 103, no. 2, pp. 143–146, Feb. 2015, doi: 10.1109/JPROC.2015.2388958.

[3] X. Jin, B. W. Wah, X. Cheng, and Y. Wang, "Significance and challenges of big data research," *Big Data Res.*, vol. 2, no. 2, pp. 59–64, Jun. 2015, doi: 10.1016/j.bdr.2015.01.006.

[4] L. Zhu, F. R. Yu, Y. Wang, B. Ning, and T. Tang, "Big data analytics in intelligent transportation systems: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 1, pp. 383–398, Jan. 2019, doi: 10.1109/TITS.2018.2815678.

[5] W. L. Chang and N. Grady, *NIST Big Data Interoperability Framework—Definitions*. Gaithersburg, MD, USA: NIST, 2019. Accessed: Jul. 14, 2020, doi: 10.6028/NIST.SP.1500-1r2

[6] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *Int. J. Inf. Manage.*, vol. 35, no. 2, pp. 137–144, Apr. 2015, doi: 10.1016/j.ijinfomgt.2014.10.007.

[7] S. Kaisler, F. Armour, J. A. Espinosa, and W. Money, "Big data: Issues and challenges moving forward," in *Proc. 46th Hawaii Int. Conf. Syst. Sci.*, Jan. 2013, pp. 995–1004.

[8] D. Izadi, J. Abawajy, S. Ghanavati, and T. Herawan, "A data fusion method in wireless sensor networks," *Sensors*, vol. 15, no. 2, pp. 2964–2979, Jan. 2015, doi: 10.3390/s150202964.

[9] H. Lee, N. Aydin, Y. Choi, S. Lekhavat, and Z. Irani, "A decision support system for vessel speed decision in maritime logistics using weather archive big data," *Comput. Oper. Res.*, vol. 98, pp. 330–342, Oct. 2018, doi: 10.1016/j.cor.2017.06.005.

[10] A. P. Plageras, K. Psannis, C. Stergiou, H. Wang, and B. B. Gupta. (2018). *Efficient IoT-Based Sensor BIG Data Collection-Processing and Analysis in Smart Buildings*. [Online]. Available: https://www.semanticscholar.org/paper/Efficient-IoT-based-sensor-BIG-Data-and-analysis-in-Plageras-Psannis/fb18e87bdfa27b3bc7a9d9337f02cd6b66d0c372

[11] K. E. Psannis, C. Stergiou, and B. B. Gupta, "Advanced media-based smart big data on intelligent cloud systems," *IEEE Trans. Sustain. Comput.*, vol. 4, no. 1, pp. 77–87, Jan. 2019, doi: 10.1109/TSUSC.2018.2817043.

[12] Y. Wang, L. Kung, W. Y. C. Wang, and C. Cegielski, "Developing a big data-enabled transformation model in healthcare: A practice based view," in *Proc. 25th Int. Conf. Inf. Syst.*, 2014, pp. 1–12.

[13] P. Aversa, L. Cabantous, and S. Haefliger, "When decision support systems fail: Insights for strategic information systems from formula 1," *J. Strategic Inf. Syst.*, vol. 27, no. 3, pp. 221–236, Sep. 2018, doi: 10.1016/j.jsis.2018.03.002.

[14] R. Häusler, D. Staegemann, M. Volk, S. Bosse, C. Bekel, and K. Turowski, "Generating content-compliant training data in big data education," in *Proc. 12th Int. Conf. Comput. Supported Edu.*, 2020, pp. 104–110.

[15] T. Nguyen, L. Zhou, V. Spiegler, P. Ieromonachou, and Y. Lin, "Big data analytics in supply chain management: A state-of-the-art literature review," *Comput. Oper. Res.*, vol. 98, pp. 254–264, Oct. 2018, doi: 10.1016/j.cor.2017.07.004.

[16] O. Müller, M. Fay, and J. vom Brocke, "The effect of big data and analytics on firm performance: An econometric analysis considering industry characteristics," *J. Manage. Inf. Syst.*, vol. 35, no. 2, pp. 488–509, Apr. 2018, doi: 10.1080/07421222.2018.1451955.

[17] S. F. Wamba, S. Akter, A. Edwards, G. Chopin, and D. Gnanzou, "How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study," *Int. J. Prod. Econ.*, vol. 165, pp. 234–246, Jul. 2015, doi: 10.1016/j.ijpe.2014.12.031.

[18] Z. A. Al-Sai, R. Abdullah, and M. H. Husin, "Critical success factors for big data: A systematic literature review," *IEEE Access*, vol. 8, pp. 118940–118956, 2020, doi: 10.1109/ACCESS.2020.3005461.

[19] D. Staegemann, M. Volk, A. Nahhas, M. Abdallah, and K. Turowski, "Exploring the specificities and challenges of testing big data systems," in *Proc. 15th Int. Conf. Signal-Image Technol. Internet-Based Syst. (SITIS)*, Nov. 2019, pp. 289–295.

[20] D. Staegemann, M. Volk, N. Jamous, and K. Turowski, "Understanding issues in big data applications—A multidimensional endeavor," in *Proc. 25th Amer. Conf. Inf. Syst.*, 2019, pp. 1–10.

[21] S. Sagiroglu and D. Sinanc, "Big data: A review," in *Proc. Int. Conf. Collaboration Technol. Syst. (CTS)*, May 2013, pp. 42–47.

[22] S. Bonesso, E. Bruni, and F. Gerli, "How big data creates new job opportunities: Skill profiles of emerging professional roles," in *Behavioral Competencies of Digital Professionals: Understanding the Role of Emotional Intelligence*, S. Bonesso, E. Bruni, F. Gerli, Eds. Cham, Switzerland: Palgrave Macmillan, 2020, pp. 21–39.

[23] M. Volk, D. Staegemann, M. Pohl, and K. Turowski, "Challenging big data engineering: Positioning of current and future development," in *Proc. 4th Int. Conf. Internet Things, Big Data Secur.*, 2019, pp. 351–358.

[24] M. Volk, D. Staegemann, S. Bosse, A. Nahhas, and K. Turowski, "Towards a decision support system for big data projects," in *WI2020 Zentrale Tracks*, N. Gronau, M. Heine, K. Poustcchi, and H. Krasnova, Eds. Berlin, Germany: GITO Verlag, 2019, pp. 357–368.

[25] R. Dontha. *The Origins of Big Data—KDnuggets*. Accessed: Jun. 16, 2020. [Online]. Available: https://www.kdnuggets.com/2017/02/origins-big-data.html

[26] O. Ylijoki and J. Porras, "Conceptualizing big data: Analysis of case studies," *Intell. Syst. Accounting, Finance Manage.*, vol. 23, no. 4, pp. 295–310, Oct. 2016, doi: 10.1002/isaf.1393.

[27] R. H. Von Alan, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quart.*, vol. 28, no. 1, pp. 75–105, 2004.

[28] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *J. Manage. Inf. Syst.*, vol. 24, no. 3, pp. 45–77, Dec. 2007.

[29] J. Webster and R. T. Watson, "Analyzing the Past to Prepare for the Future: Writing a Literature Review," *MIS Quart.*, vol. 26, no. 2, pp. 13–23, 2002. [Online]. Available: http://www.jstor.org/stable/4132319

[30] Y. Levy and T. J. Ellis, "A systems approach to conduct an effective literature review in support of information systems research," *Informing Sci., Int. J. Emerg. Transdiscipline*, vol. 9, pp. 181–212, Jan. 2006, doi: 10.28945/479.

[31] R. Bauer. *3 Reasons Why You Need Business Case Studies—PAN Communications*. Accessed: Mar. 18, 2020. [Online]. Available: https://www.pancommunications.com/blog/3-reasons-why-you-need-business-case-studies/

[32] R. K. Yin, *Case Study Research and Applications: Design and Methods*. Los Angeles, CA, USA: SAGE, 2018.

[33] F. Khalid. (2017). *Inovalation is Driving Healthcare Transformation With Pre-Engineered Infrastructure and Big Data Analytics*. Dell EMC, Bowie, MD, USA. Accessed: Jun. 1, 2020. [Online]. Available: https://www.emc.com/collateral/customer-profiles/inovalon-vscale-case-study.pdf

[34] W. L. Chang and G. Fox. (2018). *NIST Big Data Interoperability Framework—Use Cases and General Requirements*. Gaithersburg, MD, USA. Accessed: Jul. 14, 2020. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-3r1.pdf

[35] J. Vom Brocke, A. Simons, B. Niehaves, K. Reimer, R. Plattfaut, and A. Cleven, "Reconstructing the giant: On the importance of rigour in documenting the literature search process," in *Proc. ECIS*, Verona, Italy, 2009, pp. 1–14.

[36] A. Yassine, S. Singh, M. S. Hossain, and G. Muhammad, "IoT big data analytics for smart homes with fog and cloud computing," *Future Gener. Comput. Syst.*, vol. 91, pp. 563–573, Feb. 2019, doi: 10.1016/j.future.2018.08.040.

[37] D. Kenyon and J. H. P. Eloff, "Big data science for predicting insurance claims fraud," in *Proc. Inf. Secur. South Afr. (ISSA)*, Aug. 2017, pp. 40–47.

[38] Y. Zhang, M. Zhang, T. Wo, X. Lin, R. Yang, and J. Xu, "A scalable Internet-of-Vehicles service over joint clouds," in *Proc. IEEE Symp. Service-Oriented Syst. Eng. (SOSE)*, Mar. 2018, pp. 210–215.

[39] H.-M. Chen, R. Schütz, R. Kazman, and F. Matthes, "How Lufthansa capitalized on big data for business model renovation," *MIS Quart. Executive*, vol. 16, no. 1, p. 4, 2017.

[40] M. Schlesner and F. Schinkel. (2016). *Big Data Use Case: Genomic Data Research*. Fujitsu, Munich, Germany. Accessed: Sep. 4, 2019. [Online]. Available: https://www.datameer.com/wp-content/uploads/pdf/misc/cs-PF4H-Genome-Research.pdf

[41] M. Avvenuti, S. Cresci, F. Del Vigna, T. Fagni, and M. Tesconi, "CrisMap: A big data crisis mapping system based on damage detection and geoparsing," *Inf. Syst. Frontiers*, vol. 20, no. 5, pp. 993–1011, Oct. 2018, doi: 10.1007/s10796-018-9833-z.

[42] M. K. Hassan, A. I. El Desouky, S. M. Elghamrawy, and A. M. Sarhan, "Intelligent hybrid remote patient-monitoring model with cloud-based framework for knowledge discovery," *Comput. Electr. Eng.*, vol. 70, pp. 1034–1048, Aug. 2018, doi: 10.1016/j.compeleceng.2018.02.032.

[43] F. Gu, B. Ma, J. Guo, P. A. Summers, and P. Hall, "Internet of Things and big data as potential solutions to the problems in waste electrical and electronic equipment management: An exploratory study," *Waste Manage.*, vol. 68, pp. 434–448, Oct. 2017, doi: 10.1016/j.wasman.2017.07.037.

[44] Y. Arfat, M. Aqib, R. Mehmood, A. Albeshri, I. Katib, N. Albogami, and A. Alzahrani, "Enabling smarter societies through mobile big data fogs and clouds," *Procedia Comput. Sci.*, vol. 109, pp. 1128–1133, Jan. 2017, doi: 10.1016/j.procs.2017.05.439.

[45] S. Istephan and M.-R. Siadat, "Unstructured medical image query using big data—An epilepsy case study," *J. Biomed. Informat.*, vol. 59, pp. 218–226, Feb. 2016, doi: 10.1016/j.jbi.2015.12.005.

[46] D. Mourtzis, E. Vlachou, and N. Milas, "Industrial big data as a result of IoT adoption in manufacturing," *Procedia CIRP*, vol. 55, pp. 290–295, Jan. 2016, doi: 10.1016/j.procir.2016.07.038.

[47] S. Rinaldi, A. Flammini, M. Pasetti, L. C. Tagliabue, A. C. Ciribini, and S. Zanoni, "Metrological issues in the integration of heterogeneous lot devices for energy efficiency in cognitive buildings," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (I2MTC)*, May 2018, pp. 1–6.

[48] P. Ta-Shma, A. Akbar, G. Gerson-Golan, G. Hadash, F. Carrez, and K. Moessner, "An ingestion and analytics architecture for IoT applied to smart city use cases," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 765–774, Apr. 2018, doi: 10.1109/JIOT.2017.2722378.

[49] A. C. Onal, O. Berat Sezer, M. Ozbayoglu, and E. Dogdu, "Weather data analysis and sensor fault detection using an extended IoT framework with semantics, big data, and machine learning," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 2037–2046.

[50] S. Balaji, M. Patil, and C. McGregor, "A cloud based big data based online health analytics for rural NICUs and PICUs in india: Opportunities and challenges," in *Proc. IEEE 30th Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jun. 2017, pp. 385–390.

[51] L. Carnevale, A. Celesti, M. Fazio, P. Bramanti, and M. Villari, "How to enable clinical workflows to integrate big healthcare data," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Jul. 2017, pp. 857–862.

[52] S. Amini, I. Gerostathopoulos, and C. Prehofer, "Big data analytics architecture for real-time traffic control," in *Proc. 5th IEEE Int. Conf. Models Technol. Intell. Transp. Syst. (MT-ITS)*, Jun. 2017, pp. 710–715.

[53] Herrnansyah, Y. Ruldeviyani, and R. F. Aji, "Enhancing query performance of library information systems using NoSQL DBMS: Case study on library information systems of universitas indonesia," in *Proc. Int. Workshop Big Data Inf. Secur. (IWBIS)*, Oct. 2016, pp. 41–46.

[54] I. Azimi, A. Anzanpour, A. M. Rahmani, P. Liljeberg, and T. Salakoski, "Medical warning system based on Internet of Things using fog computing," in *Proc. Int. Workshop Big Data Inf. Secur. (IWBIS)*, Oct. 2016, pp. 19–24.

[55] P.-Y. Wu, C.-W. Cheng, C. D. Kaddi, J. Venugopalan, R. Hoffman, and M. D. Wang, "-Omic and electronic health record big data analytics for precision medicine," *IEEE Trans. Bio-Med. Eng.*, vol. 64, no. 2, pp. 263–273, Feb. 2017, doi: 10.1109/TBME.2016.2573285.

[56] E. Patti and A. Acquaviva, "IoT platform for smart cities: Requirements and implementation case studies," in *Proc. IEEE 2nd Int. Forum Res. Technol. Soc. Ind. Leveraging Better Tomorrow (RTSI)*, Sep. 2016, pp. 1–6.

[57] W. Seo, N. Kim, and S. Choi, "Big data framework for analyzing patents to support strategic R&D planning," in *Auckland*. IEEE, 2016, pp. 746–753. [Online]. Available: https://ieeexplore.ieee.org/document/7588929

[58] V. Dineshreddy and G. R. Gangadharan, "Towards an, 'Internet Things framework for financial services sector," in *Proc. 3rd Int. Conf. Recent Adv. Inf. Technol. (RAIT)*, Dhanbad, India, Mar. 2016, pp. 177–181.

[59] Y. Sun, H. Song, A. J. Jara, and R. Bie, "Internet of Things and big data analytics for smart and connected communities," *IEEE Access*, vol. 4, pp. 766–773, 2016, doi: 10.1109/ACCESS.2016.2529723.

[60] S. Pirttikangas, E. Gilman, X. Su, T. Leppanen, A. Keskinarkaus, M. Rautiainen, M. Pyykkonen, and J. Riekki, "Experiences with smart city traffic pilot," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2016, pp. 1346–1352.

[61] A. Aguilera, R. Grunzke, U. Markwardt, D. Habich, D. Schollbach, and J. Garcke, "Towards an industry data gateway: An integrated platform for the analysis of wind turbine data," in *Proc. 7th Int. Workshop Sci. Gateways*, Jun. 2015, pp. 62–66.

[62] A. Abusharekh, S. A. Stewart, N. Hashemian, and S. S. R. Abidi, "H-DRIVE: A big health data analytics platform for evidence-informed decision making," in *Proc. IEEE Int. Congr. Big Data*, Jun. 2015, pp. 416–423.

[63] M. Panahiazar, V. Taslimitehrani, A. Jadhav, and J. Pathak, "Empowering personalized medicine with big data and semantic Web technology: Promises, challenges, and use cases," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Oct. 2014, pp. 790–795.

[64] A. Sebaa, F. Chikh, A. Nouicer, and A. Tari, "Medical big data warehouse: Architecture and system design, a case study: Improving healthcare resources distribution," *J. Med. Syst.*, vol. 42, no. 4, p. 59, Apr. 2018, doi: 10.1007/s10916-018-0894-9.

[65] J. F. Sánchez-Rada, A. Pascual, E. Conde, and C. A. Iglesias, "A big linked data toolkit for social media analysis and visualization based on W3C Web components," in *On Move to Meaningful Internet Systems*. Valletta, Malta: Springer, 2018, pp. 498–515. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-02671-4_30

[66] A. Smirnov, A. Ponomarev, N. Teslya, and N. Shilov, "Human-computer cloud for smart cities: Tourist itinerary planning case study," in *Business Information Systems Workshops* (Lecture Notes in Business Information Processing), vol. 303, W. Abramowicz, Ed. Cham, Switzerland: Springer, 2017, pp. 179–190.

[67] A. Aguilera, R. Grunzke, D. Habich, J. Luong, D. Schollbach, U. Markwardt, and J. Garcke, "Advancing a gateway infrastructure for wind turbine data analysis," *J. Grid Comput.*, vol. 14, no. 4, pp. 499–514, Dec. 2016, doi: 10.1007/s10723-016-9376-9.

[68] R. S. Santos, T. A. Vaz, R. P. Santos, and J. M. P. de Oliveira, "Big data analytics in a public general hospital," in *Machine Learning, Optimization, and Big Data* (Lecture Notes in Computer Science), vol. 10122, P. M. Pardalos, P. Conca, G. Giuffrida, and G. Nicosia, Eds. Cham, Switzerland: Springer, 2016, pp. 433–441.

[69] H. Khazaei, S. Zareian, R. Veleda, and M. Litoiu, "Sipresk: A big data analytic platform for smart transportation," in *Proc. 1st EAI International Summit, Smart City 360°*. Bratislava, Slovakia: Springer, 2016, pp. 419–430. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-33681-7_35

[70] A. Fiannaca, L. La Paglia, M. La Rosa, A. Messina, P. Storniolo, and A. Urso, "Integrated DB for bioinformatics: A case study on analysis of functional effect of MiRNA SNPs in cancer," in *Proc. Int. Conf. Inf. Technol. Bio Med. Inform.*, Porto, Portugal, Sep. 2016, pp. 214–222.

[71] A. Majumdar and I. Bose, "Detection of financial rumors using big data analytics: The case of the bombay stock exchange," *J. Organizational Comput. Electron. Commerce*, vol. 28, no. 2, pp. 79–97, Apr. 2018, doi: 10.1080/10919392.2018.1444337.

[72] G. Escobedo, N. Jacome, and G. Arroyo-Figueroa, "Big data & analytics to support the renewable energy integration of smart grids—Case study: Power solar generation," in *Proc. 2nd Int. Conf. Internet Things, Big Data Secur. IoTBDS*, Porto, Portugal, Apr. 2017, pp. 267–275.

[73] Y. Zhuang, Y. Wang, J. Shao, L. Chen, W. Lu, J. Sun, B. Wei, and J. Wu, "D-ocean: An unstructured data management system for data ocean environment," *Frontiers Comput. Sci.*, vol. 10, no. 2, pp. 353–369, Apr. 2016, doi: 10.1007/s11704-015-5045-6.

[74] C.-M. Chen, "Use cases and challenges in telecom big data analytics," *APSIPA Trans. Signal Inf. Process.*, vol. 5, pp. 1–7, Dec. 2016, doi: 10.1017/ATSIP.2016.20.

[75] M. F. Huber, M. Voigt, and A.-C. N. Ngomo, "Big data architecture for the semantic analysis of complex events in manufacturing," in *Informatik*. Bonn, Germany: Gesellschaft fpr Informatik e.V., 2016, pp. 353–360. [Online]. Available: https://dl.gi.de/handle/20.500.12116/1139;jsessionid=D794018779FF36E5A6CBE13273EE9C67

[76] Q. Huang, G. Cervone, D. Jing, and C. Chang, "DisasterMapper," in *Proc. 4th Int. ACM SIGSPATIAL Workshop Anal. Big Geospatial Data BigSpatial*, 2015, pp. 1–6.

[77] M. Xu, S. Siraj, and L. Qi, "A Hadoop-based data processing platform for fresh Agro products traceability," in *Proc. Eur. Conf. Data Mining*, 2015, pp. 37–44. [Online]. Available: http://www.iadisportal.org/components/com_booklibrary/ebooks/201508L005.pdf

[78] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in *Proc. TextMining Workshop KDD*, May 2000, pp. 1–20.

[79] Y. Zhao, G. Karypis, and U. Fayyad, "Hierarchical clustering algorithms for document datasets," *Data Mining Knowl. Discovery*, vol. 10, no. 2, pp. 141–168, Mar. 2005, doi: 10.1007/s10618-005-0361-3.

[80] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, 99th ed. Hoboken, NJ, USA: Wiley, 2009. [Online]. Available: http://gbv.eblib.com/patron/FullRecord.aspx?p=469065

[81] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms*, 2nd ed. Piscataway, NJ, USA: IEEE Press, 2011.

[82] J. Cleve and U. Lämmel. *Data Mining. München: De Gruyter Oldenbourg*. [Online]. Available: https://doi.org/10.1515/9783110456776

[83] P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, *Introduction to Data Mining*. London, U.K.: Pearson, 2019.

[84] MATLAB. *Agglomerative Hierarchical Cluster Tree—MATLAB Linkage—MathWorks*. Accessed: Mar. 27, 2020. [Online]. Available: https://mathworks.com/help/stats/linkage.html?s_tid=mwa_osa_a#d117e514451

[85] A. L. Marra, F. Martinelli, P. Mori, and A. Saracino, "Implementing usage control in Internet of Things: A smart home use case," in *Proc. IEEE Trustcom/BigDataSE/ICESS*, Aug. 2017, pp. 1056–1063.

[86] G. Alfian, M. F. Ijaz, M. Syafrudin, M. A. Syaekhoni, N. L. Fitriyani, and J. Rhee, "Customer behavior analysis using real-time data processing," *Asia Pacific J. Marketing Logistics*, vol. 31, no. 1, pp. 265–290, Jan. 2019, doi: 10.1108/APJML-03-2018-0088.

[87] K. Vassakis, E. Petrakis, I. Kopanakis, J. Makridis, and G. Mastorakis, "Location-based social network data for tourism destinations," in *Big Data and Innovation in Tourism, Travel, and Hospitality: Managerial Approaches, Techniques, and Applications*, M. Sigala, R. Rahimi, and M. Thelwall, Eds. Singapore: Springer, 2019, pp. 105–114.

[88] S. Muthuramalingam, A. Bharathi, S. Rakesh kumar, N. Gayathri, R. Sathiyaraj, and B. Balamurugan, "IoT based intelligent transportation system (IoT-ITS) for global perspective: A case study," in *Internet of Things and Big Data Analytics for Smart Generation*, V. E. Balas, V. K. Solanki, R. Kumar, and M. Khari, Eds. Cham, Switzerland: Springer, 2019, pp. 279–300.

[89] J. vom Brocke and A. Maedche, "The DSR grid: Six core dimensions for effectively planning and communicating design science research projects," *Electron. Markets*, vol. 29, no. 3, pp. 379–385, Sep. 2019, doi: 10.1007/s12525-019-00358-7.

[90] M. J. Zaki, Eds., *On Finding the Natural Number of Topics With Latent Dirichlet Allocation: Some Observations: Advances in Knowledge Discovery and Data Mining*. Berlin, Germany: Springer, 2010.

MATTHIAS VOLK (Graduate Student Member, IEEE) studied business informatics at the Faculty of Computer Science, Otto von Guericke University Magdeburg (OVGU). He received the master's degree in 2016. He is currently pursuing the Ph.D. degree. Since then, he has been employed as a Scientific Researcher. During his studies, he gained lots of practical experience as a software developer in different companies such as Volkswagen. During his scientific career, he participated in many international scientific congresses and projects, not only as a speaker but also as a reviewer or a session chair. His research interests include domain of data-intensive systems, related projects, technologies, and the management of them.

DANIEL STAEGEMANN studied computer science at Technical University Berlin (TUB). He received the master's degree in 2017. He is currently pursuing the Ph.D. degree with the Otto von Guericke University Magdeburg. Since 2018, he has been employed as a Scientific Researcher with OVGU. His research interest includes big data, especially the testing.

IVAYLA TRIFONOVA studied business informatics at the Otto von Guericke University Magdeburg. She received the master's degree in 2019. She is currently working as an IT Consultant in the area of Life Science at a large European consulting and IT services company.

SASCHA BOSSE studied computer science at the Faculty of Computer Science, Otto von Guericke University Magdeburg. He received the master's degree in 2011, and the academic degree Doktoringenieur in 2016. Since 2012, he has been working as a Researcher with the VLBA Lab. Since 2020, he has also been working as a Subject Specialist for computer science and mathematics with University Library Magdeburg, where he is also responsible for the business applications. His research interests include IT service management, modeling, simulation, and optimization.

KLAUS TUROWSKI studied business and engineering at the University of Karlsruhe. He received the Ph.D. degree from the Institute for Business Informatics, University of Münster, and the habilitated degree in business informatics from the Faculty of Computer Science, Otto von Guericke University Magdeburg. In 2000, he deputized the Chair of business informatics at the University of the Federal Armed Forces München. Since 2001, he has been heading the Chair of business informatics and systems engineering with the University of Augsburg. Since 2011, he has also been heading the Chair of business informatics (AG WI) with the Otto von Guericke University Magdeburg, the Very Large Business Applications Lab (VLBA Lab), and the world's largest SAP University Competence Center (SAP UCC Magdeburg). Additionally, he worked as a guest lecturer at several universities around the world. He was a Lecturer with the Universities of Darmstadt and Konstanz. He was a (co-)organizer of a multiplicity of national and international scientific congresses and workshops (>30) and acted as a member of several programme commitees (>130), and expert Groups. In the context of his university activities as well as an independent consultant he gained practical experience in industry.