Knowledge and Information Systems (2023) 65:3165–3196
https://doi.org/10.1007/s10115-023-01870-1

REVIEW

Data management in digital twins: a systematic literature review

Jaqueline B. Correia[1] · **Mara Abel**[1] · **Karin Becker**[1]

Received: 24 August 2022 / Revised: 28 December 2022 / Accepted: 24 March 2023 /
Published online: 16 April 2023

Abstract
The Internet of Things (IoT) and continuous advances in data-gathering devices and techniques have significantly increased the amount of relevant data that can be leveraged for innovative real-time, data-driven applications. Digital Twins (DTs) are virtual representations of physical objects, which are fully integrated and in which the automatic data exchange occurs in a bidirectional way. Modern DTs follow a five-component architecture, which includes an explicit Data Management (DM) component that acts as a bridge between the other systems. However, there is no clarity on its role and functionalities. This article presents a Systematic Literature Review on DM solutions proposed in the DT context. We analyzed DM under the Big Data chain of activities to add value to data, highlighting key issues to be addressed: data heterogeneity, interoperability, integration, data search, and quality. In addition to surveying existing solutions for handling these issues, we contextualized them in the domain and function for which the DT was proposed, the type of data dealt with, and the technological infrastructure. Our main findings revealed that the maturity level assumed for the DM component is at an early stage. The most mature solutions were proposed for the industry domain, and many of them assume humans as the ultimate information consumers. Data integration is the prevalent DM issue addressed due to the bridging role of the DM component, and cloud computing is the key implementation technology. Among the research opportunities are reference data management architectures, adoption of industry standards and ontologies, interoperability among distinct DTs, the development of agnostic standard implementations, and data provenance mechanisms.

Keywords Digital twin · Data management · Big data · Systematic literature review

✉ Jaqueline B. Correia
jbcorreia@inf.ufrgs.br

Mara Abel
marabel@inf.ufrgs.br

Karin Becker
karin.becker@inf.ufrgs.br

[1] Institute of Informatics, Federal University of the Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil

1 Introduction

The increasing popularity of the Internet of Things (IoT), the advent of smart wearables, and the continuous advances in data-gathering techniques have significantly increased the amount of relevant data that can be leveraged for innovative real-time, data-driven applications. By exploiting these devices and various technologies, information about physical reality is seamlessly transferred into the cyber world, where it is elaborated to adapt cyber applications and services to the physical context, thus possibly modifying/adapting the physical world itself through actuators [1]. Digital Twins (DTs) are the next step in this cyber-physical convergence. DTs are virtual representations of physical objects, which are fully integrated and in which the automatic data exchange occurs in a bidirectional way [2].

DTs are at the core of disruptive innovations in diverse areas [3]. In Smart Manufacturing (Industry 4.0 - I4.0), DTs can cover all product life-cycle phases, including design, planning, assembly, and workshop optimization [2, 4]. Companies in the Oil&Gas (O&G) industry leverage innovation to increase production and maximize profit and have successful experiences in smart oilfield and pipelining, predictive maintenance, and risk assessment [5, 6]. DTs are also expected to change the concept of digital healthcare, where a virtual replica of a patient could improve health promotion and control, predict future trends using medical history, and optimize healthcare operations [7]. Smart City DTs aim to improve the efficiency and sustainability of logistics, energy consumption, urban planning, disaster management, among others [8].

Big Data [9] and DTs are mutually reinforcing technologies since huge volumes of data representing the physical/virtual worlds are collected, transformed, and generated through models (e.g., simulation, machine learning) to aggregate value to the business [4, 10]. These opportunities require dealing with data in a volume, velocity, and variety that exceed the capabilities of traditional data management systems, delivering value and veracity. In this context, data is a fundamental resource that needs to be considered in the big data value chain [11], which includes activities for data acquisition, analysis, storage, curation, and usage. Data lakes [12] is a trending topic to address Big Data issues.

Different DT frameworks are proposed in the literature [6]. Earlier DTs follow a three-component architecture that connects a physical system to a mirrored virtual one. While the Physical space represents the physical assets (e.g., sensors, actuators), the Virtual Space aims to emulate the physical environment with high fidelity. DTs following this architecture adopt *ad hoc* solutions for data management issues such as data extraction and integration of heterogeneous sources, data sanity, data transformation and enrichment, and data consumption by the virtual environment. The existence of data silos, the volume of data, and issues related to handling multiple, heterogeneous data sources, formats, and data types are often mentioned as significant challenges [13–16].

The five-component DT architecture [17] is an evolution that explicitly includes a Data Management (DM) component. The DM component acts as a bridge between all subsystems, serving as a point of ingestion of the original data and of return at the right time to direct the interactive optimization process resulting from their interaction. Existing works provide the functionality to manage different aspects of data, such as data cleaning, quality assessment, transformation, integration, search, among others. These data management functionalities are either explicitly encompassed in a dedicated DM component as proposed by [17], or scattered in other components of the DT.

This article presents a Systematic Literature Review (SLR) on the proposed solutions for DM issues within the scope of DTs, which is either implicit within the DT or explicit as part

of a DM component. This SLR was motivated by the absence of a survey or review focused on data management solutions for DTs and a lack of understanding of the role and core functionality of the DM component. Existing surveys contribute to the understanding of the concepts, properties, and primary use cases/applications in DTs [2, 10, 18, 19]. Regarding data management, [3] presents a survey from Manufacturing Automation and Networking Computing perspective, outlining architectural designs based on data-related factors (presence, coordination, and computing). Although [4] highlights an explicit component in the architecture of a DT to handle data management, it does not detail the support it must provide. We argue that the DM component can be approximated to the data and knowledge management functionality in Data Lakes [20, 21], and aim to reach a better comprehension of the state-of-the-art solutions in a fine-grained analysis.

Our SLR presents a novel perspective by considering selected data management issues extracted from the value chain of Big key Data activities [11]. An SLR is defined in [22] as "a systematic, explicit, comprehensive, and reproducible method for identifying, evaluating, and synthesizing the existing body of completed and recorded work produced by researchers, scholars, and practitioners." We surveyed existing works systematically and unbiasedly to shed light on the DM solutions proposed to deal with data heterogeneity, interoperability, integration, data search/discovery, and quality in DTs. We defined the following research questions: RQ1) *For which domains the DT solutions were proposed?*; RQ2) *Under the perspective of data usage, for which functions were the DTs proposed?*; RQ3) *What types of data do the proposed solutions consider?*; RQ4) *What solutions were proposed for the DM issues addressed?*; RQ5) *What kind of technological infrastructure is considered?*.

Our SLR complements and innovates the landscape of existing literature reviews on DTs by investigating DM aspects not yet analyzed, expressed by the research questions. The main contributions of this article are:

- The fine-grained analysis of DM under the activities within the Big Data value chain highlights key issues to be addressed by the DM component in a DT.
- An SLR surveying existing solutions for handling data heterogeneity, interoperability, integration, data search/discovery, and quality in DTs. We contextualized these solutions in the domain and function for which the DT were proposed, the type of data handled, and leveraged technological infrastructure. The compilation of these solutions sheds light on the functionality to be provided by a DM component of a DT, current trends, and opportunities.

The remainder of this article is organized as follows. Section 2 describes the theoretical background of DT. Section 3 summarizes the key DM issues derived from Big Data value chain activities. Section 4 outlines the protocol developed for the SLR, including the motivation for the selected research questions. Section 5 answers each research question defined in the protocol. Section 6 summarizes the main trends and opportunities identified. Finally, Sect. 7 draws conclusions and outlines future work.

2 Theoretical background

There is no consensus on the definition of the term "Digital Twin", and many works contributed to better comprehending this concept. A systematic review [19] compiled 29 different definitions out of 75 collected studies. The authors concluded that all definitions aim to stress specific key points, where the most common ones are virtual/mirror/replica, clone/counterpart, and integrated systems. Another study [18] systematically reviewed 35

works using topic modeling techniques, concluding that definitions are influenced by five aspects that characterize DTs: product life cycle, synchronization of the cyber/physical spaces, integration of real-time data, and behavioral modeling of the physical space and services provided.

According to [23], a DT must meet the following requirements: (a) it is some level of replica of a real thing; (b) it exists in the cyber world (i.e., it is a software entity); (c) it has a purpose of impacting an aspect of the environment in which its real counterpart exists, in a positive way; (d) it uses models to achieve its purpose; (e) it incorporates some level of subject matter expertise in the solution, which could be as simple as defining the problem, or as complex as being an integral part of the model solution; and (f) it uses data to maintain some type of synchronization with its real counterpart, where typically these data are collected in an operational environment. In this SLR, we consider DTs that meet these requirements.

Based on the manual/automatic data flows between Physical and Digital objects, [2] distinguishes between the terms Digital Models, Digital Shadows, and DTs. A *Digital Model* is a digital version of an existing or planned physical object, and no automatic exchange exists between them. A *Digital Shadow* (DS) is a digital representation of an object that has a one-way flow between the physical and the digital object, such that a change in the physical object leads to a change in the digital one, but not vice versa. In a DT, the Digital and Physical objects are fully integrated into both directions, such that a change in one leads to a change in the other. Our SLR encompasses both DTs and DSs, since the closed loop represents a stage of maturity that has not been reached by related work yet, and does affect the required data management functionality for the DM component.

Another systematic study [6] conceptualizes DTs in terms of reference architectures. In this work, we adopt the five-component architecture proposed in [17], depicted in Fig. 1. In addition to the DM central component, the Services component expands the Virtual space with other enterprise software tools (e.g., analytical and predictive resources, visualization, model calibration). The DM component serves as a point of ingestion of the original data from these systems and of return at the right time to direct the interactive optimization process resulting from the interaction among them. To that end, it has to provide different functionalities to handle the collected data and help add value to transform it into knowledge.

Organizations from different domains see benefits in capturing real-time data streams using different types of sensors (e.g., IoT, wearables), making sense of this raw data in terms of business-specific data, and leveraging models to add value that enables right-time decisions that positively impact the physical space. Existing surveys broadly classify existing DTs in domains such as I4.0, smart city, and healthcare [3, 10, 18]. Smart manufacturing is an active subdomain of I4.0, in which DT has been applied to reduce costs and production time and improve quality [4]. The O&G field expects to optimize production, anticipate failures, and produce in a cost-effective way [6, 24]. The concept of DT is key for smart cities [8, 25], as it can help to optimize the planning and management of the services provided by the city, improve the quality of life and public safety, and prevent disasters, among others. DTs will take healthcare to the next level [26], improving the early detection and prevention of diseases, enabling customized treatments, and improving the processes of healthcare providers.

Despite the idiosyncrasies of each field, challenges derived from the volume, variety, veracity, velocity, and value of data management are very similar. The functions of the DM component can be approximated to the data and knowledge management functionality in Data Lakes [20, 21]. In light of Big Data value chain activities [11], we detail in the next section the key DM issues to be handled by the DM component of a DT.

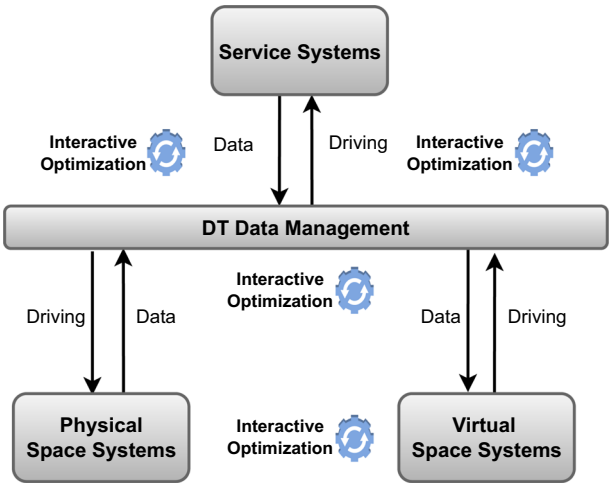

Fig. 1 Five-component DT architecture

3 Data management issues in digital twins

The term "Big Data" is used to label data management according to different attributes. Originally it was associated with three key properties [11]: volume, velocity, and variety. Volume requires dealing with large scales of data within data processing. Velocity involves dealing with high-frequency streams of incoming real-time data (e.g., sensors, IoT). Variety implies handling data using differing syntactic formats (e.g., spreadsheets, XML), structured and unstructured data (e.g., tabular data, texts, videos), schemas, and meanings. As the field matured, other properties were included [27], among them Veracity and Value. Veracity refers to the truthfulness or reliability of the data, while Value is the measurement of data usefulness that determines the discovery of hidden values from the collected data.

The Big Data value chain identifies key-level activities [11]. *Data Acquisition* covers the process of gathering, filtering, cleaning, preparing data, and making it available in some storage solution for further data analysis. *Data Analysis* is concerned with making the acquired data amenable to use in decision-making and domain-specific usage. It involves exploring, transforming, and modeling data to highlight relevant data. *Data Curation* is the active management of data over its life cycle to ensure the necessary data quality for its effective usage, and it is in charge of a data curator expert. *Data Storage* is the persistence and management of data in a scalable way that satisfies the needs of applications that require fast access to the data. *Data Usage* covers the data-driven business activities that need access to data, its analysis, and the tools required to integrate the data analysis within the business activity. In the context of a DT architecture, the functionality of the DM component can be mainly mapped to the activities encompassed in Data Acquisition, Data Analysis, and Data Curation, with the technological support of Data Storage. The Data Usage activities can be mapped into the role of the Services or Virtual Space components in creating value from data.

Data Lake is a concept that emerged to overcome the challenges related to big data scenarios [28]. According to [12], a data lake is a scalable storage and analysis system for data of any type, retained in their native format and used mainly for knowledge extraction. It should support the integration of any type of data; support for logical and physical organization of data; accessibility to various user profiles; metadata catalog to enforce quality; and scalability

in terms of storage and processing. The functions of the DM component can be approximated to the data and knowledge management functionality in Data Lakes [20, 21].

In this paper, we survey the DM solutions proposed in DTs under the perspective of Data Acquisition, Data Analysis, and Data Curation activities, considering data/knowledge management functionality similar to Data Lakes. According to this view, the DM component must address at least the following issues: heterogeneity, interoperability, integration, data discovery/search, and data quality.

The diversity of sources that generate data, such as IoT sensors, information systems, smartphones, social networks, web pages, and wearable devices, is the cause of the data variety and affects the Acquisition activities. Hence the DM component has to deal with *data heterogeneity*. According to [29], there are three levels of heterogeneity: syntactic, terminological, and semantic. Syntactic heterogeneity occurs when two data sources do not use the same formalism to represent the same data (e.g., schema). Terminological heterogeneity occurs when two data sources use different terminology to refer to the same entity. Finally, semantic heterogeneity occurs when there is no consensus of meaning or understanding about a given entity. The heterogeneity directly impacts the acquisition, integration, quality, contextualization of data and, consequently, its value. Obtaining an abstract and unified view of all these different data sets is arduous and complex. Data heterogeneity is a fundamental challenge in any application context that needs to deal with different sources that generate lots of data, such as DTs. One of the highest expectations of complex systems such as DTs is to achieve transparent integration, where data can be accessed, recovered, and treated through techniques, tools, and algorithms uniformly.

A significant challenge mentioned in the decision-driven data management context, and DTs in particular, is the existence of data silos. Data silos are isolated groupings of stored data formed by different applications and processes, where the data resides in isolation in cloud applications, on-premise databases, and application servers [30]. Fragmented data within an organization generates cores of information not shared between departments and makes generalized access to data difficult. Interoperability and data integration are vital for addressing data silos.

Data integration aims to combine data from multiple sources, providing a high-level unified view [31] that makes data amenable to use in Analysis and Usage activities. Physical data integration is the most popular approach, where the source data is combined within a new dataset or database tailored for analysis activities. In virtual data integration, the data entities remain in their original data sources and are accessed at runtime. Critical tasks for data integration are data transformation, semantic enrichment, entity resolution (data matching), entity merging, and combining and merging metadata models such as schemas and ontologies [32, 33]. Semantic data enrichment can be achieved by linking entities or metadata such as attribute names to knowledge resources (e.g., dictionaries, ontologies, knowledge graphs). Integration in the DT scenario should consider data from different application domains, which have significant semantic, terminological, and syntactic heterogeneity, lack of standards, and low quality. Overcoming these challenges allows the development of a unified view of different parts of the DT, enabling the production of actionable and valuable knowledge for the business.

Interoperability is a multidimensional concept comprising multiple perspectives and approaches from different communities according to the application domain. The IEEE Glossary defines interoperability as “the ability of two or more systems or components to exchange and use the information exchanged in a heterogeneous network" [34]. A systematic review of cyber-physical systems [35] identified ten types of interoperability. Semantic and data

interoperability are particularly relevant in the scope of DT data management and, according to the Big Data value chain, are part of Data Curation activities. Data interoperability is the ability of data to be accessible, reusable, and understandable by all transaction parties by addressing based on a shared understanding regardless of different representations, purposes, contexts, and syntax-dependent approaches. Semantic interoperability refers to a data layer that allows computers to share, understand, interpret and use the data unambiguously based on a common meaning of the data [36]. The basis of the operation of DT is the exchange of data and information between its different layers and between DTs from different application domains. Data interoperability and semantic interoperability are fundamental in the context of DTs because they allow data sharing effectively, unlocking barriers of communication and understanding, and making dependent activities and processes more fluid. Data analysis activities can also benefit from data and semantic interoperability since they contribute added-value data and information. In this way, the analyses become richer, generating more interesting and valuable insights.

Generating value from data requires finding, accessing, and making sense of datasets. *Data search and discovery* are among the analysis activities that enable users to find, understand, and trust the information used to generate value from data. Broadly speaking, a query is a semantically and syntactically correct expression of a search, which can be addressed in a range of scenarios, depending on the types of data and methods used [37]. Relevant sub-disciplines include databases, document and keyword search, entity-centric search, and semantic search, among others. The required underlying infrastructure for handling the search consists of query parsers and evaluators, indexes for various data types, metadata and ontologies, reasoners, etc. DTs can exploit helpful search and discovery in analysis activities to filter, transform, model, and extract hidden information from raw or transformed data. Notice that the data search objective is different when considered in the realm of data usage activities. From a DM component perspective, the goal is to support data consumption, such that relevant data can be submitted to models (e.g., predictive or simulation models, what-if scenarios), event managers, or consumed through dashboards, visualizations, or reports.

To guarantee quality information from a Data Curation standpoint, it is necessary to develop methods, metrics, and tools to manage *data quality*. The literature has defined specific characteristics or dimensions to manage data quality, such as timeliness, completeness, consistency, accuracy, etc. [38]. Timeliness is related to the age of the data being adequate for the task at hand. Completeness seeks to measure whether there are missing or null data. Consistency measures how consistent the data is with previous data, and accuracy measures how accurate the data is relative to actual values. Data cleaning and preprocessing operations are integral parts of integration activities, such as dealing with noisy and missing data, duplicated data, outlier detection, normalization, transformation, etc. High-quality data is essential in any context, whether in business, decision support systems, machine learning algorithms, or DTs. Big data has increased the complexity of managing data quality as the heterogeneity and volume of data have increased. In the context of DTs, data quality management gains a prominent space when considering the data flow loop and information that characterizes it. Therefore, it is necessary to prioritize data quality in developing and managing a successful DT.

As there is no consensus on the role or functionality for the DM component of a DT, we extracted from the Big Data value chain activities the key issues it should handle. DTs and Big Data are mutually reinforcing technologies, implying the need to handle Volume, Velocity, Variety, Veracity, and Value. Data heterogeneity is a key underlying property and determines the support for Acquisition. The Analysis activities are greatly influenced by the

type, completeness, richness, and quality of the data available, and hence support for data integration, interoperability, and data search are also fundamental. Finally, the value of data is intrinsic to the quality of data, a central aspect of Data Curation. We contribute by identifying an initial set of significant issues to summarize the state of the art in terms of solutions for DM in DTs and, from there, identify the challenges and possible trends.

4 Systematic literature review

The methodology adopted followed the guidelines of systematic literature reviews proposed by [39] for the software engineering field. The process is divided into three phases: planning, conducting, and reporting. The *planning phase* evaluates the need for the proposed systematic review, defines the research questions to be answered, and determines the review protocol. The protocol encompasses (i) a search strategy that maximizes relevant results; (ii) a selection process to minimize bias; and (c) how to efficiently summarize data from these studies. The next phase is *conducting the review*, in which primary studies are selected, from which data is extracted, summarized, and assessed. Finally, the *reporting phase* answers the research questions and disseminates the findings. We developed this review using two online collaborative systems: *Parsifal*,[1] a support system for conducting SLRs, and *Mendeley*,[2] a reference manager system.

4.1 Objective and research questions

An SLR fulfills the requirement of summarizing all current information about some phenomenon thoroughly and unbiasedly. The main objective of this SLR is to systematically examine works addressing DTs and summarize the solutions proposed for the critical DM issues identified in Sect. 3. According to this goal, we defined five research questions (RQ), presented in Table 1 with the respective motivation.

4.2 Search strategy

An SLR focuses on identifying the primary studies that can answer the research questions. First, we sampled papers by identifying relevant works using different Digital Libraries (DLs) and by snowballing the references from these works and surveys. We used these sample studies in three ways: a) to define the terms for an initial search string; b) as control papers in the refinement and validation of the search string; c) to delimit the search period. We defined 2014 or later as the search period because we did not identify any relevant work before 2014 in the snowballing process used to constitute this sample. The decision on the lower bound of the search period is in line with existing seminal SLRs and surveys about DTs [2, 4, 10, 19], which identify 2014-2015 as the initial year of relevant publications about DTs. Note that the pioneering work that proposed a central DM component [17] dates from 2017, and therefore our sample seems consistent.

We refined the list of keywords iteratively according to the quality and amount of studies resulting from the DLs search. The search string was composed using two categories of terms: (i) synonyms for the term Digital Twin and (ii) data management issues/functionality

[1] https://parsif.al.

[2] https://www.mendeley.com/.

Table 1 Research questions and respective motivations

ID	Research question (RQ) Motivation (M)
RQ1	*RQ: For which domains the DT solutions were proposed?*
	M: In order to identify the most active and mature DT areas for which DM solutions were proposed, this question identifies the respective domain/subdomain
RQ2	*RQ: Under the perspective of data usage, for which functions were the DTs proposed?*
	M: This question aims to identify the main functions for which a DT was proposed, considering data usage under the value chain. It provides a context on the proposed functionality for acquiring and managing data to be consumed by other components of the DT
RQ3	*RQ: What types of data the proposed solutions consider?*
	M: This question aims to characterize the heterogeneity of the data that needs to be managed in the DT. It indicates the completeness of the solution concerning data types, formats and velocity requirements
RQ4	*RQ: What solutions were proposed for the DM issues addressed?*
	M: This question aims to survey the DM solutions proposed for the key data management issues raised in Sect. 3, namely interoperability, integration, data search and discovery, and data quality. It sheds light on the specific problems addressed, and how encompassing is the scope of the DM component considered
RQ5	*RQ: What kind of technological infrastructure is considered?*
	M: This question surveys the technological infrastructure leveraged or suggested for the implementation of the proposed DM solutions. It aims to identify technological trends that support data management in DTs

Table 2 Selected digital libraries

ID	Digital library	URL
ACM	ACM Digital Library	http://dl.acm.org
IEEE	IEEE Digital Library	http://ieeexplore.ieee.org
OP	Onepetro	http://www.onepetro.org
Sco	Scopus	http://www.scopus.com
S@D	Science Direct	http://www.sciencedirect.com
WoS	ISI Web of Science	http://www.isiknowledge.com

required to handle data in DTs. We evaluated each search to reach a suitable set of studies, verifying if the results included the control papers. The final search string was:

("digital twin" OR "cyber-physical" OR "CPS" OR "digital model")
AND
("data management" OR "data integration" OR "data repository" OR "data transformation" OR "data provenance" OR "data governance" OR "heterogeneous data" OR "data interoperability" OR "metadata management" OR "data storage" OR "data quality" OR "data enrichment" OR "data modeling"))

Table 2 summarizes the DLs used. These DLs index the main journals and conferences on computer science, enable the search of papers using expressions combining keywords and logical expressions, and allow for the search to be performed in the title, abstract, and keywords.

The results reported in this paper refer to the last search performed on November 20th, 2021. With the support of Parsifal, we exported the text files containing the Bibtex references

for the articles and eliminated the duplicates. We retrieved the files for the screened papers and input them into Mendeley.

4.3 Study selection

The search retrieves potentially relevant primary studies, of which the actual relevance needs to be confirmed. The protocol defines inclusion and exclusion criteria to filter out retrieved studies not aligned with our objectives. The *inclusion criteria* are the following: (1) papers written in English; (2) papers published within the defined search period (2014 or later); (3) studies addressing DTs or DSs, according to the definition in [2]; (4) studies that explicitly propose DM solutions for DTs/DSs, (4) primary studies; (5) papers published in peer-reviewed *fori*. To discard studies that are not relevant to this SLR, we defined the following *exclusion criteria*: (1) type of publication by eliminating materials such as short papers (3 pages or less), reviews, secondary studies, reports, books, textbooks, theses and dissertations, editorial letters, brief communications, posters, commentaries, unpublished working papers; (2) non-English papers; (3) papers with full text unavailable; (4) papers published in non-peer-reviewed *fori*; (5) studies that do not explicitly address DTs/DSs; (6) studies that do not explicitly address a DM solution in the context of a DT/DS. Regarding this last criterion, we disregarded all studies describing data preprocessing designed to prepare/improve the input to a specific model (e.g., predictive model, visualization service, simulation), as well as studies focusing on the deployment of middleware or cloud computing without a specific underlying DM functionality.

The selection of studies occurred in three steps: (i) preliminary *screening* using the title, abstract, and keywords; (ii) *pre-selection* of candidate studies considering the introduction and superficial reading of the paper; (iii) final *selection* based on the full reading of the paper. Two authors independently performed the screening phase. To align the interpretation of the inclusion/exclusion criteria, these two authors also independently read all pre-selected papers, discussing all cases in which there was a doubt or disagreement. In the final selection, the authors discussed all the cases that generated concerns.

Figure 2 depicts the complete flow for selecting studies, highlighting the reasons for excluding articles at each step. The search in the six DLs identified 1,838 articles, reduced to 1,165 after eliminating duplicates. The screening of papers resulted in 664 candidate studies, from which we pre-selected 180 candidate studies based on the introduction, figures, and conclusions. After full reading, we selected 61 studies aligned with the primary objective of our SLR.

4.4 Data extraction and summary

The data extraction strategy aims at helping to answer the research questions by allowing the researchers to summarize and categorize articles, thus improving the understanding of the domain. We extracted the data for each selected paper by filling a form with the predefined fields described in Table 3.

We extracted data from the 61 selected studies according to data extraction form, and the results are summarized in Table 4. We observed that 65% of the studies were published in the last two years and that 65% of studies were developed by academia alone. We identified contributions from twenty-five countries, where the most active ones are China, Germany, the UK, Italy, and the USA. In the next section, we presented our findings in light of our research questions.

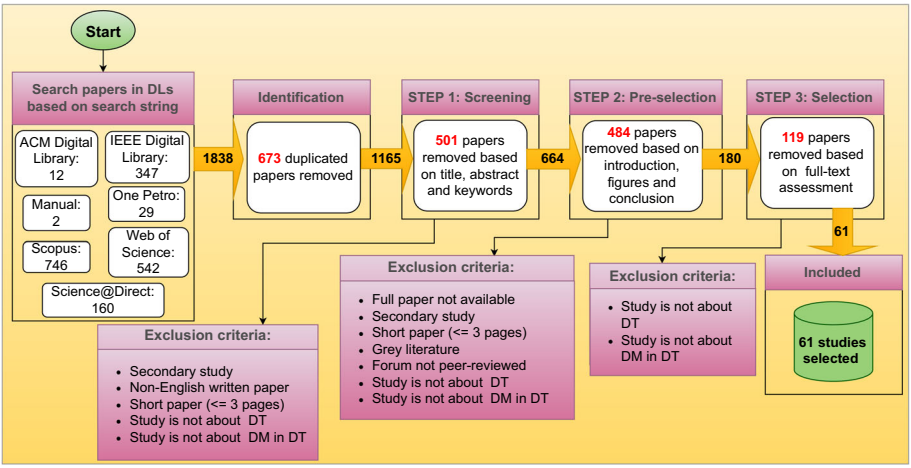

Fig. 2 Study selection results

Table 3 Extraction form

#	Study data	Description	Relevant RQ
1	Identifier	Unique id for each study	Study overview
2	Year, Authors, Country	Year of publication, authors' names, affiliations and respective country	Study overview
3	Scientific venue	Journal or Conference	Study overview
4	Institution type	Academia, Industry, or Both (based on authors' affiliation)	Study overview
5	Domain/subdomain	Generic and specific domain of the DT	RQ1
6	Function	The DT function according to data usage perspective (Semeraro et al., 2021)	RQ2
7	Type	Digital Twin or Digital Shadow (Fuller et al., 2020)	RQ2
8	Data type	Data heterogeneity information, including the types of data handled and file formats	RQ3
9	DM issues and solutions	DM issue(s) (interoperability, integration, quality and/or search/discovery) and the solution proposed	RQ4
10	Technological infrastructure	Technological infrastructure used/recommended for implementing the solution, with a focus on cloud processing and DBMS	RQ5

5 Results

5.1 RQ1: For which domains the DT solutions were proposed?

This question aims at identifying the most active and mature areas of DT research for which the DM solutions were devised. According to Table 4 (column *Domain*), most selected studies address the domains of I4.0 (61.9%), followed by smart cities (23.8%) and healthcare (6.3%). Five studies (8.2%) do not detail the domain/subdomain, and one study is a domain-agnostic proof-of-concept (POC) [40]. The following domains were identified as most relevant for DM solutions:

a) I4.0: this is the most active domain for which DM solutions were proposed. It indicates the evolution from *ad hoc* data management to the proposal of explicit DM solutions. Considering specific areas within I4.0, smart manufacturing is the predominant one (27.9%), followed by O&G (6.5%), aviation and energy (4.9% each), and the automotive industry (3.3%). Some studies [16, 41] have not defined a specific I4.0 sub-area.

Studies in the specific area of smart manufacturing are aimed at different purposes, ranging from custom manufacturing [42] to methods for building DTs for manufacturing factories [43], monitoring and predicting the carbon emission of a factory [44], optimizing hot lamination schedule [45], among others. In the specific area of O&G, the DTs address asset management [46], improvement in drilling operations [47, 48], and improvement in digital reservoir simulation [49].

b) Smart cities (SC): we identified DM solutions for DTs targeted at different levels of the urban environment, ranging from residential housing *(smart house)*, building management *(smart building)* to city level *(smart city)*. The proposals aimed at the smart city level (57.1%) either encompass the city as a whole [78, 79, 88], or a specific aspect, such as urban planning [86], disaster management [87], traffic management [85], urban water supply [83], and mobility [89]. The second most prevalent level is smart building (35.7%), where the objective is monitoring the building's energy consumption [82], controlling the internal temperature [14], detecting faults and anomalies [77], as well as monitoring the building's infrastructure [80]. Two studies [78, 79] did not define a specific city scenario for which the solution was proposed.

c) Healthcare (HC): we identified two groups of applications in this area, namely health data management and treatment of patients. In the former group, [50] proposes the integration of data from different stakeholders (e.g., hospitals, pharmaceuticals, patients) to leverage the creation of new services and applications, and [51] proposes a solution for managing the hospital and patient data based on the easier data exchange among different systems. As for the studies focused on the treatment of patients, there are POCs targeted at analyzing data on heart disease and diabetes [52] and predicting the risk of stroke [53].

5.2 RQ2: Under the perspective of data usage, for which functions were the DTs proposed?

The surveyed literature revealed different purposes for the proposed DTs. From the data usage perspective for adding value to the business through the DT, we classified these purposes as the primary DT function [18]. The functions identified are summarized in Table 4 (column *Function*). In most studies, it was classified as decision support (24.6%) since the DT allows the analysis of multiple variables and hence, supports data-driven decision-making. In addition,

Table 4 Selected studies

Study	Type	Domain	Subdomain	Function	Data type	Interop.	Integ.	Search	Qual.	Infrastr.
[50]	DS	HC	Healthcare	Decision support	Real-time, historical	✓	✓			Cloud
[51]	DS	HC	Healthcare	Decision support	Real-time, historical	✓				Cloud
[52]	DS	HC	Healthcare	Decision support	Historical		✓			Unspecified
[53]	DS	HC	Healthcare	Decision support, asset monitoring	Real-time or near real-time		✓	✓		Cloud, SQL, NoSQL
[54]	DS	I4.0	Smart manufacturing	Anomaly detection	Real-time, historical		✓			Cloud, SQL, NoSQL
[15]	DS	I4.0	Aviation industry	Asset monitoring	Historical		✓	✓		SQL
[41]	DS	I4.0	Industrial	Asset monitoring	Historical	✓				Cloud
[55]	DS	I4.0	Maint. of industrial robots	Asset monitoring	Historical		✓			Unspecified
[46]	DS	I4.0	Oil and Gas	Asset monitoring	Real-time				✓	Unspecified
[56]	DS	I4.0	Smart manufacturing	Asset monitoring	Historical CSV		✓		✓	NoSQL
[42]	DS	I4.0	Smart manufacturing	Asset monitoring	Near real-time, historical		✓			Unspecified
[43]	DT	I4.0	Smart manufacturing	Asset monitoring	Real-time, historical			✓		Unspecified
[57]	DS	I4.0	Smart structures	Asset monitoring	Real-time JSON		✓			Cloud, NoSQL
[58]	DS	I4.0	Automotive glazing industry	Decision support	Real-time		✓			Cloud, edge, SQL
[29]	DS	I4.0	Eletric energy	Decision support	Historical		✓	✓		NoSQL
[16]	DS	I4.0	Industrial	Decision support	Real-time, historical		✓	✓		Unspecified
[44]	DS	I4.0	Smart manufacturing	Decision support	Real-time historical XML, JSON				✓	Cloud, SQL
[59]	DS	I4.0	Smart manufacturing	Decision support	Real-time		✓		✓	Cloud, edge, SQL, NoSQL
[60]	DS	I4.0	Smart manufacturing	Decision support	Historical XML, tabular		✓			SQL, NoSQL
[61]	DS	I4.0	Smart grid	Decision support, simulation improv	Real-time, historical				✓	Unspecified

Table 4 continued

Study	Type	Domain	Subdomain	Function	Data type	Interop.	Integ.	Search	Qual.	Infrastr.
[47]	DS	I4.0	Oil and Gas	Event monitoring	XML, JSON, XLSX, PDF		✓			Cloud, SQL, NoSQL
[62]	DS	I4.0	Smart grid	Fault detection, predictive maint	Real-time		✓			SQL, NoSQL
[63]	DT	I4.0	Smart manufacturing	Fault diagnosis, predictive maint., decision support	Real-time XML		✓			Unspecified
[64]	DS	I4.0	Automotive manufacturing	Optimization of assembly	Real-time, historical		✓			Unspecified
[65]	DS	I4.0	Aviation industry	Optimization of assembly	Real-time CIM/XML		✓			NoSQL
[66]	DS	I4.0	Smart manufacturing	Optimization of assembly	Real-time XML	✓				SQL
[67]	DS	I4.0	Smart manufacturing	Optimization of assembly	Real-time, historical		✓			SQL, NoSQL
[13]	DS	I4.0	Steel rebars manufacturing	Optimization of logistics	Historical JSON, CSV	✓	✓			Cloud, edge, SQL, NoSQL
[68]	DS	I4.0	Smart manufacturing	Optimization of process	Real-time		✓			Cloud, NoSQL
[69]	DS	I4.0	Supply chain	Optimization of process	Real-time, historical		✓			SQL, NoSQL
[48]	DS	I4.0	Oil and Gas	Optimization of production	Real-time	✓				Cloud, edge
[70]	DS	I4.0	Smart manufacturing	Optimization of production	Real-time Tabular				✓	Unspecified
[71]	DS	I4.0	Smart manufacturing	Optimization of production	Historical XML		✓			Unspecified
[72]	DS	I4.0	Smart manufacturing	Optimization of production	Real-time, historical XML		✓			SQL, NoSQL
[24]	DS	I4.0	Metal Additive Industry	Optimization of production, decision support	Real-time, historical XML		✓			Cloud, edge

Table 4 continued

Study	Type	Domain	Subdomain	Function	Data type	Interop.	Integ.	Search	Qual.	Infrastr.
[45]	DS	I4.0	Smart manufacturing	Optimization of production, decision support	Real-time, historical		✓			Unspecified
[73]	DS	I4.0	Aviation industry	Predictive maint	Real-time, historical	✓				Blockchain storage
[74]	DS	I4.0	Smart manufacturing	Prescriptive maint., decision support	Real-time, historical		✓			SQL
[75]	DS	I4.0	Chemical industry	Quality assessment, decision support	Historical				✓	Unspecified
[49]	DS	I4.0	Oil and Gas	Simulation improv	Historical	✓			✓	Unspecified
[76]	DT	I4.0	Smart manufacturing	Simulation improv	Real-time, historical	✓	✓			Cloud, SQL
[77]	DS	SC	Smart buildings	Anomaly detection	Historical XML	✓	✓		✓	NoSQL
[78]	DS	SC	Smart city	Anomaly detection, asset monitoring	Real-time, historical		✓			Cloud, NoSQL
[79]	DS	SC	Smart city	Anomaly detection, asset monitoring	Real-time, historical		✓		✓	SQL
[80]	DS	SC	Smart Buildings	Asset monitoring	Real-time, historical CSV, XLSX		✓			SQL, NoSQL
[81]	DS	SC	Smart Buildings	Asset monitoring	Historical, JSON		✓			NoSQL
[82]	DS	SC	Smart Buildings	Asset monitoring	Near real-time, historical JSON		✓			Cloud, NoSQL
[14]	DS	SC	Smart Buildings	Asset monitoring	Real-time		✓			Cloud
[83]	DS	SC	Urban water supply	Asset monitoring, decision support	Real-time, historical				✓	Cloud, edge
[84]	DS	SC	Smart house	Decision support	Real-time CSV	✓	✓		✓	Unspecified
[85]	DS	SC	Traffic management	Decision support	Historical XML		✓	✓		SQL
[86]	DS	SC	Urban planning	Decision support	Near real-time, historical		✓	✓		Cloud, NoSQL

Table 4 continued

Study	Type	Domain	Subdomain	Function	Data type	Interop.	Integ.	Search	Qual.	Infrastr.
[87]	DS	SC	Disaster management	Event monitoring, decision support	Real-time		✓			Unspecified
[88]	DS	SC	Smart city	Event monitoring, decision support	Real-time, historical		✓	✓	✓	Unspecified
[89]	DS	SC	Urban mobility	Event monitoring, decision support	Unspecified	✓	✓		✓	Unspecified
[90]	DS		Unspecified	Decision support	Unspecified			✓		NoSQL
[40]	DS		General	Event monitoring	Real-time, XML		✓			Cloud
[91]	DS		Unspecified	Event monitoring	Real-time		✓			Unspecified
[92]	DS		Unspecified	Fault detection	Unspecified				✓	Unspecified
[93]	DS		Unspecified	Optim. of process	Historical		✓			Unspecified
[94]	DS		Unspecified	Quality assessment	Unspecified				✓	NoSQL

Table 5 Data consumption methods

Method	Study ID	%
Visualization (diagrams, graphics, maps, dashboards)	[14, 24, 43, 47, 53, 54, 57, 59, 62, 63, 70, 71, 73, 74, 77, 79, 86–88, 93]	32.8%
Models (simulation, ML, 3D)	[14, 48, 49, 64, 66, 68, 70, 78, 79, 81, 82, 89]	19.7%
Other applications and services	[13, 14, 16, 41, 51, 58, 60, 65, 69, 73, 78, 87]	19.7%
People (data specialists,decision-makers,stakeholders)	[45, 50, 52, 58, 61, 75, 78, 84, 85, 90]	16.4%
Unspecified	[42, 44, 46, 55, 56, 67, 80, 91, 92, 94]	16.4%
Database query	[15, 16, 29, 43, 53, 84, 86, 88, 90]	14.7%
GUI for data query	[40, 51, 72, 76, 77, 83]	9.8%

we identified more specific functions, namely optimization of production/process/logistics (22.9%), asset monitoring (21.3%), event monitoring (9.8%), anomaly detection (6.6%), failure detection/diagnosis (4.9%), quality assessment (3.3%), simulation improvement (3.3%), and predictive/prescriptive maintenance (1.6%).

The subdomains are related to varied functions, but we noticed a few patterns: all DTs in the healthcare domain are focused on decision support; most smart manufacturing DTs are concerned with asset monitoring and optimization in general; and smart buildings tend to monitor the asset.

From Table 4 (column *Type*), it is possible to observe that only three studies [43, 63, 76] are classified as DTs according to the definition in [2], i.e., there is a closed feedback loop (bidirectional) from the Virtual Space into the Physical Space. All the other studies are classified as DSs.

We examined the consumers of the data managed by the DT, which can be part of either the Service component or Virtual Space. Table 5 presents the data consumers identified in the selected studies. The most frequent type of data consumer is visualization tools, including dashboards, graphs, maps, and diagrams. Other studies enable users to formulate queries using customized Graphical User Interfaces (GUI) or database query languages. Some studies merely indicate that the consumers are humans (e.g., data specialists, decision-makers, stakeholders) without detailing the specific tools/services used. When data usage is ultimately targeted at humans, the prevalent DT function categories are decision support (52.1%) and asset monitoring (20.8%).

The data managed by the DT are input to models (e.g., simulation, ML, 3D visualization) or other applications/services. The prevalent DT function categories related to this type of consumers are asset monitoring and optimization in equal proportions (30.8% each) and decision support (23%). Some studies (16.4%) did not explicitly mention any method for data consumption in the DT.

5.3 RQ3: What types of data the proposed solutions consider?

The heterogeneity of data that needs to be handled in all domains is clear, which includes sensors/actuators, information systems (e.g., ERP, MES, CRM, SCADA), and data silos (data

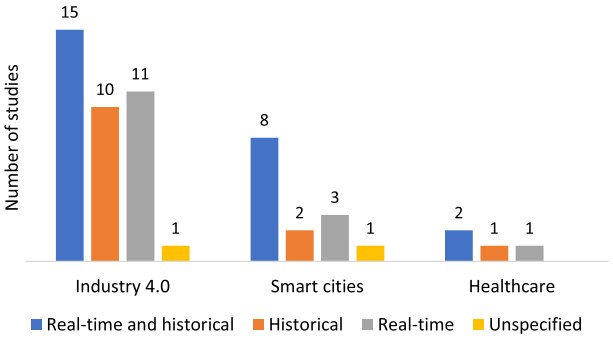

Fig. 3 Data collection and processing methods

repositories), social networks, among others. This research question aims at characterizing the heterogeneity in the selected studies in terms of sources of information, formats, and data collection/processing methods according to data analysis latency requirements. Table 4 (column *Data Type*) summarizes the type of data and formats.

a) Sources: considering the sources for the data acquisition activities, we identified that most studies (63.9%) consider both (near) real-time data streams generated by sensors/actuators, together with historical data related to all sorts of information systems and file specifications. About 34.4% of the selected studies are restricted to data streams. While the former provides support for integrated data that reflects a more complete view of the whole (e.g., different types of expertise, processes, or organization sectors), the latter, which only mentions data from sensors and actuators, typically consists of more closed-scope analyses with a specific focus.

b) Formats: about 32.8% of the selected studies explicitly mention the data format used in the proposed solution, ranging from structured data (typically tabular data), semi-structured data (XML, JSON, CSV, XLSX) to unstructured data (e.g., PDF). Among the mentioned issues is the dependency of the data format concerning the software that generated it, transformations required, and integration with other data.

c) Data Analysis latency: applications vary in the requirements of the freshness of the data for the analysis, impacting the way the need to be collected and processed. We identified three groups: real-time and historical, real-time only, and historical only. Figure 3 displays the distribution of data latency analysis per domain. DTs whose functionality is to monitor and manage events require low data latency. Examples are event monitoring in general [40, 91], management of disasters and city events [87, 88], faults in the power grid [62], and production/assembly lines optimization [48, 66, 70]. In these cases, dealing with real-time data in terms of data gathering and processing is critical, given that the data availability and processing for data analysis need to happen within a small time window. However, the data analysis latency requirements from DTs with functionalities to improve simulation [49], decision support [29, 52, 60, 85], or optimization [13, 93] are more flexible and less time-dependent, characterizing the use of historical data and also allowing the availability and analysis of data to take place over a longer time window and more detailed analysis. The combination of historical and real-time data, in general, allows for a deeper analysis of the data and a better understanding between the past and the present, using historical data as a baseline to create models for assessing real-time data. This combination is identified in DTs

with functionalities such as decision support [16, 44, 50, 51, 53, 61], anomaly detection [54, 78, 79], predictive and prescriptive maintenance [73, 74], and asset monitoring [43, 80, 83].

5.4 RQ4:What solutions were proposed for the DM issues addressed?

In this question, we summarize the DM solutions proposed by the selected studies according to the issues raised in Sect. 3, namely Interoperability, Integration, Data Search and Discovery, and Data Quality. Selected studies may propose solutions that cover more than one issue.

5.4.1 Interoperability

We identified solutions addressing data and semantic interoperability within a DT and between DTs:

a) Data interoperability: studies in this category propose data format conversions and mappings to allow seamless data exchange between data layers or components within a DT. [76] proposes data mappings to handle data in different formats and interoperate with the other parts of the DT, and [13, 51, 66] propose the conversion of the original data format into a common structure. [49] presents an architecture with built-in conversions to promote seamless data exchange between simulation tools. [50] discusses a middleware for converting files into standard formats with corresponding metadata (e.g., source, attributes and domains, security tag). Blockchain technology is leveraged in [73] to maintain data traceability after cleaning and conversion/transformation operations to a unified format, enabling the exchange of information between the DT parties.

b) Semantic interoperability: this group of studies uses domain concepts, patterns, and ontologies to describe data in terms of the respective domain. [51] develops a conceptual framework for a health CPS, which includes an interoperability manager that converts data into a standardized format using domain terminology, resulting in a more meaningful semantic structure. [89] proposes the concept of a Data Sharing Market as an information exchange model, where each market is a cloud node with agents that describe the data they provide for sharing, with the respective pipeline of cleaning and transformations to prepare data for consumption. [48] uses the OPC UA protocol, an industry automation standard, to provide semantic interoperability through data tag descriptions for the DT components. [77] created an integrated and semantically interoperable data layer, which includes a model for exchanging data with other data sources.

c) Interoperability between DTs: Some works assume that DTs will be widely adopted within sectors of the same organization or that different organizations will collaborate, and hence their DTs must interoperate. [41] proposes rules for mapping a DT source information model into a destination DT information model (data interoperability) and using domain standard data dictionaries for a common understanding of the concepts (semantic interoperability). Considering semantic interoperability, [85] leverages ontologies to relate data to domain metadata and develop a framework for exchanging data between DTs in different domains.

5.4.2 Integration

We divided the studies addressing integration into six approaches, summarized in Table 6 and described below:

Table 6 Integration approaches

Integration approach	Study ID	%
Centralized repository	[40, 54, 59, 61, 62, 74, 79, 86]	17.8%
Ad hoc modeling	[24, 42, 50, 64, 68, 76, 85, 87, 89]	20%
Modeling with standards	[14, 65, 72, 77, 78]	11.1%
Modeling with an ontological layer	[13, 15, 16, 29, 45, 56, 63, 67, 71, 80, 81, 88, 93]	28.9%
Integration Method	[55, 60, 69, 82, 91]	11.1%
Semantic enrichment	[52, 53, 57, 58, 74, 84]	13.3%

a) Centralized repository: the solution proposed by the studies in this category is limited to gathering all heterogeneous data from different sources in a single, centralized repository. Some studies do not provide details, mentioning the adoption of a data warehouse [40, 74, 79] or graph database [61]. Some of them [54, 59, 86] relied on the support of open-source tools such as Apache Spark and Apache Sedona for the logical integration of data sets. In addition to the use of a central repository, [62] also proposes the development of a data virtualization layer on top of the data lake to provide a unified view of data coming from heterogeneous sources.

b) Ad hoc modeling: studies in this category integrated heterogeneous data by proposing a unified data model, representing and interrelating different data. These works are referred to as *ad hoc* as they have a modeling solution targeted at the scope of a specific DT function and domain. Data models were proposed to support a machine learning pipeline [85, 87, 89]; a 3D constructor component to provide a unified view of assembly lines [64]; a Human Body Avatar Data Model that integrates the description of all data and provides customized assistive healthcare devices [42]; the identification of the critical product life-cycle data that influences the quality of the final product [24]; a machining process for merging and syncing the data from machines, workpiece, tools, process, and information systems [68]; the increase in the transparency and automation level of a blade test workshop [76]; and a data management layer that provides a basic data model for the data from different sources [50].

c) Modeling with standards: studies in this category have used standards as a support for the creation of an integration data model. In the domain of smart cities, [14, 77, 78] leveraged the Build Information Modeling (BIM) standard, and in the I4.0 field, [65] adopted the Bill of Materials (BOM) standard and [72] explored the ISO STEP standard for product information exchange. In addition, [65, 72] also proposed a method of data association to support quality traceability. To create a data model based on standards, [78] used a centralized non-relational repository (DynamoDB).

d) Modeling with an ontological layer: the solution proposed by the studies in this category relies on an ontological layer to represent and relate the information with domain knowledge so as to create a shareable knowledge base. Several works [16, 29, 71, 88, 93] proposed an ontology and the use of an RDF graph to relate the conceptual data model and domain data. [67] uses several knowledge graphs to integrate the different stages of the assembly process. Based on three ontologies, [45] proposed a data model based on an ontological layer and a method addressing data fusion and entity resolution. We also found four studies [13, 15, 63, 80, 81] that combine the use of ontologies and standard domain models to integrate

the information and semantic models. In addition to an ontology-based data model, [56] proposed a module for cleaning and reducing data.

e) Integration method: this group gathers the studies that propose methods or mechanisms addressing specific data integration issues. These works address the specification of mapping and entity resolution [69]; data synchronization [82]; the automatic mapping of entities [60]; the integration of two complex event models using an adapter [91] and association mappings and data fusion [55].

f) Semantic enrichment: studies in this category assume an existing data/information model, and through metadata annotation methods or semantic knowledge bases, they add the semantic level to the current data model to facilitate data integration. [53, 74] created a semantic knowledge base based on domain ontologies to enrich already processed data with more domain meaning, enabling data integration. [52, 57, 58, 84] used semantic metadata annotation methods to optimize data integration by identifying new classes, relationships, or domain descriptions.

5.4.3 Data search and discovery

A more limited number of works address functionality enabling users to find, understand and trust the value of data used to generate value from data. We divided the works addressing data search as structural and semantic search. We disregard in this section all the works that provided specific GUI interfaces for accessing data using predefined queries listed in Table 5, detailing only those which provided query functionality.

a) Structural data search: the studies in this group have developed strategies to facilitate syntactic data queries, i.e., based on the structural properties of the data. The use of Elasticsearch for indexing and discovering data is leveraged in [53, 86]. [84] developed an extension that integrates the standard Functional Mockup Unit (FMU) model into the relational model so that data scientists can more easily find the data useful for machine learning models. A few works address issues related to OWL representation of data that were included in the data model. [90] addressed the efficiency of information retrieval by transforming and storing the original OWL data representation in a NoSQL database. [15] claims that the industry is more familiar with relational structures and proposes the conversion to a relational database to enable SQL queries.

b) Semantic data search: the studies in this category leverage the semantic layer included in the data representation for enhanced semantic queries, which in all works are knowledge graphs. [29, 53, 88] deployed knowledge graphs in their data modeling solutions, arguing that one of the advantages is that queries can be performed using SPARQL, which enables to formulate queries as logical conditions over the structure of a triple Subject→Predicate→Object). [43] recommends using knowledge graphs for semantic queries through the GraphQL query language and APIs that allow queries in JSON format. [16] makes a case for knowledge graphs and semantic queries without referring to any specific query language.

5.4.4 Data quality

We divided the works proposing solutions for guaranteeing/improving the quality of data into the following categories:

a) Statistic-based methods: works in this category deal with quality issues of streaming data, for instance, due to sensor malfunctioning, communication issues, malicious data insertion, etc. Most techniques are grounded on the statistical properties of data streams/time series and are used to clean raw sensor streams. [94] compiles useful statistical methods, and [46] proposes data checking and cleaning algorithms based on statistical properties of the data. To improve the input of simulation tools, [47] proposes methods for aligning time series in different time scales. A data quality model fitted to the specific properties of signal data of industrial processes is proposed in [75].

b) Deviation/anomaly detection methods: works in this class propose solutions that assess the quality of data according to specifications, rules, models, or thresholds derived from the information model that contextualizes data streams raw data. To seamlessly transfer simulation data to other applications, [49] proposes a two-layer comparison (specification and threshold/rules). HADES [79] assumes two levels of cleaning: comparison of real-time data with historical data and assessment by predictive anomaly detection models. In addition to cleaning, [44] and [77] also propose the use of models to compare deviations from expected data. [92] proposes a quality assessment method with the respective techniques to assess input data, out data generated by models, and feedback data from the CPS.

c) Pipeline/reference architecture: works grouped into this category address a pipeline of operations to clean the data, which can be part of a reference architecture or framework. Examples of cleaning operations over raw data handle missing data, duplicate data, and outliers [56, 59, 70, 85, 88]. The pipeline in [59] and [70] are part of an ETL (Extraction, Transformation, and Loading) process designed to insert cleaned and transformed data in a data warehouse, while the one in [88] cleans data before transforming data into RDF triples. These operations may be inserted in a reference architecture or framework as layers or functional components with a specific quality-checking or preprocessing role. The reference architecture in [56] proposes two levels of cleaning (raw data and customized), while the one in [83] organizes operations and models to clean raw data, to assess data properties, and to add value into four data management layers. The architecture in [89] leverages data quality agents to perform well-known cleaning patterns.

5.5 RQ5: What kind of technological infrastructure is considered?

Our last research question aims to identify the technological infrastructure used or suggested by the studies selected for the data management component of the DT. We considered only the studies that explicitly describe the IT infrastructure for data processing and storage, grouping them into three categories: cloud computing, hybrid computing (cloud, edge, fog), and database management systems (DBMS).

a) Cloud computing: This category groups the studies that used or recommended the use of cloud data processing and/or storage resources. While some justify the adoption due to processing requirements [40, 44], others focus on storage to achieve scalability, security, and availability [68]. Most studies adopt both processing and storage [13, 14, 41, 50, 51, 53, 54, 57, 62, 76, 78, 82, 86]. Some studies adopted open-source tools and databases [53, 54, 57], such as Spark, Kafka, and Hadoop for data processing, and Elasticsearch, Hive, InfluxDB, MongoDB, Hadoop, and MariaDB for data storage. Others [14, 62, 78] have used or recommended services from commercial cloud providers, such as DynamoDB, S3, and Redshift from AWS.

b) Hybrid computing (cloud, edge, fog): studies in this category distinguished between the concepts of cloud, edge, and fog computing. In terms of cloud and edge computing, we have identified three studies [48, 58, 83]. Typically, cloud resources perform more complex computational processing and global tasks, while edge resources serve for faster computing and as a local server. Only one study [89] has mentioned the importance of the cloud, edge, and fog combination, mainly due to real-time data processing requirements. Lastly, we found three studies [13, 24, 59] that have mentioned cloud and edge computing and cloud storage infrastructure.

c) DBMS: works in this category discussed only storage requirements in terms of a DBMS, which we divided into relational, non-relational, and a combination of those. In terms of relational DBMSs, two conceptual proposals have recommended using a data warehouse [44, 74], [58, 66, 79] mentioned the support of a non-specified relational database, and [15, 76, 84] adopted specific ones (MySQL, SQL Server). Regarding non-relational approaches, most studies adopt open-source databases, such as InfluxDB [57], Hadoop and Jena [29], MongoDB [82], Elasticsearch [86], and Cassandra [81]. Others adopted AWS storage systems such as DynamoDB [77, 78] and S3 [68]. Some studies have recommended non-tabular storage systems, without specifying the specific DBMS, such as graph [67] or spatial databases [80]. Finally, some studies propose the combination of storage systems, such as data warehouse and data lakes [69], and SQL and New SQL databases [72]. Other mentioned combination-specific databases, such as AWS S3 and Redshift [62] to implement a data lake, MariaDB and Elasticsearch [53], data warehouse and Hadoop [54], Hive/Hadoop and data lake [59], SQLite and MongoDB [60], SQL Server and MongoDB [13], and MySQL, Oracle, Hadoop, and MongoDB [63].

6 Trends and opportunities

In the previous section, we summarized the contributions of the systematically selected works that handle data heterogeneity and explicitly propose solutions for one or more identified data management issues: interoperability, integration, data search and discovery, and quality. This section discusses the trends and opportunities identified in our summarization.

No single work proposes an encompassing solution addressing all the data management issues considered in our survey. Integration is the most discussed one (75.4%), which is an expected result due to the central role of the DM component, acting as a bridge between the other DT components. The integration of the data in a centralized and unified repository/data model is combined with mechanisms for ensuring data quality in 17.4% of the studies, with data/semantic interoperability as part of the integration requirements (8.7%), and with functionality for searching and discovering data in 15.2% of them, particularly those that add some level of semantics to the information model. The smart city domain is the one for which the most encompassing solutions were identified, where [85, 89] address integration, interoperability, and data quality, and [88] involves integration, data search, and data quality.

Most of the selected studies address Digital Shadows according to the definition by [2] since they consider the automatic flow of data from the physical to the virtual space only. In addition, we observed that most works assume that knowledge workers, stakeholders and decision-makers are the ultimate consumers of the managed data. In our opinion, this represents an initial maturity level concerning data management as a specific concern in DTs. All surveyed works proposed valuable functionality, methods, data models, and processes for the DM component of a DT.

The next steps in the maturity level of data management in DTs are to understand the role it plays in the closed feedback loop, and the impact the changes reflected in the real space, or insights gained from the models and tools in the other components have in the data managed by DM component. Another issue that needs to evolve is the flow of data in/out of the DM component. Current DTs typically assume the data flows from the Physical Space into the DM component, and from there to the Services/Virtual Space model; however, actually, there are changes and decisions made within the realm of these other components that can impact the data managed. Hence, the flow from the consumers into the DM component, as well as from the DM component back into the Physical world also needs to be assessed for the full implementation of the five-component reference architecture (Fig. 1).

The maturity level is also reflected in the domains for which these solutions were proposed. The I4.0 is the most active field for the DT since the early days. As the interaction between the Physical/Virtual worlds is more understood in this domain, it is natural to expand the concerns to DM issues explicitly. Our survey confirmed that all domains share similar problems concerning velocity, volume, variety, value, and veracity, and moving forward, an interesting effort is to generalize these solutions to provide a domain-agnostic framework.

We observed some common trends and opportunities:

(a) *Reference architectures for Data Management* many works (e.g., [14, 54, 78, 79]) suggest a reference architecture that organizes the data/information in different abstraction layers, with components to perform operations that add quality and value to the raw data at each level (separation of concerns). At the lowest level, the architecture deals with ingesting raw data of different types, sources, data formats, and protocols, possibly with components/operators, to deal with noise and quality issues that are proper to this level. The subsequent layers represent the information model, as the raw data is successively cleaned, transformed, integrated, aggregated, and properly stored. The quality assessment components at this level are often more complex (e.g., models). Many of these architectures additionally encompass a semantic layer, in which the information model is enriched and transformed into a shared knowledge model that represents the characteristics of the domain. It also provides components to access the data/information/knowledge to enable the usage of the managed data, either by humans, models, services, or applications. In addition to layered architectures, other alternative ways of organizing the data/information and its access are proposed, such as the Data Shared Market [89], based on clouds and agents, and the Decision Information Packages, which organizes information according to the different stakeholders [72].

Reference architectures provide patterns, building blocks, interconnections, and a common vocabulary [95]. As there is an opportunity to approximate the DM component to the data and knowledge management functionality of Data Lakes [20, 21], reference architectures can provide a starting point basic structure and best practices for constructing solutions in specific scenarios. It can accelerate the development and implementation of the DM component by reusing existing solutions and providing a basis for governance ensuring their consistency and applicability. It can also lead to the development of general-purpose data management platforms. To ensure the definition of standardized reference architectures and their acceptance, it is crucial that such an effort results from the collaboration between academia and industry. An example is the Digital Twin Capabilities Periodic Table (CPT), [3] proposed by the Digital Twin Consortium. The CPT is a technology-agnostic requirements definition framework aimed at organizations who want to design, develop, deploy, and operate DTs based on use case capability

[3] https://www.digitaltwinconsortium.org/initiatives/capabilities-periodic-table/.

requirements versus the features of technology solutions. It defines Data Management capabilities (referred to as *Data Services*), and it is an example of collaboration initiatives capable of guiding the development of reference architectures of DTs, including data management.

(b) *Industry standards* industry standards were leveraged for different purposes in the selected works. For data exchange, some studies used standards such as OPC UA and interoperable file formats (e.g., WITSML, for the O&G industry). Domain standards also guided data modeling and integration (e.g., BIM, BOM, STEP), providing basic concepts for organizing the data in an information model or explored by accompanying process/methods. Some standards can be useful in different domains (e.g., CFIHOS, VID). Leveraging industry standards is an important step toward generalizing the proposed solutions beyond the specific scope for which a DT is proposed and achieving customizable solutions. It is also important to increase the industry's acceptance to facilitate the deployment of the proposed solutions in real settings.

(c) *Semantic enrichment and ontologies* an ontology formalizes the intended meaning of the terms of a vocabulary according to a certain view of the world [96] and has been leveraged in DTs for distinct DM purposes. The use of standard ontologies representing DT entities (e.g., sensors, power plants, manufacturing) can reduce the semantic heterogeneity, such that different/similar concepts can be understood regardless of the differences in modeling (e.g., [29, 81]). Existing consolidated ontologies can be leveraged for this purpose, such as SOSA (Sensor, Observation, Sample, and Actuator) and SSN (Semantic Sensor Network) for IoT. It can also help establish the correspondence between different industry standards available [13, 15, 63, 80, 81]. It also enables a common understanding and interrelation of concepts in different domains or disciplines, which can support the interoperation of DTs [41] or stakeholders. In summary, ontologies can provide in the context of DTs an organizing view over the domain that helps professionals with distinct technical profiles to navigate and integrate data from several provenances. Several functionalities can be based on semantic enrichment, among them expansion of the search service to semantic characteristics; enrichment of the data transformation and lineage process with semantic metadata; domain inferences based on prior knowledge; improvement in quality assessment; etc.

(d) *Data management across DTs* data management between DTs will become a significant issue, since an organization can rely on more than one DT, sharing information through them. In domains like smart cities, for instance, the interaction between DTs from different subdomains (smart buildings, urban planning) highlights the need for interoperability at all levels (semantic, data, and others) to provide fully integrated and optimized services for citizens. Initial ideas were proposed for the smart cities domain [85] and for I4.0 [41]. Future work will have to address more complex issues considering federations of DTs.

(e) *Cloud/hybrid computing and open-source tools* our results have shown that cloud, edge, and fog computing are a trend in the context of DT. Cloud computing assumes a leading role as it provides several services and resources, ranging from network and communication management to data storage and processing, required to handle the different, often geographically distributed, spaces of a DT. In addition, cloud computing offers scalability, availability, agility, and high speed to deal with data volume issues. [3] contributed with an analysis of cloud-based architectural designs considering data presence

(localized vs. ubiquitous), coordination (centralized vs. hierarchical), and computation (concentrated vs. distributed).
Another trend is the adoption of open-source tools for processing and storing data in the cloud (e.g., Apache Spark, Hadoop, Elasticsearch). Compared to proprietary software, in addition to the low licensing costs, open-source tools promote interoperability with a wide range of software and freedom of customization to meet the needs of the DT infrastructure. Due to community engagement, free software tools are accompanied by extensive supporting documentation, and updates occurs in a faster pace compared to proprietary software. Therefore, we understand that the use of open-source tools, cloud, edge, and fog computing will be increasingly present in DT solutions.

(f) *Standard infrastructures and implementations* we identified extensive use of cloud computing for data processing and storage. This leads to the opportunity of providing core data management functionality for DTs using standard interfaces. An interesting example in the O&G domain is the OSDU[4] (Open Subsurface Data Universe) data management platform. Based on a micro-service architecture, the platform provides standard interfaces for a range of functions covering the life cycle of the data management, from ingestion to use, and cloud providers supply specific implementations. [21] investigated the potential of OSDU functionality for developing the DM component in DTs for that industry. The development and application of standard infrastructures for data management can facilitate the implementation and use of the DM component of DTs, as standard interfaces can simplify the management of the various resources required for DT implementation by providing a unified way for the user.

(g) *Data provenience and blockchain* As raw data follows a big data value chain transformation in the DT, it is also important to keep track of the original sources of the data, the changes made over time, and how it was manipulated [97, 98]. This contributes to transparency and provides context for the results/decisions they generated. It must also be possible to follow the quality and reliability of the data, audit data traces, allowing the replication of procedures, assigning properties or responsibilities (e.g., error), or providing informational context that can be consulted and analyzed [99]. This is an important gap among the selected studies, as only three of them addressed the traceability of the data [65, 72, 73]. A promising technology for this purpose is blockchain [100], which is a shared, immutable ledger that facilitates the process of recording transactions and tracking assets in a business network. [73] envisioned a blockchain-based framework for the I4.0 that enables to follow the whole product life cycle events once data collected from trustworthy sources are recorded in the blockchain, allowing process monitoring, diagnostics, and optimized control. This could be combined with state-of-the-art in data traceability and lineage.

7 Conclusions

In this paper, we reviewed data management in the context of DTs in a systematic and unbiased way, selecting 61 studies that explicitly proposed conceptual or implemented solutions for data interoperability, integration, quality, and search/discovery. Our protocol defined research questions to systematically investigate the domains in which explicit DM solutions for DT were proposed, the DT function according to data usage, the types of data addressed, the

[4] https://osduforum.org/.

proposed solutions according to identified DM issues, and the technological infrastructure leveraged.

Our study allowed us to assess the state of the art of data management in the context of DTs and identify trends and opportunities. We conclude that I4.0 is the most mature domain, most solutions are actually developed in the scope of DSs assuming humans are the ultimate consumers of the information, integration is the prevalent DM issue addressed due to the bridging role of the DM component, and that cloud computing and storage is the key technology leveraged for implementing the DM solutions. We also observed that the difficulty in dealing with velocity, volume, variety, value, and veracity is common to all domains and that the maturity level assumed for the DM component is at an early stage. Among the trends and research opportunities are reference architectures to guide the development and implementation of the DM component through best practices ensuring its consistency and applicability, the adoption of industry standards and ontologies, the interaction among DTs, the use of cloud computing and open-source tools, the development of infrastructures and standard implementations, data provenance mechanisms and blockchain.

The present survey was developed in 2022, and the results reported cover studies between 2014 and 2021. Some relevant, more recent studies might not have been included, but collecting and summarizing state-of-the-art proposals is an ongoing effort. As DTs have become a real trend topic, the number of studies is expected to increase significantly. We acknowledge that collecting new studies as future work is important, particularly because it enables measuring the advances in data management in the context of DTs.

As future work, our SLR could be complemented by considering more recent literature. Other aspects could also be investigated, such as data architecture, data and operations storage, data security and metadata management, content management, and so on. The data management area is broad and includes various fundamental activities for a DT's success; therefore, it is important to have a comprehensive overview of all its aspects.

Acknowledgements Research supported by CAPES and the PETWIN Project (FINEP financing and LIBRA Consortium).

References

1. Conti M et al (2012) Looking ahead in pervasive computing: challenges and opportunities in the era of cyber-physical convergence. Pervasive Mobile Comput 8(1):2–21. https://doi.org/10.1016/j.pmcj.2011.10.001
2. Fuller A, Fan Z, Day C, Barlow C (2020) Digital twin: enabling technologies, challenges and open research. IEEE Access 8:108952–108971. https://doi.org/10.1109/ACCESS.2020.2998358
3. Raptis TP, Passarella A, Conti M (2019) Data management in industry 4.0. IEEE Access 7:97052–97093. https://doi.org/10.1109/ACCESS.2019.2929296
4. Tao F, Zhang H, Liu A, Nee AY (2019) Digital twin in industry: state-of-the-art. IEEE Trans Industr Inf 15(4):2405–2415. https://doi.org/10.1109/TII.2018.2873186
5. Lu H, Guo L, Azimi M, Huang K (2019) Oil and gas 4.0 era: a systematic review and outlook. Computers Industry 111:68–90. https://doi.org/10.1016/j.compind.2019.06.007
6. Wanasinghe TR et al (2020) Digital twin for the oil and gas industry: overview, research trends, opportunities, and challenges. IEEE Access 8:104175–104197. https://doi.org/10.1109/ACCESS.2020.2998723
7. Elayan H, Aloqaily M, Guizani M (2021) Digital twin for intelligent context-aware iot healthcare systems. IEEE Internet Things J 8(23):16749–16757. https://doi.org/10.1109/JIOT.2021.3051158
8. Deng T, Zhang K, Shen Z-JM (2021) A systematic review of a digital twin city: a new pattern of urban governance toward smart cities. J Manage Sci Eng 6(2):125–134. https://doi.org/10.1016/j.jmse.2021.03.003

9. Rao TR, Mitra P, Bhatt R, Goswami A (2019) The big data system, components, tools, and technologies: a survey. Knowledge and Information Systems 60(3):1165–1245. https://doi.org/10.1007/s10115-018-1248-0
10. Jones D, Snider C, Nassehi A, Yon J, Hicks B (2020) Characterising the digital twin: a systematic literature review. CIRP J Manuf Sci Technol 29:36–52. https://doi.org/10.1016/j.cirpj.2020.02.002
11. Curry E (2016) The big data value chain: definitions, concepts, and theoretical approaches, 29–37. Springer International Publishing, Cham
12. Sawadogo PN, Darmont J (2021) On data lake architectures and metadata management. J Intell Inf Syst 56(1):97–120. https://doi.org/10.1007/s10844-020-00608-7
13. Sun S, Zheng X, Villalba-Díez J & Ordieres-Meré J (2020) Data handling in industry 4.0: Interoperability based on distributed ledger technology. Sensors (Switzerland) **20** (11). https://doi.org/10.3390/s20113046
14. Vivi Q L, Parlikad A K, Woodall P, Ranasinghe G D.& Heaton J (2019) Developing a dynamic digital twin at a building level: Using Cambridge campus as case study, 67–75 (ICE Publishing, 2019)
15. Singh S et al (2021) Data management for developing digital twin ontology model. Proc Instit Mech Eng, Part B: J Eng Manuf 235(14):2323–2337. https://doi.org/10.1177/0954405420978117
16. Sahlab N, Kamm S, Muller T, Jazdi N & Weyrich M (2021) Knowledge graphs as enhancers of intelligent digital twins, 19–24 (Institute of Electrical and Electronics Engineers Inc.,)
17. Tao F, Zhang M (2017) Digital twin shop-floor: a new shop-floor paradigm towards smart manufacturing. IEEE Access 5:20418–20427. https://doi.org/10.1109/ACCESS.2017.2756069
18. Semeraro C, Lezoche M, Panetto H, Dassisti M (2021) Digital twin paradigm: A systematic literature review. Computers in Industry 130:87. https://doi.org/10.1016/j.compind.2021.103469
19. Barricelli BR, Casiraghi E, Fogli D (2019) A survey on digital twin: definitions, characteristics, applications, and design implications. IEEE Access 7:167653–167671. https://doi.org/10.1109/ACCESS.2019.2953499
20. Kronberger P, Dabrowski P, Chacon J, & Bangert P (2020) The Digitalization Journey of the Brage Digital Twin, Vol. Day 2 Tue, November 03, 2020 of Procceeding of the SPE Norway Subsurface Conference
21. Correia JB, Rodrigues F, Santos N, Abel M, Becker K (2022) Data management in digital twins for the oil and gas industry: beyond the osdu data platform. J Inf Data Manage 13:3
22. Okoli, C (2015) A guide to conducting a standalone systematic literature review. Commun Assoc Inf Syst 37:43. https://doi.org/10.17705/1cais.03743
23. Moyne J et al (2020) A requirements driven digital twin framework: specification and opportunities. IEEE Access 8:107781–107801. https://doi.org/10.1109/ACCESS.2020.3000437
24. Liu C et al (2022) Digital twin-enabled collaborative data management for metal additive manufacturing systems. J Manuf Syst 62:857–874. https://doi.org/10.1016/j.jmsy.2020.05.010
25. Deren L, Wenbo Y, Zhenfeng S (2021) Smart city based on digital twins. Comput Urban Sci 1(1):1–11
26. Ahmadi-Assalemi G (2020) Digital twins for precision healthcare. Springer, Cham
27. Al-Mekhlal, M. & Khwaja, A. A. A synthesis of big data definition and characteristics, 314–322 (IEEE, 2019)
28. Couto J, Borges O T, Ruiz D D, Marczak S & Prikladnicki R Perkusich A (ed.) A mapping study about data lakes: An improved definition and possible architectures. (ed.Perkusich, A.) Proc. of the 31st International Conference on Software Engineering and Knowledge Engineering, SEKE 2019, Hotel Tivoli, Lisbon, Portugal, July 10-12, 2019, 453–578 (KSI Research Inc. and Knowledge Systems Institute Graduate School, 2019)
29. Jirkovsky V, Obitko M, Marik V (2017) Understanding data heterogeneity in the context of cyber-physical systems integration. IEEE Trans Industr Inf 13(2):660–667. https://doi.org/10.1109/TII.2016.2596101
30. Patel J (2019) Bridging data silos using big data integration. Int J Database Manage Syst 11(3):01–06
31. Rahm E (2016) The case for holistic data integration, Vol. 9809 of Lecture Notes in Computer Science, 11–27 (Springer). https://doi.org/10.1007/978-3-319-44039-2_2
32. Doan A, Halevy A, Ives Z (2012) Principles of data integration. Elsevier, Amsterdam
33. Dong XL, Srivastava D (2015) Big data integration. Morgan & Claypool Publishers, New England
34. Geraci. Ieee standard computer dictionary: A compilation of ieee standard computer glossaries. IEEE Std 610 1–217 (1991). https://doi.org/10.1109/IEEESTD.1991.106963
35. Gürdür D, Asplund F (2018) A systematic review to merge discourses: Interoperability, integration and cyber-physical systems. J Industr Inf Integr 9:14–23. https://doi.org/10.1016/j.jii.2017.12.001
36. Heiler S (1995) Semantic interoperability. ACM Comput Surv (CSUR) 27(2):271–273
37. Chapman A et al (2020) Dataset search: a survey. VLDB J 29(1):251–272. https://doi.org/10.1007/s00778-019-00564-x

38. Sidi F et al. (2012) Data quality: a survey of data quality dimensions, 300–304 (IEEE)
39. Kitchenham B, Charters S (2007) Guidelines for performing systematic literature reviews in software engineering. Keele University and Durham University Joint Report, Tech Rep
40. Dao M S et al (2014) A real-time complex event discovery platform for cyber-physical-social systems, 201–208
41. Platenius-Mohr M, Malakuti S, Grüner S, & Goldschmidt T (2019) Interoperable digital twins in IIoT systems by transformation of information models: a case study with asset administration shell (ICST)
42. Landolfi G et al (2018) Intelligent value chain management framework for customized assistive healthcare devices. Elsevier, Amsterdam
43. Oakes B, Meyers B, Janssens D, Vangheluwe H (2021) Structuring and accessing knowledge for historical and streaming digital twins 2941:1–13
44. Zhang C, Ji W (2019) Digital twin-driven carbon emission prediction and low-carbon control of intelligent manufacturing job-shop 83:624–629. https://doi.org/10.1016/j.procir.2019.04.095
45. Chen S et al (2020) Top-down human-cyber-physical data fusion based on reinforcement learning. IEEE Access 8:134233–134245. https://doi.org/10.1109/ACCESS.2020.3011254
46. Agarwal, P. & McNeill, S. Real-time cleaning of time-series data for a floating system digital twin, Vol. 2019-May (2019)
47. Andia P, & Israel R R (2018) A cyber-physical approach to early kick detection, Vol. 2018-March, 6–8
48. Brackel H U, Macpherson J, Mieting R, & Wassermann I (2018) An open approach to drilling systems automation
49. Al-Ismael M, Al-Turki A & Al-Darrab A (2020) Reservoir simulation well data exchange towards digital transformation and live earth models
50. Zhang Y, Qiu M, Tsai CW, Hassan MM, Alamri A (2017) Health-CPS: healthcare cyber-physical system assisted by cloud and big data. IEEE Syst J 11(1):88–95. https://doi.org/10.1109/JSYST.2015.2460747
51. Alhumud M A, Hossain M A & Masud M (2016) Perspective of health data interoperability on cloud-based medical cyber-physical systems, 1–6 (Institute of Electrical and Electronics Engineers Inc.,)
52. Núñez-Valdez E, Solanki VK, Balakrishna S, Thirumaran M (2020) Incremental hierarchical clustering driven automatic annotations for unifying IoT streaming data. Int J Interact Multim Artif Intell 6(2):15. https://doi.org/10.9781/ijimai.2020.03.001
53. Hussain I, Park SJ (2021) Big-ECG: cardiographic predictive cyber-physical system for stroke management. IEEE Access 9:123146–123164. https://doi.org/10.1109/ACCESS.2021.3109806
54. Hinojosa-Palafox E A, Rodriguez-Elias O M, Hoyo-Montano J A & Pacheco-Ramirez J H (2019) Towards an architectural design framework for data management in industry 4.0, 191–200 (Institute of Electrical and Electronics Engineers Inc.,)
55. Wang T & Cheng L (2021) Large-scale semantic knowledge acquisition and application for cyber-physical-social systems, 282–285 (Institute of Electrical and Electronics Engineers Inc.,)
56. Kong T, Hu T, Zhou T, Ye Y (2021) Data construction method for the applications of workshop digital twin. System 58:323–328. https://doi.org/10.1016/j.jmsy.2020.02.003
57. Zonzini F et al (2020) Structural health monitoring and prognostic of industrial plants and civil structures: a sensor to cloud architecture. IEEE Instrum Measure Magazine 29(9):21–27. https://doi.org/10.1109/MIM.2020.9289069
58. Brecher, C. et al. Gaining IIoT insights by leveraging ontology-based modelling of raw data and digital shadows, 231–236 (Institute of Electrical and Electronics Engineers Inc., 2021)
59. Yu W, Dillon T, Mostafa F, Rahayu W, & Liu Y (2019) Implementation of industrial cyber physical system: challenges and solutions, 173–178
60. Hoos E, Hirmer P & Mitschang B, Kirikova M, Nørvåg K, & Papadopoulos G A (eds) Context-aware decision information packages: An approach to human-centric smart factories. (eds Kirikova, M., Nørvåg, K. & Papadopoulos, G. A.) Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 10509 LNCS of Lecture Notes in Computer Science, 42–56 (Springer International Publishing, 2017)
61. Dai J et al. (2017) Cyber physical power system modeling and simulation based on graph computing, Vol. 2018-January, 1–6
62. Cardoso B B et al. (2021) Data lake architecture for distribution system operator, 1–5 (Institute of Electrical and Electronics Engineers Inc.,)
63. Liu J et al. (2020) The research of ontology-based digital twin machine tool modeling, 2130–2134 (Institute of Electrical and Electronics Engineers Inc.,)
64. Kousi N et al (2019) Digital twin for adaptation of robots' behavior in flexible robotic assembly lines. Proc Manuf 28:121–126. https://doi.org/10.1016/j.promfg.2018.12.020
65. Zhuang C, Gong J, Liu J (2021) Digital twin-based assembly data management and process traceability for complex products. J Manuf Syst 58:118–131. https://doi.org/10.1016/j.jmsy.2020.05.011

66. Lv Q, Zhang R, Sun X, Lu Y, Bao J (2021) A digital twin-driven human-robot collaborative assembly approach in the wake of COVID-19. J Manuf Syst 60:837–851. https://doi.org/10.1016/j.jmsy.2021.02.011
67. Jiang Y, Chen C & Liu X (2021) Assembly process knowledge graph for digital twin, Vol. 2021-Augus, 758–763 (IEEE Computer Society)
68. Hänel A et al (2021) Impact of cyber-physically enhanced manufacturing on the product requirement documentation in high-tech applications 102:210–215. https://doi.org/10.1016/j.procir.2021.09.036
69. Pernici B et al. (2020) AgileChains: agile supply chains through smart digital twins, 2678–2684
70. Blum M, & Schuh G (2017) Towards a data-oriented optimization of manufacturing processes a real-time architecture for the order processing as a basis for data analytics methods, Vol. 1, 257–264 (SciTePress)
71. Gómez-Berbís, J. M. & de Amescua-Seco, A. Sedit: Semantic digital twin based on industrial iot data management and knowledge graphs **1124 CCIS**, 178–188 (2019). https://doi.org/10.1007/978-3-030-34989-9_14
72. Liu J et al (2021) A digital twin-driven approach towards traceability and dynamic control for processing quality. Adv Eng Inf 50:87. https://doi.org/10.1016/j.aei.2021.101395
73. Suhail, S., Hussain, R., Jurdak, R. & Hong, C. S. Trustworthy Digital Twins in the Industrial Internet of Things with Blockchain. IEEE Internet Computing 1–8 (2021). https://doi.org/10.1109/MIC.2021.3059320, arXiv:2010.12168
74. Ansari F, Glawar R, Nemeth T (2019) PriMa: a prescriptive maintenance model for cyber-physical production systems. Int J Computer Integr Manuf 32(4–5):482–503. https://doi.org/10.1080/0951192X.2019.1571236
75. Kirchen I, Schutz D, Folmer J, & Vogel-Heuser B (2017) Metrics for the evaluation of data quality of signal data in industrial processes, 819–826
76. Zhang Q, Yang Z, Duan J, Liu Z, Qin J (2021) Three-dimensional visualization interactive system for digital twin workshop. J Southeast Univ (English Edition) 37(2):137–152. https://doi.org/10.3969/j.issn.1003-7985.2021.02.003
77. Lu Q, Xie X, Parlikad AK, Schooling JM (2020) Digital twin-enabled anomaly detection for built asset monitoring in operation and maintenance. Autom Constr 118:78. https://doi.org/10.1016/j.autcon.2020.103277
78. Lu Q et al (2020) Developing a digital twin at building and city levels: case study of west cambridge campus. J Manage Eng 36:3. https://doi.org/10.1061/(asce)me.1943-5479.0000763
79. Alwan A A, Ciupala M A, Baravalle A, & Falcarin P (2020) HADES: a hybrid anomaly detection system for large-scale cyber-physical systems, 136–142
80. Jouan P, Hallot P (2020) Digital twin: research framework to support preventive conservation policies. ISPRS Int J Geo-Inf 9:4. https://doi.org/10.3390/ijgi9040228
81. Chevallier, Z., Finance, B. & Boulakia, B. C. A reference architecture for smart building digital twin, Vol. 2615 (2020)
82. Acquaviva, A. et al. Forecasting heating consumption in buildings: A scalable full-stack distributed engine. Electronics (Switzerland) **8** (5) (2019). https://doi.org/10.3390/electronics8050491
83. Wu D, Wang H, & Seidu R (2020) Toward a sustainable cyber-physical system architecture for urban water supply system, 482–489 (IEEE)
84. Rybnytska O, Šikšnys L, Pedersen T B, & Neupane B (2020) PGFMU: Integrating data management with physical system modelling, Vol. 2020-March, 109–120 (APA)
85. Kiourtis A, Mavrogiorgou A, Kyriazis D, Maglogiannis I, Themistocleous M (2018) Exploring the complete data path for data interoperability in cyber-physical systems 12(4):339–349. https://doi.org/10.1504/IJHPCN.2018.096714
86. Bujari A, Calvio A, Foschini L, Sabbioni A, & Corradi A (2021) IPPODAMO: a digital twin support for smart cities facility management, 49–54 (Association for Computing Machinery, Inc,)
87. Fan C, Zhang C, Yahja A, Mostafavi A (2021) Disaster city digital twin: a vision for integrating artificial and human intelligence for disaster management. Int J Inf Manage 56:871. https://doi.org/10.1016/j.ijinfomgt.2019.102049
88. Azzam, A. et al. The CitySpin platform: a CPSS environment for city-wide infrastructures, Vol. 2530, 57–64 (2019). https://www.w3.org/TR/sparql11-query/
89. Kasrin N et al (2021) Data-sharing markets for integrating IoT data processing functionalities. CCF Trans Pervasive Comput Inter 3(1):76–93. https://doi.org/10.1007/s42486-020-00054-y
90. Huang W, & Dai W (2017) Knowledge storage and acquisition for industrial cyber-physical systems based on non-relational database, Vol. 2017-January, 6671–6676
91. Wang, Y. & Zhou, X. Spatio-temporal semantic enhancements for event model of cyber-physical systems, 813–818 (2014)

92. Gifty R, Bharathi R, Krishnakumar P (2020) Faulty-data detection and data quality measure in cyber-physical systems through Weibull distribution. Computer Commun 150:262–268. https://doi.org/10.1016/j.comcom.2019.11.036
93. Proper HA, Bork D, Poels G (2021) Towards an ontology-driven approach for digital twin enabled governed IT management 2941:14
94. Sha K, Zeadally S (2015) Data quality challenges in cyber-physical systems. J Data Inf Qual 6:2. https://doi.org/10.1145/2740965
95. Cloutier R et al (2010) The concept of reference architectures. Syst Eng 13(1):14–27. https://doi.org/10.1002/sys.20129
96. Guarino N (1998) Formal ontology and information systems, 3–15. IOS Press, Amsterdam
97. Herschel M, Diestelkämper R, Lahmar HB (2017) A survey on provenance: what for? what form? what from? VLDB J 26(6):881–906
98. Pérez B, Rubio J, Sáenz-Adán C (2018) A systematic review of provenance systems. Knowl Inf Syst 57(3):495–543
99. Simmhan YL, Plale B, Gannon D (2005) A survey of data provenance in e-science. SIGMOD Rec 34(3):31–36. https://doi.org/10.1145/1084805.1084812
100. Zheng Z, Xie S, Dai H, Chen X, Wang H (2018) Blockchain challenges and opportunities: a survey. Int J Web Grid Serv 14:352–375

Jaqueline B. Correia is currently PhD candidate at the Institute of Informatics, Federal University of the Rio Grande do Sul (UFRGS), Brazil. She completed Master's degree in Computer Science from UFRGS in 2022. She is currently member of the Petwin research group. Her current research interests include data management, data integration, data science, data lineage in digital twin systems.

Mara Abel has a degree in Geology (1982), and MSc. and Ph.D. degrees in Computer Science (1988 and 2001), all from the Federal University of the Rio Grande do Sul - -UFRGS. She is a Full Professor at the Informatics Institute - -UFRGS, the leader of the research group on Computing Systems for Petroleum Exploration and Production, and vice- leader of the Artificial Intelligence Group. Mara has authored nearly 120 research articles published in journals, books, and conference proceedings. She coordinated several projects on knowledge management and knowledge engineering for building ontologies, especially those applied to Petroleum Geology. She is the co-founder of ENDEEPER, a spin-off of her research projects. She is very involved with creating Start-ups, acting as the Head of the IT Incubator at UFRGS, and a member of the Board that led to the creation of Zenit Technological Park at UFRGS. She is a member of the Brazilian Computer Society (SBC), the International Association of Ontology Applications (IAOA), and the Professional Petroleum Data Management Association (PPDM).

Karin Becker is an Associate Professor in the Department of Computer Science at the Institute of Informatics, Federal University of the Rio Grande do Sul (UFRGS), Brazil. With a PhD from Namur University in Belgium (1993) and a master's degree in Computer Science from UFRGS (1989), she has gained extensive experience in both academic and industry research, specifically in the areas of database, data mining, and data science. Karin has coordinated approximately 20 research projects with external funding and supervised nearly 50 graduate students. She has authored almost 140 research articles published in journals, books, and conference proceedings. Karin's current research focuses on data management in digital twin systems and data mining applications in social media data. She acts as a reviewer for various conferences and journals in her domain expertise and has chaired conferences and workshops in the database and data science areas. She was also an invited speaker at different events. Karin is a Senior Member of the Brazilian Database Board and has been distinguished with prizes for best research papers. She is a member of IEEE and the Brazilian Computer Science Society.