UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

JAQUELINE BITENCOURT CORREIA

Data Management in Digital Twins: a Systematic Literature Review

Thesis presented in partial fulfillment of the requirements for the degree of Master of Computer Science

Advisor: Profa. Dra. Karin Becker

Porto Alegre
December 2022

CIP — CATALOGING-IN-PUBLICATION

Correia, Jaqueline Bitencourt

Data Management in Digital Twins: a Systematic Literature Review / Jaqueline Bitencourt Correia. – Porto Alegre: PPGC da UFRGS, 2022.

91 f.: il.

Thesis (Master) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2022. Advisor: Karin Becker.

1. Digital Twin. 2. Data management. 3. Big Data. 4. Systematic Literature Review. I. Becker, Karin. II. Título.

ACKNOWLEDGMENTS

First, I thank God for everything. Without God's love and mercy, I would not have come here.

I thank my family, especially my mother, siblings, and husband, for their support, encouragement, and unconditional understanding during this period.

I thank my excellent advisor, Prof. Dr. Karin Becker, for the opportunities and teachings during these three years.

This work was developed in the scope of the Petwin Project (PeTWIN.org), funded by the Brazilian Federal Agency for Innovation (FINEP) and the Libra Consortium for petroleum production. I deeply appreciate the financial support received for this research. I was also partially financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Finally, I thank everyone from INF-UFRGS, especially the other teachers, for their teachings and contributions throughout this academic journey.

ABSTRACT

The Internet of Things (IoT), personal and wearable devices, and 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. 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. DTs and big data are mutually reinforcing technologies since huge volumes of data representing the physical/virtual worlds are collected, transformed, and generated through models to aggregate value to the business. Modern DTs follow a five-component architecture, which includes a Data Management (DM) component that bridges a physical system, a mirrored virtual one, and services components. However, there is no clarity on the functionality required for the DM component. This work presents a Systematic Literature Review on DM issues and proposed solutions in the DT context. We analyzed DM under the big data value chain activities, highlighting key issues to be addressed: data heterogeneity, interoperability, integration, data search/discovery, 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 technical infrastructure. The compilation of these solutions sheds light on the functionality of the DM component in a DT, trends, and opportunities. 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.

Gestão de dados em Gêmeos Digitais: uma Revisão Sistemática da Literatura

RESUMO

A Internet das Coisas, dispositivos pessoais e vestíveis e avanços contínuos nas técnicas de coleta de dados aumentaram significativamente a quantidade de dados relevantes que podem ser aproveitados para aplicativos inovadores orientados a dados em tempo real. Os gêmeos digitais (GDs) são representações virtuais de objetos físicos, que são totalmente integrados e nos quais a troca automática de dados ocorre de maneira bidirecional. GDs e Big Data são tecnologias que se reforçam mutuamente, uma vez que grandes volumes de dados que representam os mundos físicos/virtuais são coletados, transformados e gerados por meio de modelos para agregar valor ao negócio. Os GDs modernos seguem uma arquitetura de cinco componentes, que inclui um componente de gestão de dados que faz a ponte entre o sistema físico, o componente virtual espelhado, o componente dos serviços e as conexões. No entanto, não há clareza sobre a funcionalidade necessária para o componente de gestão de dados. Este trabalho apresenta uma revisão sistemática da literatura sobre questões de gestão de dados e soluções propostas no contexto do GD. Analisamos o componente de gestão de dados sob a perspectiva das atividades da cadeia de valor de Big Data, destacando os principais problemas a serem abordados: heterogeneidade de dados, interoperabilidade, integração, pesquisa/descoberta de dados e qualidade. Além de pesquisar soluções existentes para lidar com esses problemas, contextualizamos-os no domínio e na função para os quais o GD foi proposto, o tipo de dados tratados e a infraestrutura tecnológica. A compilação dessas soluções lança luz sobre a funcionalidade do componente de gestão de dados em um GD, tendências e oportunidades. Nossas principais descobertas revelaram que o nível de maturidade assumido para o componente de gestão de dados está em um estágio inicial. As soluções mais maduras foram propostas para o domínio da indústria, e muitas delas assumem os seres humanos como os consumidores finais das informações. A integração de dados é o problema de gestão de dados mais abordado devido à função de ponte do componente de gestão de dados, e a computação em nuvem é a principal tecnologia de implementação. Entre as oportunidades de pesquisa estão as arquiteturas de gerenciamento de dados de referência, a adoção de padrões e ontologias do setor, a interoperabilidade entre GDs distintos, o desenvolvimento de implementações de padrão agnóstico e mecanismos de proveniência de dados.

Palavras-chave: Gêmeo Digital, Gestão de dados, Big Data, Revisão Sistemática da Literatura.

LIST OF ABBREVIATIONS AND ACRONYMS

ACM	Association for Computing Machinery Digital Library
AI	Artificial Intelligence
API	Application Program Interface
BIM	Building Information Modeling
BOM	Bill of Materials
CPS	Cyber-Physical Systems
CRM	Customer Relationship Management
CSV	Comma Separated Values
DBMS	Database Management System
DLs	Digital Libraries
DM	Data Management
DS	Digital Shadow
DT	Digital Twin
ERP	Enterprise Resource Planning
ETL	Extract, Transform, Load
FCA	Formal Concept Analysis
FMU	Functional Mockup Unit
GUI	Graphical User Interface
GPS	Global Positioning System
GPU	Graphic Processing Unit
HCA	Hierarchical Clustering Algorithm
IEEE	Institute of Electrical and Electronics Engineers
IoT	Internet of Things
IIoT	Industrial Internet of Things

I4.0	Industry 4.0
JSON	JavaScript Object Notation
LDA	Latent Dirichlet Allocation
MES	Manufacturing Execution Systems
NoSQL	Not Only SQL
O&G	Oil and Gas
OP	One Petro Digital Library
OPC-UA	Open Platform Communications Unified Architecture
OSDU	Open Subsurface Data Universe
OWL	Ontology Web Language
PDF	Portable Document Format
PICO	Population, Intervention, Comparison, Outcome
PICOC	Population, Intervention, Comparison, Outcome, Contex
POC	Proof-of-concept
QA	Quality Assessment
QRcode	Quick Response Code
RDF	Resource Description Framework
RFID	Radio Frequency Identification
RQ	Research Question
SCADA	Supervisory Control And Data Acquisition
SCO	Scopus Digital Library
SLR	Systematic Literature Review
SOA	Service Oriented Architecture
SOSA	Sensor, Observation, Sample, and Actuator
SQL	Standard Query Language
SSN	Semantic Sensor Network

STEP	STandard for the Exchange of Product model data
S@D	Science Direct Digital Library
UK	United Kingdom
USA	United States of America
WITSML	Wellsite Information Transfer Standard Markup Language
WOS	ISI Web of Science Digital Library
WSN	Wireless Sensor Network
XLSX	Office Open XML spreadsheet
XML	eXtensible Markup Language

LIST OF FIGURES

Figure 2.1 Five components DT architecture 24
Figure 2.2 Big Data Value Chain. 27
Figure 2.3 A Conceptual Framework of Data Quality 33

Figure 5.1 Study selection and quality assessment 44
Figure 5.2 Quality questions average score 45
Figure 5.3 Number of papers per year 46
Figure 5.4 Number of institutions per country 46
Figure 5.5 Data collection and processing methods 54

LIST OF TABLES

Table 2.1 Comparison between SLR and Systematic Mapping 17

Table 3.1 Research questions of related literature reviews 37
Table 3.2 Literature reviews comparison 38

Table 4.1 Research questions and respective motivations........ 39
Table 4.2 Selected digital libraries 41
Table 4.3 Quality assessment criteria........ 42
Table 4.4 Extraction form 43

Table 5.1 Studies grouped based on scientific vehicle and institution type 46
Table 5.2 Selected studies 47
Table 5.2 Selected studies 48
Table 5.2 Selected studies 49
Table 5.3 Data consumption methods 52
Table 5.4 Integration approaches 56

Table A.1 Quality assessment score of selected studies 83

CONTENTS

1 INTRODUCTION..13
2 THEORETICAL FOUNDATION..16
2.1 Types of Systematic Literature Reviews..16
2.2 Guidelines for developing an Systematic Literature Review..........................18
2.2.1 Planning phase..18
2.2.2 Conducting phase...20
2.2.3 Reporting phase...21
2.3 DT Definitions..22
2.4 DTs Application domains..23
2.4.1 Industry 4.0..24
2.4.2 Smart cities..25
2.4.3 Healthcare...25
2.5 Data Management Issues in Digital Twins..26
2.5.1 Data heterogeneity..28
2.5.2 Data integration..29
2.5.3 Interoperability...30
2.5.4 Data search and discovery..31
2.5.5 Data quality..32
3 RELATED WORK...35
3.1 Literature Reviews about Digital Twins...35
3.2 Final considerations and comparison..37
4 SYSTEMATIC LITERATURE REVIEW: PROTOCOL DEFINITION.....................39
4.1 Objective and Research Questions..39
4.2 Search Strategy...40
4.3 Study Selection..41
4.4 Quality Assessment..42
4.5 Data Extraction and Summary...42
5 RESULTS..44
5.1 Study Selection..44
5.2 Quality Assessment..44
5.3 Data Extraction Results..45
5.4 RQ1: For which domains the DT solutions were proposed?............................49
5.5 RQ2: Under the perspective of data usage, for which functions were the DTs proposed?...51
5.6 RQ3: What types of data the proposed solutions consider?.............................52
5.7 RQ4:What solutions were proposed for the DM issues addressed?..................54
5.7.1 Interoperability...54
5.7.2 Integration...56
5.7.3 Data Search and Discovery...58
5.7.4 Data Quality...59
5.8 RQ5: What kind of technological infrastructure is considered?.......................60
6 DISCUSSIONS...63
6.1 Data Management Issues...63
6.2 Trends and Opportunities..64
7 CONCLUSION AND FUTURE WORKS..69
REFERENCES..73
APPENDIX A — QUALITY ASSESSMENT TABLE OF SELECTED STUDIES 83

APPENDIX B — RESUMO EXPANDIDO 85
B.1 Introdução e contribuições da dissertação 85
B.2 Metodologia 88
B.3 Resultados e conclusão 88

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 (CONTI et al., 2012). 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 (FULLER et al., 2020).

DTs are at the core of disruptive innovations in diverse areas (RAPTIS; PASSARELLA; CONTI, 2019). In Smart Manufacturing (Industry 4.0 - I4.0), DTs can cover all product life-cycle phases, including design, planning, assembly, and workshop optimization (TAO et al., 2019; FULLER et al., 2020). 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 (LU et al., 2019; WANASINGHE et al., 2020). 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 (ELAYAN; ALOQAILY; GUIZANI, 2021). Smart City DTs aim to improve the efficiency and sustainability of logistics, energy consumption, urban planning, disaster management, among others (DENG; ZHANG; SHEN, 2021).

Big Data and DTs are mutually reinforcing technologies (RAO et al., 2019) 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 (TAO et al., 2019; JONES et al., 2020). 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 (CURRY, 2016), which includes activities for data acquisition, analysis, storage, curation, and usage. Data lakes are a trending topic to address Big Data issues (SAWADOGO; DARMONT, 2021).

Different DT frameworks are proposed in the literature (WANASINGHE et al., 2020). 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 (SUN et al., 2020; VIVI et al., 2019; SINGH et al., 2021; SAHLAB et al., 2021).

The five-component DT architecture (TAO; ZHANG, 2017) 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 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 (TAO; ZHANG, 2017), or scattered in other components of the DT.

This work 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 (SEMERARO et al., 2021; BARRICELLI; CASIRAGHI; FOGLI, 2019; FULLER et al., 2020; JONES et al., 2020). Regarding data management, (RAPTIS; PASSARELLA; CONTI, 2019) presents a survey from Manufacturing Automation and Networking Computing perspective, outlining architectural designs based on data-related factors (presence, coordination, and computing). Although (TAO et al., 2019) 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 (KRONBERGER et al., 2020; CORREIA et al., 2022), 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 (CURRY, 2016). An SLR is defined in (OKOLI, 2015) 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 SLR are:

- The fine-grained analysis of DM under the activities within the Big Data value chain, highlighting 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 remaining of this document is organized as follows. Chapter 2 describes the theoretical foundation on types of systematic literature reviews, DT definitions, active domains and summarizes the key DM issues derived from Big Data value chain activities. Chapter 3 describes the literature reviews in the DT area. Chapter 4 outlines the protocol developed for the SLR, including the motivation for the selected research questions and the quality assessment. Chapter 5 details the study selection, the quality assessment process and answers each research question defined in the protocol. Chapter 6 summarizes the main trends and opportunities identified. Finally, Chapter 7 draws conclusions and outlines future work.

2 THEORETICAL FOUNDATION

This chapter describes the aspects and compares systematic literature reviews and systematic mappings. It also describes other relevant subjects to this work related to DT definitions and application domains. Finally, are described data management aspects, such as data integration, interoperability, data search, and data quality.

2.1 Types of Systematic Literature Reviews

Literature reviews are an overview of a specific subject (DENNEY; TEWKSBURY, 2013). Their aim is to summarize all the main concepts and terminologies related to a specific topic in a single study, allowing one to identify trends and specific areas of the topic in question that need further research (ROWLEY; SLACK, 2004). Therefore, a literature review is very useful in summarizing the state of the art of a specific subject, making it easier to understand and derive knowledge about it.

In addition to identifying research gaps, a literature review also allows one to evaluate and compare certain theories by examining relationships between variables. It provides a conceptual basis for creating new conceptual models or theories, among many other uses. The method or type of review may vary according to the objective of the literature review. Kitchenham and Charters (2007) evaluate two types of secondary study: SLR, and systematic mapping study (also known as scope study).

An SLR is a methodological and formal way to identify, evaluate and analyze primary studies to answer a specific research question (STAPLES; NIAZI, 2007). According to Kitchenham and Charters (2007), an SLR is a form of secondary study that has a well-defined methodology to identify, evaluate and interpret all available evidence related to a specific research question and that is to some extent repeatable. The following goals can justify the development of an SLR (KITCHENHAM; CHARTERS, 2007): (i) to summarize the existing evidence on a method, treatment, concept, theory, or approach exposing limitations and benefits; (ii) to identify any research gaps showing to the academia which themes need further research and new solutions, and (iii) to provide a basis for new research activities and also to examine to what extent empirical evidence is supported or contradicted.

Unlike a traditional literature review, an SLR follows an explicit protocol to identify primary studies and analyze them in a thorough and unbiased manner (MACDONELL et al., 2010). The process is typically divided into three main phases: planning, conducting, and reporting. In a few words, the planning phase involves identifying the need for the SLR, defining the research questions and developing a review protocol to select the studies, extract the necessary data to answer the questions, and synthesize the data so that the questions can be answered. In the conducting phase, the activities are carried out according to the protocol. The final reporting phase aims at producing a document to efficiently present and communicate the results. In Section 2.2, we detail the process outlined in (KITCHENHAM; CHARTERS, 2007).

Another type of secondary study is systematic mapping. It shares in common with an SLR the adoption of a methodology for the collection and selection of studies to reduce bias, but differs in the objective (PETERSEN; VAKKALANKA; KUZNIARZ, 2015). A systematic mapping aims to provide a broad view of a research area and identify which evidence is available, and indicate the number of evidence (KITCHENHAM; CHARTERS, 2007). Therefore, systematic mapping aims to find the topics covered in the literature, classify them, and categorize them to structure a research area, resulting in the visual summary.

The process for a systematic mapping also includes the definition of research questions, keywords (search string), choice of databases and the collection of studies, the selection of articles based on inclusion and exclusion criteria, data extraction to classify the articles (thematic analysis), analysis of studies and summarization of results, and finally, the writing of the report. It is important to highlight that all phases of this process are performed more broadly. Since the objective is to abstract a high-level overview, there is no depth in the analysis (KITCHENHAM; CHARTERS, 2007).

Table 2.1 – Comparison between SLR and Systematic Mapping

Features	**Systematic Literature Review**	**Systematic Mapping**
Focus of the Review	Identify, analyze and interpret all available evidence related to a specific RQ	Identify and classify what evidence is available (broad review) in a specific topic of area
	Identify best practices based on empirical evidence	Establish the state of evidence
Research Questions	Narrow RQs	Broader RQs
	Specific RQs	Multiple RQs
	Consider population; intervention; comparison, and outcomes (PICO)	Consider only population and intervention
Methods for Searching	Search string highly focused	Search string less highly focused
Methods for Selection	Generally few studies are considered	A large number of studies are considered (broad coverage)
	The studies are evaluated in details	The studies are not evaluated in details
Methods for data extraction	The primary studies are assessed regarding their quality (the main goal is to establish the state of evidence)	The primary studies are not assessed regarding their quality
	Include data extraction procedures	Much broader (classification and categorization stage)
	It is a time-consuming task	It is not a time-consuming task
Synthesis	Include depth analysis techniques, e.g., meta-analysis and narrative synthesis	Include no-depth analysis techniques, e.g., total and summaries
Dissemination of the results	Higher importance for practitioners (relevant to industry)	May be more limited, the aim is to influence the future of the research in a specific topic

Source: (NAPOLEÃO et al., 2017)

Table 3.2 compares the characteristics of these two forms of secondary study. As we can see, although SLRs and mapping studies have a methodology with the same phases, they differ in the focus granularity of the review, research issues, research methods, selection and extraction of data, as well as in synthesis and dissemination techniques of results. We chose to develop our literature review as an SLR rather than a systematic mapping due to the need to investigate specific aspects of data management in DTs (integration, data interoperability, quality, and data search) using a methodology that minimizes bias and allows this analysis in a fine granular level. Therefore, we consider an SLR the most appropriate investigation methodology.

2.2 Guidelines for developing an Systematic Literature Review

In this work, we adopted the guidelines for developing an SLR described in (KITCHENHAM; CHARTERS, 2007). Their method was originally proposed for reviews in the Software Engineering field, but it has been widely the foundation for SLR in Computing Science in general. Below, we outline the activities of each phase proposed for the process.

2.2.1 Planning phase

The five stages associated with planning the review are: identification of the need for a review, commissioning a review, specification of research questions, development of review protocol, and its evaluation. We detail each of these stages below:

- **Identification of the need for a review:** the first stage is to evaluate whether there is a real need for the development of an SLR. It is important to seek by any systematic reviews inherent to the subject or phenomenon to be investigated. Khan et al. (2001) suggest a list of questions that should be answered before starting a systematic review.
- **Commissioning a review:** commissioning refers to the development of a formal document that describes and contextualizes the SLR to be developed. From this need, organizations can hire a research group to execute this work. This stage is only necessary when the SLR needs to be performed by an outsourced organization.
- **Specifying the research question(s):** research questions drive the methodology of an SLR. Guidelines define some types of questions that are frequently used for medical studies (GLASZIOU et al., 2000). However, in other areas, such as soft-

ware engineering or computing science in general, the types of questions are not yet clear. In the medical area, it is very common to structure research questions through PICOC (Population, Intervention, Comparison, Outcome, Context) criteria. Population refers to the population affected by the intervention, Intervention refers to alternative treatments, Comparison refers to the comparison of intervention treatments, Outcomes refer to the clinical and economic factors that will be used to compare the interventions, and Context refers to the context to which the intervention is delivered. Kitchenham, Mendes and Travassos (2006) adapted this structure to the software engineering field and proposed PICO (Population, Intervention, Comparison, Outcome). However, these criteria for structuring research questions are just an attempt to assist in this process, and depending on the subject or context of SLR; they are not always suitable (JØRGENSEN, 2007).

- **Developing a review protocol:** the methods used to perform a particular SLR are defined at this stage. Defining such a protocol is critical to minimizing the researchers' bias. The protocol of an SLR consists of all components of a review and some additions, such as the logic of the research, the questions that the research aims to answer, and strategies to identify and select relevant studies, and extract data from them.

  The search strategy is defined by the search string, and the sources to be sought for primary articles, including digital libraries, journals, and conferences. For selecting articles, the protocol should define inclusion and exclusion criteria, how these criteria should be applied, how many researchers should perform this task, and how the assessment disagreements will be resolved.

  The protocol should include a quality assessment, including the criteria to be used and how to value them based on the goal for such an evaluation. There are different motivations that justify the quality assessment of primary studies, such as refinement of the study selection process, to investigate whether quality differences provide an explanation for differences in studies results, as a means of weighing the individual importance of each study, to interpret discoveries, among others. There is not a standard set of questions, and the protocol can define its own criteria according to the context of the study and goal (FINK, 2019).

  Researchers should finally define how and what data will be extracted from the selected articles, define the synthesis strategy and the techniques to be used, as well as define the dissemination strategy by choosing the conference or journal to which the work will be submitted for publication. It is recommended to establish a schedule for the execution of each activity.

- **Evaluating the review protocol:** evaluating the SLR protocol is an important task that should be discussed between participants and stakeholders. However, this is an optional task within the SLR process.

2.2.2 Conducting phase

After the planning phase has been defined and agreed upon among all participants, the conducting phase of the SLR can be started in practice. There are three stages that consist in executing the protocol, and which are described below:

- **Identification of research:** the main objective is to find as many primary articles relevant to the research topic based on an impartial search strategy. Therefore, it is important that the definite search string contains synonyms, abbreviations, and alternative spellings connected by logical operators (AND, OR). To minimize a systematic bias in the search, one can identify non-published results by examining gray literature and conferences or consulting experts and researchers. Control articles, identified manually, can help adjust the search string to automatically retrieve the most relevant articles. In addition, the search for articles should not be performed automatically only. It can be considered a manual search because the most important is to include the most relevant articles in SLR.

- **Selection of primary studies:** once potentially relevant primary studies are obtained according to the search string, their actual relevance must be assessed using the inclusion/exclusion criteria. The selection of studies is a process that can be carried out throughout several iterations, throughout a process of refinement, in order to handle the volume of articles. For instance, only specific parts of the article can be considered in an initial screening (e.g., title, abstract), deferring the full reading to a later moment when there is more evidence about the suitability of the study. Quality assessment can be used as a further criterion for selection to maintain only the relevant articles for the research. Interestingly, the selection of studies is performed by more than one person: disagreements must be discussed and resolved in order to generate more consistency concerning possible uncertainties.

- **Study quality assessment:** although the concept of quality can be subjective, quality assessment is an essential tool to minimize bias and systematic errors regarding internal and external validity. Each article should be assessed according to the verification list of the protocol according to the motivation of such an assessment.

- **Data extraction and monitoring:** at this stage, the data extraction form is created and accurately filled by researchers. The form is designed to collect all the necessary information that addresses the review issues and the study quality criteria. Numerical data are essential because meta-analysis (statistical techniques) helps summarize and integrate the results found in primary studies. The data extraction form must contain at least the researcher's name, data extraction date, article data (title, authors, journal, institution, etc.), and a space for additional notes.

- **Data synthesis:** data synthesis involves the collection and summary of the results of selected primary studies. Synthesis may be descriptive (non-quantitative), quantitative (meta-analysis), or both. It is usually possible to complement a descriptive synthesis with a quantitative synthesis. Based on the information extracted from the primary studies, it is important to identify and tabulate the results to detect patterns or disparities between studies. Quantitative data should also be tabulated and synthesized comparably.

2.2.3 Reporting phase

The main objective of this phase is to effectively spread the results obtained from the review. The stages of the report phase are detailed below:

- **Specifying dissemination mechanisms:** the last phase of SLR involves writing and disseminating the review results in the communication vehicles defined in the planning phase. It is crucial to communicate the results of an SLR effectively. Usually, forms of dissemination are through journals and scientific conferences. However, it can also occur through non-scientific magazines and newspapers, web pages, posters, popular and specialized press, and white papers.

- **Formatting the main report:** usually, SLRs are written in the technical report format or a section of a doctoral thesis, or the scientific article format of a journal or conference. SLRs written in the scientific article format may have size restrictions;

therefore, it is essential to organize the content properly, maintaining the rigor and validity of the study.

- **Evaluating the report:** this stage is aimed at evaluating and reviewing the report. This task becomes necessary when the result of the SLR is in the format of a technical report, taking into account that it is not common the peer review. Ideally, experts and researchers with experience should evaluate the report and can use the quality verification lists for SLRs.

2.3 DT Definitions

There is no consensus on the definition of the term “Digital Twin", and many works have contributed to a better comprehension of this concept. Barricelli, Casiraghi and Fogli (2019) compiled 29 different definitions out of 75 systematically collected studies. These 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. Semeraro et al. (2021) 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, behavioral modeling of the physical space and services provided. Details of these works are presented in Chapter 3.

According to Moyne et al. (2020), a DT must meet the following requirements: (a) it is some level of a 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; (f) it uses data to maintain some type of synchronization with its real counterpart, where typically this data is collected in an operational environment.

Based on the manual/automatic data flows between Physical and Digital objects, Fuller et al. (2020) distinguish 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-versus. In a DT, the Digital and Physical objects are fully integrated in both directions, such that a change in one leads to a change in the other.

Wanasinghe et al. (2020) conceptualize DTs in terms of reference architectures. Figure 2.1 presents the five-component architecture proposed in (TAO; ZHANG, 2017). 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.

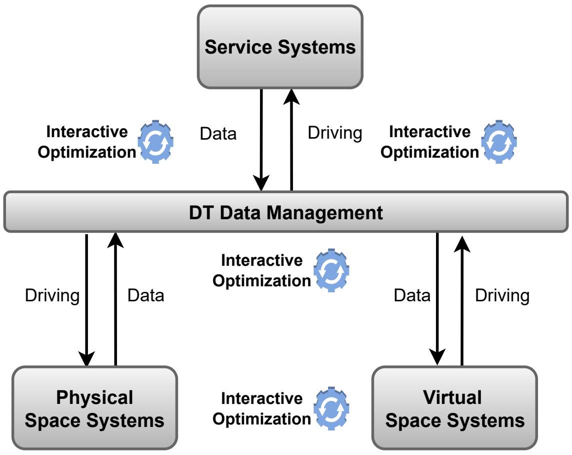
In this SLR, we consider digital twins that meet the requirements proposed in (MOYNE et al., 2020). We also consider both DTs and DSs according to the distinction introduced in (FULLER et al., 2020), since the closed loop represents a stage of maturity that has not been reached by related work yet, and does affect the required data management functionality for the DM component. We also adopt the five-component architecture proposed in (TAO; ZHANG, 2017) since it highlights a functional component responsible for data management.

There are many uncertainties and unawareness about the data management component of a DT. Its role and functionalities are issues that need to be delimited, and we aim to understand more about these issues through this SLR. As a starting point, we understand that the functions of the DM component can be approximated to the data and knowledge management functionality in Data Lakes (KRONBERGER et al., 2020). In the light of Big Data value chain activities (CURRY, 2016), we detail in Section 2.5 the key DM issues to be handled by the DM component of a DT.

2.4 DTs Application domains

A DT can explore and generate descriptive, diagnostic, predictive, and prescriptive analyses by transferring behavior from the physical to the virtual world. 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 enable right-time decisions that

Figure 2.1 – Five components DT architecture



Source: adapted from (TAO; ZHANG, 2017)

positively impact the physical space (RAPTIS; PASSARELLA; CONTI, 2019; SEMERARO et al., 2021; JONES et al., 2020). In this section, we briefly describe the potential of DT applications in the domains that most exploit their benefits to I4.0, Healthcare, and Smart Cities. Our goal is to highlight that, despite the idiosyncrasies of each field, the data management challenges derived from the volume, variety, veracity, speed, and value are very similar.

2.4.1 Industry 4.0

I4.0 encompasses many industry sectors and is based on three fundamental components: the IoT, cloud computing, and cyber-physical systems (CPS) (XU; XU; LI, 2018). Smart manufacturing is a subdomain of I4.0, in which the requirements are directly related to manufacturing, production, and assembly processes to reduce costs and manufacturing time. Organizations such as Bridgestone[1], Boeing[2] and the Change2Twin European initiative[3] report successful use cases.

In the energy and the O&G industry, DTs have the potential to optimize exploration processes, reduce the environmental impact and safety risks, and improve reservoir simulation models, well drilling, and production processes (SIRCAR et al., 2022). DTs have become a key technology for anticipating failures, determining maintenance needs, and minimizing losses. In the electricity production and distribution sector (e.g., Smart Grid, Power Grid, Wind Farm), DTs play a key role in predictive maintenance, and health

[1] https://www.bridgestone.com/corporate/news/2019121901.html
[2] https://www.boeing.com/features/innovation-quarterly/feb2019/btj-global.page
[3] https://www.change2twin.eu/digital-twin/

monitoring of the assets (SIVALINGAM et al., 2018), as demonstrated by Siemens[4].

2.4.2 Smart cities

The goal of a smart city is to provide better services and infrastructure to its citizens. The concept of DT can help visualize all the city's aspects and optimize the city's planning, management, and services (DEREN; WENBO; ZHENFENG, 2021). Many elements relevant to the town (e.g., traffic) can be collected using sensors (e.g., IoT, personal devices). The main challenge is consolidating all gathered data and city infrastructure to achieve a unified view. The application of DT in a smart city enables planners to collect data from sensors and use them for creating and simulating scenarios to analyze people's flow, understand how the traffic flows, and suggest alternative ways for the citizens. Through the socioeconomic and localization data, the city's leaders can make decisions efficiently and focus on providing further resources to needy communities. They can also monitor the quality of services, such as water supply. Concrete examples are Virtual Singapore[5], Seoul S-Map[6] and 51World[7].

2.4.3 Healthcare

DTs in healthcare can take disease prevention, early detection of diseases, and patient care improvement to the next level. It can help resolve issues such as lack of convergence between physical and medical information systems and the absence of interactive patient life cycle monitoring (AHMADI-ASSALEMI et al., 2020). It can also bring new opportunities to the health domain since it can provide decision support for personalized treatments based on the patient's data, minimizing mistakes and ineffective therapies, and paving the way to precision medicine. Hospitals can use the DT concept to optimize their processes, reduce the patient's wait time for attending, better manage risk cases and provide further control and knowledge about their resources and processes, enabling new strategies (AHMADI-ASSALEMI et al., 2020). Some interesting applications are

[4] https://new.siemens.com/global/en/products/energy/energy-automation-and-smart-grid/electrical-digital-twin.html

[5] https://www.nrf.gov.sg/programmes/virtual-singapore/

[6] https://smartcity.go.kr/en/

[7] https://www.unrealengine.com/en-US/spotlights/51world-creates-digital-twin-of-the-entire-city-of-shanghai

Carestation Insights[8] and the HeartModel[9]. Healthcare data derives from various sources and representations, such as sensors, wearable devices, monitoring equipment, images (e.g., x-ray, MRI, CT, ultrasound), texts (e.g., exam results, clinical and biomedical notes) and corporate systems. This wide heterogeneity of data requires different techniques for collecting and analyzing large volumes of data and the ability to deal with (near) real-time data and integration requirements.

2.5 Data Management Issues in Digital Twins

Qi and Tao (2018) compare DT and Big Data and note that both concepts share the same technologies. However, while the former is more related to technologies on cyber-physical integration such as simulation, virtual reality, augmented reality, and CPS, Big Data is more related to data technologies such as cloud computing, data cleaning, data mining, machine learning and etc. Therefore, DTs and Big Data are complementary concepts.

The term “Big Data" is used to label data management according to different attributes. Originally it was associated with three key properties (CURRY, 2016): 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 (AL-MEKHLAL; KHWAJA, 2019), 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.

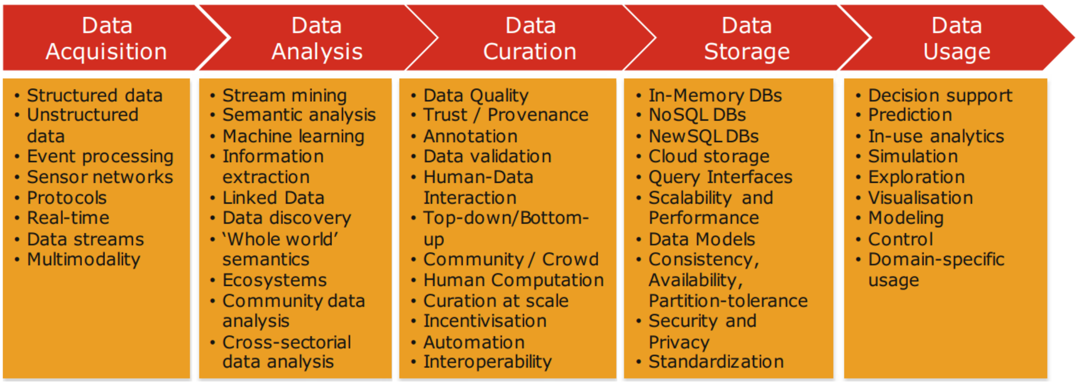
The Big Data value chain identifies key level activities (CURRY, 2016), shown in Figure 2.2. Below we detail these activities:

- *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,

[8] https://www.gehealthcare.com/products/anesthesia-delivery/carestation-insights

[9] https://www.philips.com/a-w/about/news/archive/blogs/innovation-matters/20181112-how-a-virtual-heart-could-save-your-real-one.html

Figure 2.2 – Big Data Value Chain.



Source: (CURRY, 2016)

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 mapped mostly 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 (COUTO et al., 2019). According to Sawadogo and Darmont (2021), 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 kinds of users; 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 (KRONBERGER et al., 2020; CORREIA et al., 2022).

In this work, 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, we consider the DM component has to address the following issues: heterogeneity, interoperability, integration, data discovery/search, and data quality.

2.5.1 Data heterogeneity

Increased computational power and the development of IoT devices have caused a large-scale data processing change, thus also allowing an explosion in the variety of data types and sources. Therefore, access to these various types and data sources has become complex, and the DM component has to be able to deal with this *data heterogeneity*.

According to Jirkovsky, Obitko and Marik (2017), there are three levels of heterogeneity:

- *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.

- *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, their integrity. Therefore, obtaining a unified view of all these different types of data sets is an arduous and complex task. Data heterogeneity is an essential challenge in any application context that needs to deal with different sources that generate lots of data, such as the application domains mentioned in Section 2.4. One of the highest expectation of complex systems such as DTs is to achieve transparent integration, where data can be accessed, recovered, and treated through techniques, tools, and algorithms in a uniform way.

2.5.2 Data integration

Data integration aims to combine data from multiple sources, providing a high-level unified view that makes data amenable to use in Analysis activities (RAHM, 2016a). The role of data integration is fundamental to an organization's efficiency (KADADI et al., 2014). The data themselves do not provide value and need to be processed. The use of specific techniques and approaches is required to extract relevant information.

There are several approaches to integrate the data. Physical data integration is the most popular approach, where the source data is combined within a new dataset or database (data lake, data warehouse) tailored for analysis activities. In virtual data integration, the data entities remain in their original data sources and are accessed at run time.

Other critical approaches for data integration are data transformation, semantic enrichment, entity resolution (data matching), entity merging, and combining and merging metadata models such as schemata and ontologies (DOAN; HALEVY; IVES, 2012; DONG; SRIVASTAVA, 2015). Semantic data enrichment can be achieved by linking entities or metadata such as attribute names to knowledge resources (e.g., dictionaries, ontologies, knowledge graphs).

Entity Resolution is a data integration approach that aims to identify different descriptions or entity profiles that correspond to the same object in the real world and can be applied to any type of data (structured, unstructured, semi-structured) (PAPADAKIS et al., 2020). This approach is also known as data fusion, data merging, data consolidation, or finding representations (BLEIHOLDER; NAUMANN, 2009).

Data transformation is a very useful approach when the goal is mapping schemata or integrating schemata. For this, the data is transformed according to the global schema of the integrated information system. Schema mapping assumes a particular destination schema and, by identifying correspondence with origin schemata, generates a set of elements to determine how data should be transformed. Schema integration, however, aims to generate a global scheme from individual schemata creating a new and correct new schema (BLEIHOLDER; NAUMANN, 2009).

Ontologies and knowledge graphs are approaches that integrate different data and provide a semantic understanding. According to Studer, Benjamins and Fensel (1998), "an ontology is the formal and explicit specification of a shared conceptualization", which is the hierarchical definition of generalization and specialization of concepts. Knowledge graphs provide a rich knowledge base for data integration, in which they organize entities hierarchically, categorically or by classes and interconnect them through various semantic relationships (RAHM, 2016b).

Integration in the DT scenario should consider data from different application domains that 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 a context and even the different parts of the DT, enabling the production of actionable and valuable knowledge for the business.

2.5.3 Interoperability

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" (GERACI, 1991). A systematic review in cyber-physical systems (GüRDüR; ASPLUND, 2018) identified ten types of interoperability. Due to the focus on the DM component of a DT, we focused on semantic and data interoperability.

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 (MYKKÄNEN; TUOMAINEN, 2008). In (RENNER, 2001), data interoperability is defined as the ability to correctly interpret data that cross the system or organizational limits.

Semantic interoperability is when the meaning of the information model in the context of an area or domain is understood, with a common understanding of concepts, taxonomies and meanings (PLATENIUS-MOHR et al., 2020). Interoperating between different systems and applications is a problem orthogonal to all domains and becomes even more relevant within the context of DTs (LEAL; GUÉDRIA; PANETTO, 2019). Therefore, developing solutions to promote interoperability and shared semantics requires developing and deploying open standards and ontologies.

Since ontologies have the ability to explicitly represent the semantics of the domain, define entities, relationships and their properties, and establish a common understanding between data from different data sources, they can be a useful approach to providing semantic interoperability (OBRST, 2003). DTs deal with data from various sources and systems, and ontologies can assist to understand the different types of taxonomies, identify and solve distortions of meaning, and also provide a pattern of information sharing.

The basis of the operation of DT is the exchange of data and information between its different layers and between DTs from different domains of application. Therefore, data interoperability and semantic interoperability are fundamental in the context of DTs because they allow data sharing in an effective way, unlocking barriers of communication and understanding, and making dependent activities and processes more fluid. Data analysis activities can also benefit from data interoperability and semantic interoperability since they contribute to more data becoming available. In this way, the analyzes get richer, generating the most interesting insights.

2.5.4 Data search and discovery

Generating value from data requires the ability to find, access, and make 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 (CHAPMAN et al., 2020). Relevant sub-disciplines include databases, document and keyword search, entity-centric search, semantic search, tabular search, among others. The required underlying infrastructure for handling the search includes query parsers and evaluators, optimizers, indexes for various data types, metadata and ontologies, reasoners, etc.

A structured query requires the user to know and understand the database schema. When using a structured query language like SQL or XQuery, the user can create questions for the database and thus obtain a data set as an answer (YU; JAGADISH, 2007). The structured query is a powerful way to access data in a database. However, formulating complex queries is difficult for most users and, in addition, has no support for semantic data recovery (MUNIR; ANJUM, 2018).

Some approaches use domain ontologies adapted for database modeling to support the search by the semantics of the data. By taking advantage of the ability of ontologies to define a semantic data model, we have the advantage of improving search capacity, including the enrichment of queries in traditional databases, making it possible to query and recover the semantics of the data (MUNIR; ANJUM, 2018).

In the Big Data context, most data are not structured, so the search process, although fundamentally similar to that of structured searches, differs because it has its scope distributed (GROVER; KAR, 2017). The Apache Lucene[10] is a powerful, performative and scalable full-text search library that provides advanced indexing and search resources such as phrase queries, wildcard queries, proximity queries, range queries, and others. Apache Lucene is the basis for famous and widely used search engines like Solr and Elasticsearch (VENKATESH et al., 2019).

Elasticsearch[11] is a distributed search and data analysis mechanism built on Lucene. It is able to deal with all types of data, including textual, numerical, geospatial, structured, and unstructured. It is the central component of the Elastic Stack, which is a set of free tools for the ingestion, enrichment, storage, analysis, and visualization of data.

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 data search objective is different when considered in the realm of data usage activities and data usage activities. From a DM component perspective, the goal is to support data consumption, such that relevant data can be submitted to analysis models (e.g., machine learning models, what-if scenarios), event managers, or consumed through dashboards, visualizations or reports.

2.5.5 Data quality

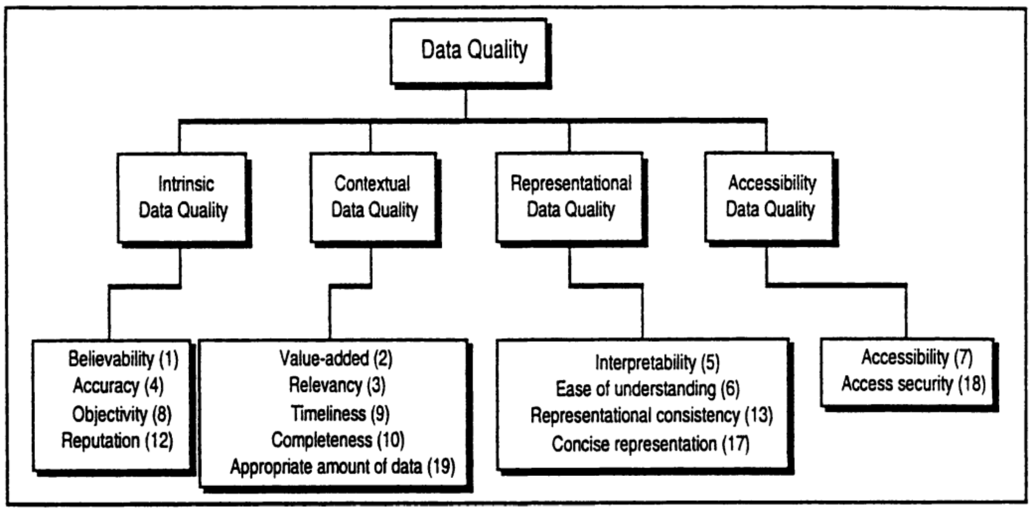
To guarantee quality information from a data curation standpoint, it is necessary to develop methods, metrics, and tools to manage *data quality*. The literature has defined specific characteristics or dimensions to manage data quality, such as timeliness, completeness, consistency, accuracy, etc (SIDI et al., 2012). 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 compatible the data is with previous data, and accuracy measures how accurate the data is relative to actual values.

Wang and Strong (1996) developed a conceptual data quality framework that has been divided into four categories intrinsic, contextual, representational, and accessibility

[10] https://lucene.apache.org/

[11] https://www.elastic.co/pt/what-is/elasticsearch

Figure 2.3 – A Conceptual Framework of Data Quality



Source: (WANG; STRONG, 1996)

as shown in Figure 2.3. The intrinsic quality category includes the dimensions of *credibility* that indicates whether the accepted data is true, *accuracy* indicates whether the data is correct and free of errors, *objectivity* that indicates whether the data It is impartial and *reputation* indicates whether the data is reliable or considered reliable concerning its origin.

The contextual data quality category includes the dimension of *value-added* that indicates whether the data present advantages for the user, *relevancy* that indicates whether the data is useful for a given context, *timeliness* indicates whether the data is appropriate for a given context, *completeness* indicates whether the data is complete or not null and *appropriate amount of data* shows whether the amount of data available is adequate.

The representational data quality category includes *interpretability* that indicates whether the data is in a clear and clearly defined language, *ease of understanding* indicates whether the data is clear and easy to understand, *representational consistency* indicates whether the data is always represented in the same format and *concise representation* indicates whether the data is compactly represented in a context. And finally, the category of *accessibility* indicates whether the data is available and is easily accessible and if a *access security* to the data is appropriate.

Data cleaning and preprocessing are examples of operations that improve the data quality. Besides, they are integral parts of integration activities, such as dealing with noisy and missing data, duplicated data, outlier detection, normalization, transformation, etc. Big data has increased the complexity of managing data quality as the heterogeneity and volume of data have increased. In addition, there is a high degree of complexity in identifying events that reveal incorrect data or abnormal conditions in data streams, requiring advanced solutions (e.g., machine learning).

Machine learning techniques can improve the accuracy and efficiency of data cleaning algorithms from the statistical perspective to deduplicate data, repair incorrect and missing values, and remove erroneous or irrelevant data. However, there are some challenges inherent to data cleaning research for big data. The scalability of data cleaning techniques is a challenge, given the rapid growth of data sets. Hence the development of automated and scalable approaches is an important matter. Cleaning techniques focused on semi-structured and unstructured data are an open field of research, as well as developing data cleaning approaches from the qualitative perspective for streaming data (CHU et al., 2016).

High-quality data is essential in any context, whether in business, decision support systems, machine learning algorithms, or DTs. Being able to rely on data and information is critical to the success of data-oriented activities and processes. In the context of DTs, data quality management gains a prominent space when considering the data flow loop and information that characterizes a DT. Therefore, it is necessary to prioritize data quality in developing and managing a successful DT.

3 RELATED WORK

This chapter describes existing reviews in the context of DTs that motivated our SLR. We also compare them with our work, justifying the need for our SLR.

3.1 Literature Reviews about Digital Twins

The first step in the planning phase of an SLR is to identify the need by comparing it with existing literature reviews. In this section, we summarize secondary studies (SLR, surveys, reviews) in the context of DTs that motivated the development of ours. We describe these secondary studies in chronological order. The goal and the research questions addressed by each work are summarized in Table 3.1. Table 3.2 summarizes the source of information, the search string, and the number of selected studies.

The survey described in (RAPTIS; PASSARELLA; CONTI, 2019) investigates the state of the art of data management at the level of cyber-physical industrial networks. The motivation behind this survey is to provide researchers coming from both the communications/networking/computation fields and the industrial/manufacturing/automation fields an overview of data management issues at the intersection of these two large domains. The main contributions are related to evaluating selected studies to identify data properties (volume, variety, traffic, and criticality) and investigate in different use cases corresponding data technologies used to handle these properties. The authors also outline the architectural design of I4.0 with respect to their data management philosophy, more specifically, data presence, data coordination, and data computing at the network level. In addition, they provide a taxonomy that shows the latest technologies at the level of data-centric networks and services. The authors conclude their survey by discussing future challenges and open research on data management in cyber-physical industrial environments. This work discussed data management from a network and manufacturing interdisciplinary level, and hence in a completely different perspective and granularity compared to our SLR.

The SLR in (BARRICELLI; CASIRAGHI; FOGLI, 2019) seeks to extract from the primary studies the characteristics of DTs and the design implications concerning the DT life cycle. They compile 29 different definitions for the term Digital Twin and argue that these definitions highlight specific key points (e.g., clone/counterpart, integrated system). Another contribution is investigating the most interesting applications of DTs that have been published in the scientific literature. They explore beyond the I4.0 domain, highlighting applications in aviation and healthcare. In addition, the authors identified DT's design implications at different stages of their life cycle, detecting two possible types of life cycles for DTs. The first cycle begins in the design phase of its physical twin, which does not yet exist, while the second cycle begins when the physical twin already exists and has been operating for some time. Although it provides a better understanding of the term Digital Twin and its properties, this work does not explicitly address data management.

The main contribution of the SLR in (JONES et al., 2020) is to extract from primary studies in an exhaustive way information that characterizes a DT and all its elements. For the authors, identifying all the characteristics of a DT is fundamental, as it helps the academy to consolidate knowledge about how to best represent it. This review also identifies the benefits of DTs, the gaps, and future directions reported in primary studies. This work also does not explicitly address data management.

The literature review described in (FULLER et al., 2020) also aims to compile all the characteristics of a DT from existing primary studies. It extends knowledge about the concept of DT by addressing misconceptions and explicitly defining what a DT is and what it is not. The authors sought in the primary studies all DTs definitions and, based on cyber-physical integration level, they distinguished between digital model, digital shadow, and DT, as already mentioned in Section 2.3. This helps to define a scope and allows to establish a manner to measure different degrees or levels of cyber-physical integration maturity to reach a DT. The authors also discuss open challenges from different perspectives: Data Analytics, IoT/IIoT, and enabling technologies. Among the data data analytics challenges, the authors mention IT Infrastructure, privacy and security, trust, and expectations in terms of added value.

The SLR reported in (SEMERARO et al., 2021) is a review with a broader scope compared to the aforementioned ones. Using a 5W+1H approach[1], the authors sought to collect information from primary studies to define what a DT is, what the technologies and components are used to implement a DT, what are the possible functions for which DTs are used, who is developing DTs, and for which stage of the product life cycle. The authors seek to understand how to design a DT based on the architectures proposed by selected studies. They highlighted different layers for a DT architecture (physical layer, network layer, and computing layer) and components (sensors, systems, communication

[1] What, Where, Who, Why, When, How

protocols, middleware, APIs, data-driven methods, physical, geometric model, behavior, fidelity, modularity, etc.). In summary, the main contribution of this SLR is the provision of an overview of the concept of DT from different aspects.

3.2 Final considerations and comparison

Tables 3.1 and 3.2 summarize the existing surveys and compare them with our SLR. From Table 3.1, it is possible to identify how different the goal and the research questions of each work are and how focused they are in understanding the terminology and DT properties, the existing applications, the domains for which they are built, and the underlying technology. Two literature reviews (RAPTIS; PASSARELLA; CONTI, 2019; JONES et al., 2020) did not define specific research questions, as they chose to perform a thematic analysis in the corpus. Our SLR aims to better understand the role and the functionality provided by the DM component in a DT, with research questions that have not been addressed in the literature.

Table 3.1 – Research questions of related literature reviews

Study	**Research questions**	**Goal**
(RAPTIS; PASSARELLA; CONTI, 2019)	Thematic analysis	Investigates management at network level, communication and industrial automation
(BARRICELLI; CASIRAGHI; FOGLI, 2019)	(i) What are the definitions of Digital Twin that have been published in literature?; (ii) What are the main characteristics that should be present in a Digital Twin?; (iii) What are the domains in which Digital Twin applications have been developed and described in scientific literature?	Investigates DT definitions, its characteristics, applications and design implications
(JONES et al., 2020)	Thematic analysis	Aims to characterize the concept of DT and its parts
(FULLER et al., 2020)	(i) What is a Digital Twin and what are some of its misconceptions with current and previous definitions?; (ii) What are the applications, challenges, and enabling technologies associated with IoT/IIoT, data analytics and Digital Twins?; (iii) Is there a link between IoT, IIoT and data analytics with Digital Twin technology?; (iv) What are the open research and challenges with Digital Twins?	Investigates the technologies associated with DT and its applications
(SEMERARO et al., 2021)	(i) What is a Digital Twin?; (ii) Where is appropriate to use a Digital Twin?; (iii) Who is doing Digital Twins?; (iv) When has a Digital Twin to be developed?; (v) Why should a Digital Twin be used?; (vi) How to design and implement a Digital Twin?; (vii) What are the main challenges of implementing a Digital Twin?	Investigates the concept of DT at the level of functions, components and technologies, context, life cycle and architecture
Our work	**(i) For which domains the DT solutions were proposed?;** **(ii) Under the perspective of data usage, for which functions were the DTs proposed?** **(iii) What types of data do the proposed solutions consider?;** **(iv) What solutions were proposed for the DM issues addressed?;** **(v) What kind of technical infrastructure is considered?**	**Investigates data management in the context of DTs**

Source: The author

Table 3.2 complements the previous one with more details about the literature reviews: the digital libraries used to search for primary studies, the search strings, and the number of selected studies. It is possible to see that all studies vary enormously in the search and selection strategies, resulting in a completely different set of primary studies. They also do not explicitly address data management issues in their search string.

Table 3.2 – Literature reviews comparison

Study	Source	Search string	Corpus
(RAPTIS; PASSARELLA; CONTI, 2019)	Selected journals from manufacturing and network domains	Unspecified	Unspecified
(BARRICELLI; CASIRAGHI; FOGLI, 2019)	Google Scholar	(”digital twin artificial intelligence” and ”digital twin model”)	75
(JONES et al., 2020)	Google Scholar	(”digital twin”)	92
(FULLER et al., 2020)	Google Scholar, ACM, IEEE, Science Direct	(”Digital-Twin”, ”Digital Twins”, ”Industrial Digital Twin”, ”Healthcare Digital Twin”, ”Smart Cities Digital Twin”)	43
(SEMERARO et al., 2021)	Scopus, Elsevier, Science Direct	(”digital twin”, ”factory of future”, ”industry 4.0 technologies”, ”cyber-physical system”, ”predictive manufacturing”)	115
Our work	**ACM, IEEE, OnePetro, Scopus, Science Direct, Web of Science**	**(“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”))**	**57**

Source: The author

Our SLR is the only one that investigates the data management aspects in DTs, expressing it in the search string, goal, and research questions. This finding justifies the need for the development of an SLR such as ours, which explicitly examines aspects of data management such as data integration, data interoperability, data search, and data quality in the layer closer to applications, i.e., closer to the applications that consume and use the data produced by different parts of the DT. Therefore, our SLR complements and innovates this landscape of literature reviews in the context of DTs when investigating data management aspects not yet analyzed.

4 SYSTEMATIC LITERATURE REVIEW: PROTOCOL DEFINITION

In this chapter, we discuss the protocol defined for this SLR. The methodology adopted followed the guidelines of systematic literature reviews proposed by (KITCHENHAM; CHARTERS, 2007), described in Section 2.2.

4.1 Objective and Research Questions

An SLR must summarize 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 Section 2.5. As summarized in Chapter 3, related work has not addressed this subject.

According to this goal, we defined five research questions (RQ), presented in Table 4.1 with the respective motivation.

Table 4.1 – 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 Section 2.5, 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 implementing the proposed DM solutions. It aims to identify technological trends that support data management in DTs.

Source: The author

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 (JONES et al., 2020; TAO et al., 2019; BARRICELLI; CASIRAGHI; FOGLI, 2019; FULLER et al., 2020), which identify 2014-2015 as the initial year of relevant publications about DTs. Note that the pioneering work that proposed a central DM component (TAO; ZHANG, 2017) dates from 2017, and therefore our sample seems consistent.

Then 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 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 4.2 summarizes the DLs used. These DLs index the main journals and conferences on computer science, enable to search papers using expressions combining keywords and logical expressions, and allow for the search to be performed in the title, abstract, and keywords.

We developed this review using two online collaborative systems: *Parsifal*[1], a

[1] <https://parsif.al>

Table 4.2 – 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>

Source: The author

support system for conducting SLRs, *Mendeley*[2], a reference manager system. 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 the Mendeley system. The results reported in this work refer to the last search performed on November 20th, 2021.

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.

We defined the following *inclusion criteria*: (1) papers written in English; (2) papers published in 2014 or later; (3) studies addressing DTs or DSs, according to the definition in (FULLER et al., 2020); (4) studies that explicitly propose DM solutions for DTs/DSs, (5) primary studies; (6) 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 criteria, we disregarded all studies describing data pre-processing designed to prepare/improve an 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.

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

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.

4.4 Quality Assessment

Our protocol includes quality assessment motivated by the following goals: identifying the individual contribution of each study to our research and assessing whether quality differences help explain the results.

We selected nine criteria to evaluate each study from the perspective of methodological quality and contribution level for data management in the context of DTs. The quality questions are summarized in Table 4.3.

Table 4.3 – Quality assessment criteria

Quality Questions
QQ1. Is there a clear statement of the goals of the research (DERMEVAL et al., 2016)?
QQ2. Is the problem to be solved by the technique/method/approach/framework clearly explained (TIWARI; GUPTA, 2015)?
QQ3. Is there sufficient discussion of related work (TIWARI; GUPTA, 2015)? (Are competing techniques discussed and compared with the present technique?)
QQ4. Is the proposed technique/method/approach/framework clearly described (DERMEVAL et al., 2016)?
QQ5. Is there an adequate description of the context (industry, laboratory setting, products used and so on) in which the research was carried out (DERMEVAL et al., 2016)?
QQ6. Is there a discussion about the results of the study (DERMEVAL et al., 2016)?
QQ7. Are the limitations of this study explicitly discussed (DERMEVAL et al., 2016)?
QQ8. Is the broader relevance of the work discussed (TIWARI; GUPTA, 2015)?
QQ9. Is the study increase the knowledge about data management in DTs research (TIWARI; GUPTA, 2015)?

Source: The author

Each quality question (QQ) was judged against three possible answers: “Yes” (score = 1), “Partially” (score = 0.5), or “No” (score = 0). Consequently, the quality score for each particular study is computed by taking the sum of the scores of the answers to the questions and can reach up to nine points. We defined a minimum quality threshold (i.e., minimum of five); otherwise, the study should be discarded.

4.5 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 under-

Table 4.4 – 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	Technical infrastructure used/recommended for implementing the solution, with a focus on cloud processing and DBMS	RQ5

Source: The author

standing of the domain. We extracted the data for each selected paper by filling a form with the predefined fields described in Table 4.4. We used spreadsheets to tabulate data extracted from selected studies, summarize the results and generate tables and graphs.

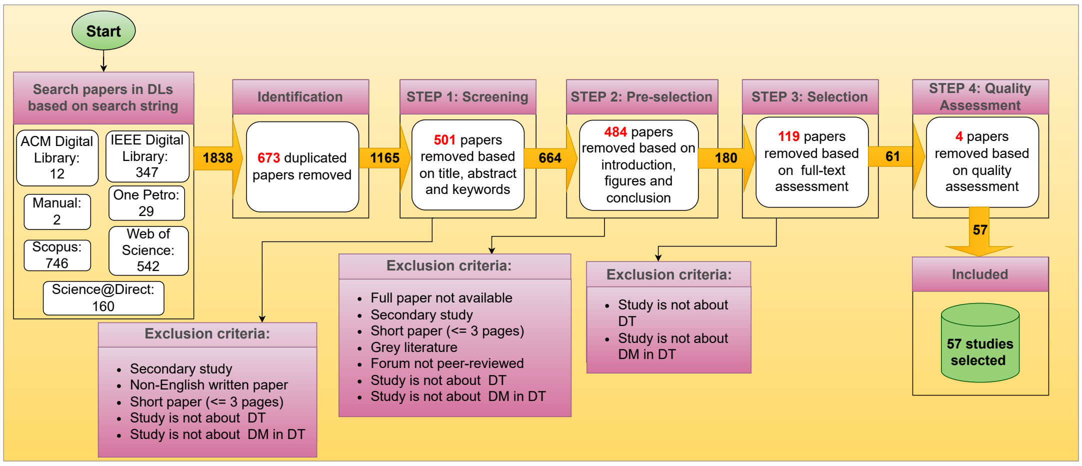
5 RESULTS

This chapter describes our findings by answering the five research questions we addressed in our systematic literature review.

5.1 Study Selection

Figure 5.1 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, which was reduced to 1,165 after eliminating duplicates. The screening of articles resulted in 664 candidate studies, of which we pre-selected 180 candidate studies based on the introduction, figures, tables, and conclusions. From the full reading, we selected 61 studies, which were then submitted for quality assessment.

Figure 5.1 – Study selection and quality assessment



Source: The author

5.2 Quality Assessment

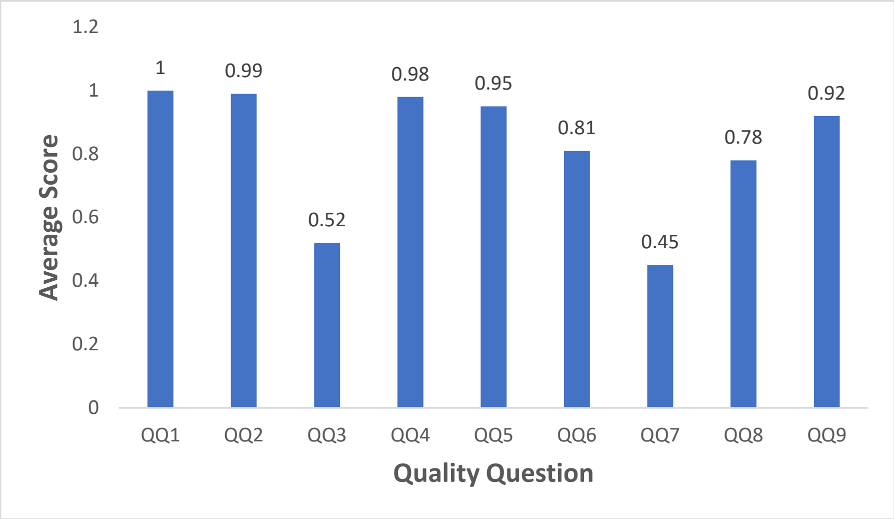
We performed the quality assessment on the 61 selected studies, using the questions in Table 4.3. The quality scores for each study are presented in Appendix A (Table A.1).

As depicted in Figure 5.1, we additionally removed four studies (SHA; ZEADALLY, 2015; DAI et al., 2017; JIANG; CHEN; LIU, 2021; HÄNEL et al., 2021) that obtained a score inferior to five, which in our judgment, denotes irrelevance to our

SLR. Thus, our final selection includes 57 studies that are aligned with the primary objective of our SLR.

The average score of the remaining 57 papers is 7.39. Figure 5.2 details the average score for each QQ. Quality assessment has allowed us to identify that two QQs, 3 and 7, presented the lowest score compared to the others. The average score of QQ3 is only 0.52, which indicates the immaturity of data management solutions in the context of DTs, given that through this question, we evaluate if the studies present a discussion of sufficient related works and if there is a comparison with competing techniques. The average score of QQ7 is only 0.45. Through this question, we assess if the studies discuss their limitations explicitly. This result denotes the lack of mastery of the problem and the impact of the results of the solutions proposed by the studies.

Figure 5.2 – Quality questions average score



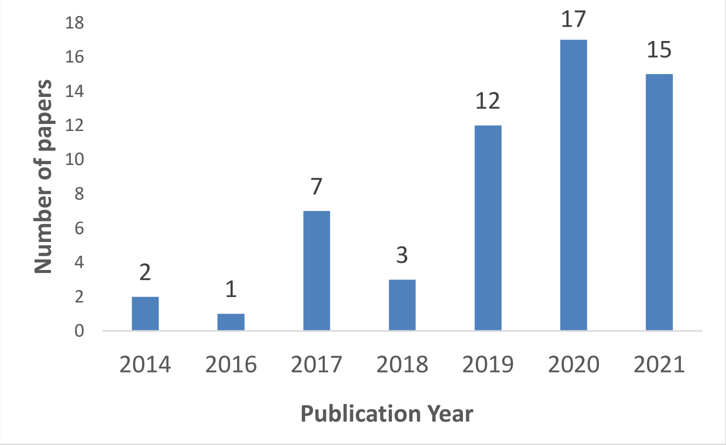
Source: The author

5.3 Data Extraction Results

We extracted data from the 57 selected studies according to the data extraction form defined in the protocol (Table 4.4). The results are summarized in Table 5.2. Before we present the results for each research question, we present some statistics from the selected studies.

Figure 5.3 shows the distribution of publications by year. We can observe that the number of publications explicitly addressing DM issues significantly increase from 2019 on. We hypothesize that the evolution in the state-of-the-art and state-of-the-practice

Figure 5.3 – Number of papers per year



Source: The author

Table 5.1 – Studies grouped based on scientific vehicle and institution type

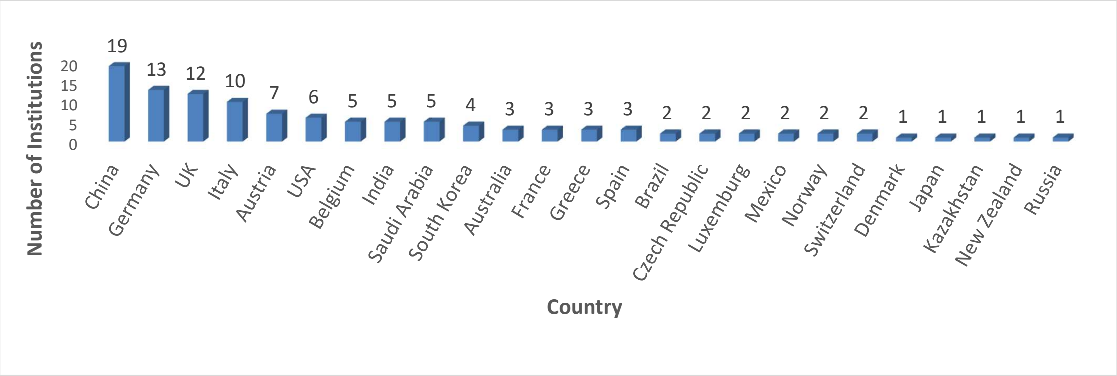
Scientific vehicle	Academia	Industry	Both
Conference	18	5	10
Journal	20	0	4

Source: The author

in DTs has motivated us to address DM issues explicitly to take this concept to its full realization (SINGH et al., 2021) and construct the necessary conceptual, methodological, and technological data management foundations for DT development.

Figure 5.4 displays the countries developing research in this area, which were summarized based on the involved organizations. We identified contributions from twenty-five countries in total, where the most active ones are China, Germany, the United Kingdom (UK), and Italy. These countries have the largest companies in the industry, manufacturing, supply chain, aviation, and energy, which may explain these results.

Figure 5.4 – Number of institutions per country



Source: The author

We also categorized studies by type of scientific vehicle and type of institution, as shown in Table 5.1. The results show that most studies (66.66%) were carried out by academia, five studies (8.77%) were developed by industry alone, and fourteen (24.56%) through partnerships between academia and industry. Given the prevalence of academic works, we conclude that this is a topic still in its infancy regarding actual implementations. The interest in practical applications of DTs is revealed by the number of works that involve the industry alone or in partnerships.

Table 5.2 – Selected studies

ID	Study	Type	Domain	Subdomain	Function	Data type	Interop.	Integ.	Search	Qual.	Infrastructure
1	(ZHANG et al., 2017)	DS	HC	Healthcare	Decision support	Real-time, Historical	✓	✓			Cloud
2	(ALHUMUD; HOSSAIN; MASUD, 2016)	DS	HC	Healthcare	Decision support	Real-time, Historical	✓				Cloud
3	(NÚÑEZ-VALDEZ et al., 2020)	DS	HC	Healthcare	Decision support	Historical		✓			Unspecified
4	(HUSSAIN; PARK, 2021)	DS	HC	Healthcare	Decision support, Asset monitoring	Real-time, Near real-time		✓	✓		Cloud, SQL, NoSQL
5	(HINOJOSA-PALAFOX et al., 2019)	DS	I4.0	Smart manufacturing	Anomaly detection	Real-time, Historical		✓			Cloud, SQL, NoSQL
6	(SINGH et al., 2021)	DS	I4.0	Aviation industry	Asset monitoring	Historical		✓	✓		SQL
7	(PLATENIUS-MOHR et al., 2019)	DS	I4.0	Industrial	Asset monitoring	Historical	✓				Cloud
8	(WANG; CHENG, 2021)	DS	I4.0	Maintenance of industrial robots	Asset monitoring	Historical		✓			Unspecified
9	(AGARWAL; MCNEILL, 2019)	DS	I4.0	Oil and Gas	Asset monitoring	Real-time				✓	Unspecified
10	(KONG et al., 2021)	DS	I4.0	Smart manufacturing	Asset monitoring	Historical CSV		✓		✓	NoSQL
11	(LANDOLFI et al., 2018)	DS	I4.0	Smart manufacturing	Asset monitoring	Near real-time, Historical		✓			Unspecified
12	(OAKES et al., 2021)	DT	I4.0	Smart manufacturing	Asset monitoring	Real-time, Historical			✓		Unspecified
13	(ZONZINI et al., 2020)	DS	I4.0	Smart structures	Asset monitoring	Real-time. JSON		✓			Cloud, NoSQL
14	(BRECHER et al., 2021)	DS	I4.0	Automotive glazing industry	Decision support	Real-time		✓			Cloud, Edge, SQL
15	(JIRKOVSKY; OBITKO; MARIK, 2017)	DS	I4.0	Eletric energy	Decision support	Historical		✓	✓		NoSQL
16	(SAHLAB et al., 2021)	DS	I4.0	Industrial	Decision support	Real-time, Historical		✓	✓		Unspecified
17	(ZHANG; JI, 2019)	DS	I4.0	Smart manufacturing	Decision support	Real-time, Historical. XML, JSON				✓	Cloud, SQL
18	(YU et al., 2019)	DS	I4.0	Smart manufacturing	Decision support	Real-time		✓		✓	Cloud, Edge, SQL, NoSQL

Table 5.2 – Selected studies

ID	Study	Type	Domain	Subdomain	Function	Data type	Interop.	Integ.	Search	Qual.	Infrastructure
19	(HOOS; HIRMER; MITSCHANG, 2017)	DS	I4.0	Smart manufacturing	Decision support	Historical. XML, tabular		✓			SQL, NoSQL
20	(ANDIA; ISRAEL, 2018)	DS	I4.0	Oil and Gas	Event monitoring	XML, JSON, XLSX, PDF		✓			Cloud, SQL, NoSQL
21	(CARDOSO et al., 2021)	DS	I4.0	Smart grid	Fault detection, Predictive maintenance	Real-time		✓			SQL, NoSQL
22	(LIU et al., 2020)	DT	I4.0	Smart manufacturing	Fault diagnosis, Predictive maintenance, Decision support	Real-time. XML		✓			Unspecified
23	(KOUSI et al., 2019)	DS	I4.0	Automotive manufacturing	Optimization of assembly	Real-time, Historical		✓			Unspecified
24	(ZHUANG; GONG; LIU, 2021)	DS	I4.0	Aviation industry	Optimization of assembly	Real-time. CIM/XML		✓			NoSQL
25	(LV et al., 2021)	DS	I4.0	Smart manufacturing	Optimization of assembly	Real-time. XML	✓				SQL
26	(SUN et al., 2020)	DS	I4.0	Steel rebars manufacturing	Optimization of logistics	Historical. JSON, CSV	✓	✓			Cloud, Edge, SQL, NoSQL
27	(PERNICI et al., 2020)	DS	I4.0	Supply chain	Optimization of process	Real-time, Historical		✓			SQL, NoSQL
28	(BRACKEL et al., 2018)	DS	I4.0	Oil and Gas	Optimization of production	Real-time	✓				Cloud, Edge
29	(BLUM; SCHUH, 2017)	DS	I4.0	Smart manufacturing	Optimization of production	Real-time. Tabular				✓	Unspecified
30	(GÓMEZ-BERBÍS; AMESCUA-SECO, 2019)	DS	I4.0	Smart manufacturing	Optimization of production	Historical. XML		✓			Unspecified
31	(LIU et al., 2021)	DS	I4.0	Smart manufacturing	Optimization of production	Real-time, Historical. XML		✓			SQL, NoSQL
32	(LIU et al., 2022)	DS	I4.0	Metal Additive Industry	Optimization of production, Decision support	Real-time, Historical. XML		✓			Cloud, Edge
33	(CHEN et al., 2020)	DS	I4.0	Smart manufacturing	Optimization of production, Decision support	Real-time, Historical		✓			Unspecified
34	(SUHAIL et al., 2021)	DS	I4.0	Aviation industry	Predictive maintenance	Real-time, Historical	✓				Blockchain storage
35	(ANSARI; GLAWAR; NEMETH, 2019)	DS	I4.0	Smart manufacturing	Prescritive maintenance, Decision support	Real-time, Historical		✓			SQL
36	(KIRCHEN et al., 2017)	DS	I4.0	Chemical industry	Quality assessment, decision support	Historical				✓	Unspecified
37	(AL-ISMAEL; AL-TURKI; AL-DARRAB, 2020)	DS	I4.0	Oil and Gas	Simulation improvement	Historical	✓			✓	Unspecified
38	(ZHANG et al., 2021)	DT	I4.0	Smart manufacturing	Simulation improvement	Real-time, Historical	✓	✓			Cloud, SQL
39	(LU et al., 2020b)	DS	SC	Smart buildings	Anomaly detection	Historical. XML	✓	✓		✓	NoSQL

Table 5.2 – Selected studies

ID	Study	Type	Domain	Subdomain	Function	Data type	Interop.	Integ.	Search	Qual.	Infrastructure
40	(LU et al., 2020a)	DS	SC	Smart city	Anomaly detection, Asset monitoring	Real-time, Historical		✓			Cloud, NoSQL
41	(ALWAN et al., 2020)	DS	SC	Smart city	Anomaly detection, Asset monitoring	Real-time, Historical		✓		✓	SQL
42	(JOUAN; HALLOT, 2020)	DS	SC	Smart buildings	Asset monitoring	Real-time, Historical. CSV, XLSX		✓			SQL, NoSQL
43	(CHEVALLIER; FINANCE; BOULAKIA, 2020)	DS	SC	Smart buildings	Asset monitoring	Historical. JSON		✓			NoSQL
44	(ACQUAVIVA et al., 2019)	DS	SC	Smart buildings	Asset monitoring	Near real-time, Historical. JSON		✓			Cloud, NoSQL
45	(VIVI et al., 2019)	DS	SC	Smart buildings	Asset monitoring	Real-time		✓			Cloud
46	(WU; WANG; SEIDU, 2020)	DS	SC	Urban water supply	Asset monitoring, Decision support	Real-time, Historical				✓	Cloud, Edge
47	(RYBNYTSKA et al., 2020)	DS	SC	Smart house	Decision support	Real-time. CSV	✓	✓		✓	Unspecified
48	(KIOURTIS et al., 2018)	DS	SC	Traffic management	Decision support	Historical. XML		✓	✓		SQL
49	(BUJARI et al., 2021)	DS	SC	Urban planning	Decision support	Near real-time, Historical		✓	✓		Cloud, NoSQL
50	(FAN et al., 2021)	DS	SC	Disaster management	Event monitoring, Decision support	Real-time		✓			Unspecified
51	(AZZAM et al., 2019)	DS	SC	Smart city	Event monitoring, Decision support	Real-time, Historical		✓	✓	✓	Unspecified
52	(KASRIN et al., 2021)	DS	SC	Urban mobility	Event monitoring, Decision support	Unspecified	✓	✓		✓	Unspecified
53	(HUANG; DAI, 2017)	DS		Unspecified	Decision support	Unspecified			✓		NoSQL
54	(DAO et al., 2014)	DS		General	Event monitoring	Real-time. XML		✓			Cloud
55	(WANG; ZHOU, 2014)	DS		Unspecified	Event monitoring	Real-time		✓			Unspecified
56	(GIFTY; BHARATHI; KRISHNAKUMAR, 2020)	DS		Unspecified	Fault detection	Unspecified				✓	Unspecified
57	(PROPER; BORK; POELS, 2021)	DS		Unspecified	Optimization of process	Historical		✓			Unspecified

Source: The author

5.4 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 5.2 (column *Domain*), most selected studies address the domains of I4.0 (59.64%), followed by smart cities (24.56%) and healthcare (7.01%). Five studies (8.77%) do not detail the domain/subdomain, and one study is a domain-agnostic proof-of-concept (POC) (DAO et al., 2014).

- **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 (44.11%), followed by O&G (11.76%), aviation (8.82%), automotive industry (5.88%), and the electric energy industry (2.94%). Some studies (PLATENIUS-MOHR et al., 2019; SAHLAB et al., 2021) 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 (LANDOLFI et al., 2018) to methods for building DTs for manufacturing factories (OAKES et al., 2021), monitoring and predicting the carbon emission of a factory (ZHANG; JI, 2019), optimizing hot lamination schedule (CHEN et al., 2020), among others. In the specific area of O&G, the DTs address asset management (AGARWAL; MCNEILL, 2019), improvement of drilling operations (ANDIA; ISRAEL, 2018; BRACKEL et al., 2018), and improvement of digital reservoir simulation (AL-ISMAEL; AL-TURKI; AL-DARRAB, 2020).

- **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 (57.14%) either encompass the city as a whole (LU et al., 2020a; ALWAN et al., 2020; AZZAM et al., 2019), or a specific aspect, such as urban planning (BUJARI et al., 2021), disaster management (FAN et al., 2021), traffic management (KIOURTIS et al., 2018), urban water supply (WU; WANG; SEIDU, 2020) and mobility (KASRIN et al., 2021). The second most prevalent level is smart building (35.71%), where the objective is monitoring the building's energy consumption (ACQUAVIVA et al., 2019), controlling the internal temperature (VIVI et al., 2019), detecting faults and anomalies (LU et al., 2020b), as well as monitoring the building's infrastructure (JOUAN; HALLOT, 2020). Two studies (LU et al., 2020a; ALWAN et al., 2020) did not define a specific city scenario for which the solution was proposed.

- **Healthcare (HC)**: we identified two groups of applications in this area, namely health data management and treatment of patients. In the former group, (ZHANG et al., 2017) proposes the integration of data from different stakeholders (e.g., hospitals, pharmaceuticals, patients) to leverage the creation of new services and applications, and (ALHUMUD; HOSSAIN; MASUD, 2016) 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 (NÚÑEZ-VALDEZ et al., 2020) and predicting the risk of stroke (HUSSAIN; PARK, 2021).

5.5 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 (SEMERARO et al., 2021). The functions identified are summarized in Table 5.2 (column *Function*). The function in most studies was classified as decision support (38.59%) since the DT allows the analysis of multiple variables and hence, supports data-driven decision-making. In addition, we identified more specific functions, namely optimization of production/process/logistics (21.05%), asset monitoring (24.56%), event monitoring (10.52%), anomaly detection (7.01%), failure detection/diagnosis (5.26%), quality assessment (1.75%), simulation improvement (3.50%), and predictive/prescriptive maintenance (7.01%).

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 5.2 (column *Type*), it is possible to observe that only three studies (OAKES et al., 2021; LIU et al., 2020; ZHANG et al., 2021) are classified as DTs according to the definition in (FULLER et al., 2020), 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 5.3 presents the data consumers identified in the selected studies, using the Study Ids specified in Table 5.2. The most frequent type of data consumer is visualization tools (35.08%) (e.g., dashboards, graphs, maps, and diagrams). Other studies enable users to formulate queries using customized

Table 5.3 – Data consumption methods

Method	Study ID	%
Visualization (diagrams, graphics, maps, dashboards)	4, 5, 12, 13, 18, 20, 21, 22, 29, 30, 32, 34, 35, 39, 41, 45, 49, 50, 51, 57	35.08%
Other applications and services	2, 7, 14, 16, 19, 24, 26, 27, 34, 40, 45, 50	21.05%
Models (simulation, ML, 3D)	23, 25, 28, 29, 37, 40, 41, 43, 44, 45, 52	19.29%
People (data specialists, decision-makers, stakeholders)	1, 3, 14, 33, 36, 40, 47, 48, 53	15.78%
Database query	4, 6, 12, 15, 16, 47, 49, 51, 53	15.78%
Unspecified	8, 9, 10, 11, 17, 42, 55, 56	14.03%
GUI for data query	2, 31, 38, 39, 46, 54	10.52%

Source: The author

Graphical User Interface - GUI (10.52%) and database query languages (15.78%). Many studies (15.78%) 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 (50%) 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 in 40.34% of the studies. The prevalent DT function categories related to this type of consumers are asset monitoring (30.76%) and optimization in equal proportions (26.92%) and decision support (23.07%). Some studies (14.03%) did not explicitly mention any method for data consumption in the DT.

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

The heterogeneity of data that needs to be handled in DTs in all domains is clear from Section 2.4, which includes sensors/actuators, information systems (e.g., ERP, MES, CRM, SCADA), and data silos (data repositories), social networks, among others. This research question aims at characterizing the heterogeneity in the selected studies in terms of sources of information, formats, sensor technology, and data collection/processing methods according to velocity requirements. Table 5.2 (column *Data Type*) summarizes the type of data and formats.

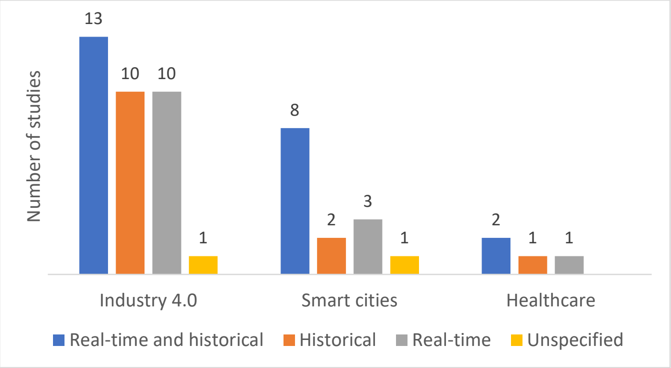
- **Sources:** considering the sources for the data acquisition activities, we identified that most studies (40.35%) consider both (near) real-time data streams generated by sensors/actuators and historical data related to all sorts of information systems and file specifications. About 24.56% of the 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.

- **Formats:** about 35.08% 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.

- **Sensor technology:** while the specific sensor technology is not detailed in most studies, the mentioned ones are RFID, Quick Response Code (QRcode), Global Positioning System (GPS), pipe flow meters, temperature and pressure sensors, and personal/wearable devices (e.g., smartwatches). While in some applications, the sensors are static (e.g., temperature sensors in buildings or production wells), in others, it is necessary to identify the location where the data was produced. The latter was identified in smart cities (e.g., traffic, public transportation) and production/logistics (e.g., product tracking).

- **Data Analysis latency:** applications vary in the requirements of freshness of the data for the analysis, impacting the way in which needs to be collected and processed. We identified three groups: real-time and historical, real-time only, and historical only. Figure 5.5 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 (DAO et al., 2014; WANG; ZHOU, 2014), management of disasters and city events (FAN et al., 2021; AZZAM et al., 2019), faults in the power grid (CARDOSO et al., 2021), and production/assembly lines optimization (BRACKEL et al., 2018; BLUM; SCHUH, 2017; LV et al., 2021). 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 (AL-ISMAEL; AL-TURKI; AL-DARRAB, 2020), decision support (NÚÑEZ-VALDEZ et al., 2020; JIRKOVSKY; OBITKO; MARIK, 2017; HOOS; HIRMER; MITSCHANG, 2017; KIOURTIS et al., 2018) or optimization (PROPER; BORK; POELS, 2021; SUN et al., 2020) are more flexible and less time-dependent, characterizing the use of historical data and also allowing the availability and analysis

Figure 5.5 – Data collection and processing methods



Source: The author

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 identified in DTs with functionalities such as decision support (ZHANG et al., 2017; ALHUMUD; HOSSAIN; MASUD, 2016; HUSSAIN; PARK, 2021; SAHLAB et al., 2021; ZHANG; JI, 2019), anomaly detection (LU et al., 2020a; ALWAN et al., 2020; HINOJOSA-PALAFOX et al., 2019), predictive and prescriptive maintenance (SUHAIL et al., 2021; ANSARI; GLAWAR; NEMETH, 2019), asset monitoring (OAKES et al., 2021; JOUAN; HALLOT, 2020; WU; WANG; SEIDU, 2020).

5.7 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 Section 2.5, namely Interoperability, Data Integration, Data Search and Discovery, and Data Quality. Selected studies may propose solutions that cover more than one issue. The remaining of this section details our finding with regard to each of these issues.

5.7.1 Interoperability

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

- **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. (ZHANG et al., 2021) proposes data mappings to handle data in different formats and interoperate with the other parts of the DT, and (LV et al., 2021; SUN et al., 2020; ALHUMUD; HOSSAIN; MASUD, 2016) propose the conversion of the original data format into a common structure. (AL-ISMAEL; AL-TURKI; AL-DARRAB, 2020) presents an architecture with built-in conversions to promote seamless data exchange between simulation tools. (ZHANG et al., 2017) 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 (SUHAIL et al., 2021) to maintain data traceability after cleaning and conversion/transformation operations to a unified format, enabling the exchange of information between the DT parties.

- **Semantic interoperability:** this group of studies uses domain concepts, patterns, and ontologies to describe data in terms of the respective domain. (ALHUMUD; HOSSAIN; MASUD, 2016) 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. (KASRIN et al., 2021) 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. (BRACKEL et al., 2018) uses the OPC UA protocol, an industry automation standard, to provide semantic interoperability through data tag descriptions for the DT components. (LU et al., 2020b) created an integrated and semantically interoperable data layer, which includes a model for exchanging data with other data sources.

- **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. (PLATENIUS-MOHR et al., 2019) 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, (KIOURTIS et al., 2018) leverages on-

Table 5.4 – Integration approaches

Integration approach	Study ID	%
Centralized repository	5, 18, 21, 35, 41, 49, 54	16.27%
Ad hoc modeling	11, 11, 23, 32, 38, 48, 50, 52	18.60%
Modeling with standards	24, 31, 39, 40, 45	11.62%
Modeling with an ontological layer	6, 10, 15, 16, 22, 26, 30, 33, 42, 43, 51, 57	27.90%
Integration Method	8, 19, 27, 44, 55	11.62%
Semantic enrichment	3, 4, 13, 14, 35, 47	13.95%

Source: The author

tologies to relate data to domain metadata and develop a framework for exchanging data between DTs belonging to different domains.

5.7.2 Integration

We divided the studies addressing integration into six approaches, summarized in Table 5.4 using the Study Ids of Table 5.2, and described below:

- **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, only mentioning the adoption of a data warehouse (ANSARI; GLAWAR; NEMETH, 2019; ALWAN et al., 2020; DAO et al., 2014). (YU et al., 2019; HINOJOSA-PALAFOX et al., 2019; BUJARI et al., 2021) 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, (CARDOSO et al., 2021) 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.
- **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 *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 (KIOURTIS et al., 2018; FAN et al., 2021; KASRIN et al., 2021); a 3D constructor component to provide a unified view of assembly lines (KOUSI et al., 2019); a Human Body Avatar Data Model that integrates the description of all data and provides customized assistive healthcare devices (LANDOLFI et al., 2018); the identification of the critical product life-cycle data that

influences the quality of the final product (LIU et al., 2022); the increase of the transparency and automation level of a blade test workshop (ZHANG et al., 2021); and a data management layer that provides a basic data model for the data from different sources (ZHANG et al., 2017).

- **Modeling with standards:** studies in this category used standards as a support for the creation of an integration data model. In the domain of smart cities, (LU et al., 2020a; LU et al., 2020b; VIVI et al., 2019) leveraged the Build Information Modeling (BIM) standard, and in the I4.0 field, (ZHUANG; GONG; LIU, 2021) adopted the Bill of Materials (BOM) standard and (LIU et al., 2021) explored the ISO STandard for the Exchange of Product model data (STEP) standard for product information exchange. In addition, (LIU et al., 2021; ZHUANG; GONG; LIU, 2021) also proposed a method of data association to support quality traceability. To create a data model based on standards, Lu et al. (2020a) used a centralized non-relational repository (DynamoDB).

- **Modeling with an ontological layer:** the solution proposed by the studies in this category rely on an ontological layer to represent and relate the information with domain knowledge, so as to create a shareable knowledge base. Several works (PROPER; BORK; POELS, 2021; GÓMEZ-BERBÍS; AMESCUA-SECO, 2019; SAHLAB et al., 2021; AZZAM et al., 2019; JIRKOVSKY; OBITKO; MARIK, 2017) proposed an ontology, and the use of a RDF graph to relate the conceptual data model and domain data. Based on three ontologies, (CHEN et al., 2020) proposed a data model based on an ontological layer and a method addressing data fusion and entity resolution. We also found four studies (CHEVALLIER; FINANCE; BOULAKIA, 2020; SUN et al., 2020; LIU et al., 2020; JOUAN; HALLOT, 2020; SINGH et al., 2021) 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, (KONG et al., 2021) proposed a module for cleaning and reducing data.

- **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 (PERNICI et al., 2020); data synchronization (ACQUAVIVA et al., 2019); the automatic mapping of entities (HOOS; HIRMER; MITSCHANG, 2017); the integration of two complex event models using an adapter (WANG; ZHOU, 2014) and association mappings and data fusion

(WANG; CHENG, 2021).

- **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. (ANSARI; GLAWAR; NEMETH, 2019; HUSSAIN; PARK, 2021) created a semantic knowledge base based on domain ontologies to enrich already processed data with more domain meaning, enabling data integration. (BRECHER et al., 2021; NÚÑEZ-VALDEZ et al., 2020; RYBNYTSKA et al., 2020; ZONZINI et al., 2020) used semantic metadata annotation methods to optimize data integration by identifying new classes, relationships, or domain descriptions.

5.7.3 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 pre-defined queries listed in Table 5.3, detailing only those which provided query functionality.

- **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 (HUSSAIN; PARK, 2021; BUJARI et al., 2021). (RYBNYTSKA et al., 2020) 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 Ontology Web Language (OWL) representation of data that were included in the data model. (HUANG; DAI, 2017) addressed the efficiency of information retrieval by transforming and storing the original OWL data representation in a NoSQL database. (SINGH et al., 2021) claims that the industry is more familiar with relational structures and proposes the conversion to a relational database to enable SQL queries.

- **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. (JIRKOVSKY; OBITKO; MARIK, 2017; HUSSAIN; PARK, 2021; AZZAM et al., 2019) 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). (OAKES et al., 2021) recommends using knowledge graphs for semantic queries through the GraphQL query language and APIs that allow queries in JSON format. (SAHLAB et al., 2021) makes a case for knowledge graphs and semantic queries, without referring to any specific query language.

5.7.4 Data Quality

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

- **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 statistical properties of data streams/time series and are used to clean raw sensor streams. (AGARWAL; MCNEILL, 2019) proposes data checking and cleaning algorithms based on statistical properties of the data. To improve the input of simulation tools, (ANDIA; ISRAEL, 2018) 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 (KIRCHEN et al., 2017).

- **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 contextualize data streams raw data. To seamlessly transfer simulation data to other applications, (AL-ISMAEL; AL-TURKI; AL-DARRAB, 2020) proposes a two-layer comparison (specification and threshold/rules). HADES (ALWAN et al., 2020) assumes two levels of cleaning: comparison of real-time data with historical data and assessment by predictive anomaly detection models. In addition to cleaning, (ZHANG; JI, 2019) and (LU et al., 2020b) also propose the use of models to compare deviations from expected data. (GIFTY; BHARATHI; KRISHNAKUMAR, 2020) proposes a quality assessment method with the respective techniques to assess input data, out data generated by models, and feedback data from the CPS.

- **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 (KONG et al., 2021)(KIOURTIS et al., 2018) (AZZAM et al., 2019)(YU et al., 2019)(BLUM; SCHUH, 2017). The pipeline in (YU et al., 2019) and (BLUM; SCHUH, 2017) are part of an Extract, Transform, Load (ETL) process designed to insert cleaned and transformed data in a data warehouse, while the one in (AZZAM et al., 2019) 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 pre-processing role. The reference architecture in (KONG et al., 2021) proposes two levels of cleaning (raw data and customized), while the one in (WU; WANG; SEIDU, 2020) organizes operations and models to clean raw data, assess data properties, and add value into four data management layers. The architecture in (KASRIN et al., 2021) leverages data quality agents to perform well-known cleaning patterns.

5.8 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).

- **Cloud computing:** This category groups the studies that used or recommended the use of cloud data processing and/or storage resources. Some justify this adoption due to processing requirements (ZHANG; JI, 2019; DAO et al., 2014). Most studies adopt both processing and storage (ZHANG et al., 2017; ALHUMUD; HOSSAIN; MASUD, 2016; PLATENIUS-MOHR et al., 2019; SUN et al., 2020; ZHANG et al., 2021; HINOJOSA-PALAFOX et al., 2019; CARDOSO et al., 2021; LU et al., 2020a; VIVI et al., 2019; HUSSAIN; PARK, 2021; ZONZINI et al., 2020; AC-

QUAVIVA et al., 2019; BUJARI et al., 2021). Some studies adopted open-source tools and databases (HINOJOSA-PALAFOX et al., 2019; HUSSAIN; PARK, 2021; ZONZINI et al., 2020), such as Spark, Kafka, and Hadoop for data processing, and Elasticsearch, Hive, InfluxDB, MongoDB, Hadoop, and MariaDB for data storage. Others (LU et al., 2020a; VIVI et al., 2019; CARDOSO et al., 2021) have used or recommended services from commercial cloud providers, such as DynamoDB, S3 and Redshift from AWS.

- **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 (BRECHER et al., 2021; BRACKEL et al., 2018; WU; WANG; SEIDU, 2020). 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 (KASRIN et al., 2021) has mentioned the importance of the cloud, edge, and fog combination, mainly due to real-time data processing requirements. Lastly, we found three studies (YU et al., 2019; SUN et al., 2020; LIU et al., 2022) that have mentioned cloud and edge computing and cloud storage infrastructure.

- **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 (ANSARI; GLAWAR; NEMETH, 2019; ZHANG; JI, 2019), (BRECHER et al., 2021; ALWAN et al., 2020; LV et al., 2021) mentioned the support of a non-specified relational database, and (SINGH et al., 2021; ZHANG et al., 2021; RYBNYTSKA et al., 2020) adopted specific ones (MySQL, SQL Server). Regarding non-relational approaches, most studies adopt open-source databases, such as InfluxDB (ZONZINI et al., 2020), Hadoop and Jena (JIRKOVSKY; OBITKO; MARIK, 2017), MongoDB (ACQUAVIVA et al., 2019), Elasticsearch (BUJARI et al., 2021), and Cassandra (CHEVALLIER; FINANCE; BOULAKIA, 2020). Others adopted AWS storage systems such as DynamoDB (LU et al., 2020a; LU et al., 2020b). Some studies have recommended non-tabular storage systems, without specifying the specific DBMS, such as spatial databases (JOUAN; HALLOT, 2020). Finally, some studies propose the combination of storage systems, such as data warehouse and data lakes (PERNICI et al., 2020), and SQL and New SQL databases (LIU et al., 2021). Others mentioned a combination of specific databases, such as AWS S3 and Redshift (CARDOSO et al., 2021) to implement a data lake, MariaDB and Elasticsearch (HUSSAIN; PARK, 2021), data warehouse and Hadoop (HINOJOSA-PALAFOX et al., 2019), Hive/Hadoop and data lake (YU et al., 2019), SQLite and MongoDB (HOOS; HIRMER; MITSCHANG, 2017), SQL Server and MongoDB (SUN et al., 2020), and MySQL, Oracle, Hadoop and MongoDB (LIU et al., 2020).

6 DISCUSSIONS

In the previous chapter, 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 data quality. This chapter summarizes findings concerning these issues and discusses the trends and opportunities identified in our investigation.

6.1 Data Management Issues

No single work proposes an encompassing solution addressing all the data management issues considered in our survey. Integration is the most discussed one (75.43%), which is an expected result due to the central role of the DM component, acting as a bridge between the other DT components. Data integration with the centralized repository approach was reported in 16.27% of the studies, ad hoc modeling in 18.60%, modeling with standards and integration method in (11.67% each), modeling with an ontological layer in 27.90%, and semantic enrichment as part of data integration was reported in 13.95% of the studies.

The smart city domain is the one for which the most encompassing solutions were identified, where (KASRIN et al., 2021; KIOURTIS et al., 2018) address integration, interoperability, and data quality, and (AZZAM et al., 2019) involves integration, data search, and data quality.

Most of the selected studies address Digital Shadows according to the definition by (FULLER et al., 2020) since they consider the automatic flow of data from the physical to the virtual space only. In addition, we observed that most studies 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 data flow 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, changes and decisions made within the realm of these other components can impact the data managed. Hence, the flow from the consumers into the DM component and from the DM component back into the Physical world also needs to be assessed for the full implementation of the five-component reference architecture (Figure 2.1).

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

6.2 Trends and Opportunities

Regarding the solutions, we observed some common trends and opportunities:

- **Reference architectures for data management:** many works (HINOJOSA-PALAFOX et al., 2019; ALWAN et al., 2020; LU et al., 2020a; VIVI et al., 2019) 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. The reference architecture 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 are proposed, such as the Data

Shared Market (KASRIN et al., 2021), based on clouds and agents, and the Decision Information Packages, which organizes information according to the different stakeholders (LIU et al., 2021). Reference architectures provide patterns, building blocks, interconnections, and a common vocabulary (CLOUTIER et al., 2010). As there is an opportunity to approximate the DM component to the data and knowledge management functionality of Data Lakes (CORREIA et al., 2022; KRONBERGER et al., 2020), 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)[1], 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 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.

- **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.

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

- **Semantic enrichment and ontologies:** an ontology formalizes the intended meaning of the terms of a vocabulary according to a certain view of the world (GUARINO, 1998) 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., (JIRKOVSKY; OBITKO; MARIK, 2017; CHEVALLIER; FINANCE; BOULAKIA, 2020)). 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 (CHEVALLIER; FINANCE; BOULAKIA, 2020; SUN et al., 2020; LIU et al., 2020; JOUAN; HALLOT, 2020; SINGH et al., 2021). It also enables a common understanding and interrelation of concepts in different domains or disciplines, which can support the interoperation of DTs (PLATENIUS-MOHR et al., 2019) 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 of quality assessment; etc.

- **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 (KIOURTIS et al., 2018) and for I4.0 (PLATENIUS-MOHR et al., 2019). Future work will have to address more complex issues considering federations of DTs.

- **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. (RAPTIS; PASSARELLA; CONTI, 2019) 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 occur faster than proprietary software. Therefore, we understand that using open-source tools, cloud, edge, and fog computing will increasingly present in DT solutions.

- **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 Open Subsurface Data Universe[2] (OSDU) 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. Correia et al. (2022) 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. Standard interfaces can simplify the management of the various resources required for DT implementation by providing a unified way for the user.

- **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 (HERSCHEL;

[2] <https://osduforum.org/>

DIESTELKÄMPER; LAHMAR, 2017). This contributes to transparency and provides context for the results/decisions they generate. It must also be possible to follow the quality and reliability of the data, audit data traces, allow the replication of procedures, assign properties or responsibilities (e.g., error), or provide informational context that can be consulted and analyzed (SIMMHAN; PLALE; GANNON, 2005). This is an important gap among the selected studies, as only three of them addressed the traceability of the data (LIU et al., 2021; ZHUANG; GONG; LIU, 2021; SUHAIL et al., 2021).

A promising technology for this purpose is blockchain (ZHENG et al., 2018), which is a shared, immutable ledger that facilitates the process of recording transactions and tracking assets in a business network. Suhail et al. (2021) envisioned a blockchain-based framework for the I4.0 that enables following 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 CONCLUSION AND FUTURE WORKS

This work aimed to investigate state-of-the-art in terms of data management in the context of DTs through a systematic literature review process. Based on the knowledge produced by selected primary studies, we answered the five defined research questions, which allowed us to synthesize knowledge about the domains of DTs application, their functions and uses, the types of data involved, what solutions were proposed related to data management and the type of technological infrastructure. In addition, we were able to identify the challenges and research opportunities.

This SLR was motivated by the absence of a survey or review on data management in DTs. Most surveys focus on understanding the concepts, properties, and main use cases/applications (SEMERARO et al., 2021; BARRICELLI; CASIRAGHI; FOGLI, 2019; FULLER et al., 2020; JONES et al., 2020). In terms of data management, (TAO et al., 2019) highlights an explicit component in the architecture of a DT to handle data management, and (RAPTIS; PASSARELLA; CONTI, 2019) outlines architectural designs considering data-related factors (presence, coordination, and computing). Therefore, our SLR presents a novel perspective by considering selected data management issues extracted from the value chain of Big Data activities: data integration, interoperability, data quality and search for data close to the layer of applications in the context of DTs.

In a nutshell, we summarize the answers to the defined research questions as follows:

- *RQ1: For which domains the DT solutions were proposed?* The three major areas/domains of DTs application are I4.0, smart cities, and healthcare. Most selected primary studies target I4.0 (59.64%), which indicates DTs are a more mature concept in this area even from a DM standpoint.

- *RQ2: Under the perspective of data usage, for which functions were the DTs proposed?* The main uses and functions of the studies reported by the studies were decision support (38.59%), asset monitoring (24.56%), and optimization of production/process/logistics (21.05%). The last two functions show that DT is often used or designed to monitor or optimize functions over time, which deal with real-time or right-time. Therefore, this also reflects on data management requirements, and the DM component needs to be able to meet such needs adequately.

  In addition, we realized most systems are actually Digital Shadows, considering the distinction as defined in (FULLER et al., 2020), which means that most propositions consume data produced by physical systems without an explicit closed loop from the virtual system to the physical one.

- *RQ3: What types of data the proposed solutions consider?* We identified that 40.35% of studies deal with both real-time and historical data, while 24.56% strictly reported data streams. Although few studies have explained the data types, we identified the presence of heterogeneous data and heterogeneous sensing technology. Therefore, data and source types are a reflection of DT's functions and usage. If the DT is designed for asset monitoring, dealing with data streams is a crucial requirement.

- *RQ4: What solutions were proposed for the DM issues addressed?* Among the studies related to interoperability, we find solutions focused on data interoperability, semantics, and between DTs. Studies that proposed integration solutions were categorized as centralized repository, ad hoc modeling, modeling with standards, modeling with an ontological layer, integration method, and semantic enrichment. For the issue of data search, we find solutions based on structural data search and semantic search. Finally, we find solutions for data quality using statistical methods, rules-based methods to detect deviations, and pipeline operations to pre-process and improve the quality of data.

- *RQ5: What kind of technological infrastructure is considered?* The types of technological infrastructure considered by studies were categorized in cloud computing, hybrid computing (cloud, edge, fog), and DBMS. This categorization was based on the infrastructure for data processing and storage.

Quality assessment enabled us to refine the selection of the studies and to explain the results by the quality differences. The selected studies have an average score of 7.39, which is reasonable. However, our expectation was to find more robust solutions regarding the data management aspects. We concluded that explicitly detailing the data management functions in DTs is a relatively new subject of study, particularly because the comparison with the related works and discussion of the limitations of the studies' propositions were quality questions with low scores on average. Therefore, more research is needed in both theoretical and practical terms to advance knowledge about data management to leverage the creation of practical and useful solutions for implementing more functional and effective DTs.

The selection process was a major challenge for this study. We try to be exhaustive and collect as many relevant articles as possible, but finding all relevant work already published is impossible (BROCKE et al., 2015). To handle the huge volume of articles found in the digital libraries (1,838 studies in our case), we proposed a three-step process for screening and selecting the studies to handle this volume. In this way, the selection was iteratively refined based on the summary (title, abstract, keywords), then on the overview of the article (mainly introduction and selected portions/figures of the studies), and finally, by full-reading the papers. This process enabled minimizing the bias in screening the papers using two independent readers and consolidating the criteria for study selection/exclusion later, based on the full reading. Therefore, we are confident that our strategy and methodological process adopted have been well-defined and executed, which allowed us to maximize the scope and quality of our review.

The results reported in this work cover studies between 2014 and 2021. As DTs have become a real trend topic, it is expected that the number of studies will significantly increase. We attempted to update our search to cover studies published in 2022, but the volume from the DLs search was so significant (about 2.300 studies) that we realized that we would not be able to process them in due time. We recognize that collecting new studies as future work is important, including to measure data management advances in the context of DTs.

As future work, our SLR could be complemented by considering more recent literature. However, 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, having an overview of all its aspects is important.

Our research resulted in three publications, detailed below. In addition, an article summarizing this SLR was submitted to a journal (Knowledge and Information Systems - KAIS) and is currently under peer review.

1. CORREIA, Jaqueline B. et al. Data Management in Digital Twins for the Oil and Gas Industry: beyond the OSDU Data Platform. Journal of Information and Data Management, v. 13, n. 3, 2022.

2. CORREIA, Jaqueline B.; ABEL, Mara; BECKER, Karin. Nucleo de Fusão de Dados de um Gêmeo Digital da Indústria de Petróleo e Gás. In: Anais do XXXVI Simpósio Brasileiro de Bancos de Dados. SBC, 2021. p. 343-348.

3. CORREIA, Jaqueline B. et al. Comparing ARIMA and LSTM models to predict time series in the oil industry. In: Anais do IX Symposium on Knowledge Discovery, Mining and Learning. SBC, 2021. p. 129-136.

4. CORREIA, Jaqueline B.; ABEL, Mara; BECKER, Karin. Data Management in Digital Twins: a Systematic Literature Review. Submitted to KAIS.

REFERENCES

ACQUAVIVA, A. et al. Forecasting heating consumption in buildings: A scalable full-stack distributed engine. **Electronics (Switzerland)**, MDPI AG, v. 8, n. 5, 2019. ISSN 20799292.

AGARWAL, P.; MCNEILL, S. Real-time cleaning of time-series data for a floating system digital twin. In: **Proc. of the Annual Offshore Technology Conference**. [S.l.: s.n.], 2019. v. 2019-May. ISBN 9781613996416. ISSN 01603663.

AHMADI-ASSALEMI, G. et al. Digital twins for precision healthcare. In: **Cyber defence in the age of AI, Smart societies and augmented humanity**. [S.l.]: Springer, 2020. p. 133–158.

AL-ISMAEL, M.; AL-TURKI, A.; AL-DARRAB, A. Reservoir simulation well data exchange towards digital transformation and live earth models. In: **Proc. of the 2020 International Petroleum Technology Conference (IPTC)**. [S.l.: s.n.], 2020. ISBN 9781613996751.

AL-MEKHLAL, M.; KHWAJA, A. A. A synthesis of big data definition and characteristics. In: IEEE. **Proc. of the IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)**. [S.l.], 2019. p. 314–322.

ALHUMUD, M. A.; HOSSAIN, M. A.; MASUD, M. Perspective of health data interoperability on cloud-based Medical Cyber-Physical Systems. In: **Proc. of the 2016 IEEE International Conference on Multimedia and Expo Workshop (ICMEW)**. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2016. p. 1–6. ISBN 9781509015528.

ALWAN, A. A. et al. HADES: A Hybrid Anomaly Detection System for Large-Scale Cyber-Physical Systems. In: **proc. of the 5th International Conference on Fog and Mobile Edge Computing (FMEC)**. [S.l.: s.n.], 2020. p. 136–142. ISBN 9781728172163.

ANDIA, P.; ISRAEL, R. R. A cyber-physical approach to early kick detection. In: **Society of Petroleum Engineers - Proc. of the IADC/SPE Drilling Conference and Exhibition, DC 2018**. [S.l.: s.n.], 2018. v. 2018-March, p. 6–8. ISBN 9781613995686.

ANSARI, F.; GLAWAR, R.; NEMETH, T. PriMa: a prescriptive maintenance model for cyber-physical production systems. **International Journal of Computer Integrated Manufacturing**, v. 32, n. 4-5, p. 482–503, 2019.

AZZAM, A. et al. The CitySpin platform: A CPSS environment for city-wide infrastructures. In: **CEUR Workshop Proceedings**. [s.n.], 2019. v. 2530, p. 57–64. ISSN 16130073. Available from Internet: <https://www.w3.org/TR/sparql11-query/>.

BARRICELLI, B. R.; CASIRAGHI, E.; FOGLI, D. A survey on digital twin: Definitions, characteristics, applications, and design implications. **IEEE Access**, IEEE, v. 7, n. Ml, p. 167653–167671, 2019. ISSN 21693536.

BLEIHOLDER, J.; NAUMANN, F. Data fusion. **ACM computing surveys (CSUR)**, ACM New York, NY, USA, v. 41, n. 1, p. 1–41, 2009.

BLUM, M.; SCHUH, G. Towards a data-oriented optimization of manufacturing processes a real-time architecture for the order processing as a basis for data analytics methods. In: **Proc. of the 19th International Conference on Enterprise Information Systems (ICEIS)**. [S.l.]: SciTePress, 2017. v. 1, p. 257–264. ISBN 9789897582479.

BRACKEL, H. U. et al. An open approach to drilling systems automation. In: **Society of Petroleum Engineers - Proc. of the SPE Asia Pacific Oil and Gas Conference and Exhibition**. [S.l.: s.n.], 2018. ISBN 9781613995952.

BRECHER, C. et al. Gaining IIoT insights by leveraging ontology-based modelling of raw data and digital shadows. In: **Proc. of the 2021 4th IEEE International Conference on Industrial Cyber-Physical Systems (ICPS)**. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2021. p. 231–236. ISBN 9781728162072.

BROCKE, J. V. et al. Standing on the shoulders of giants: Challenges and recommendations of literature search in information systems research. **Communications of the association for information systems**, v. 37, n. 1, p. 9, 2015.

BUJARI, A. et al. IPPODAMO: A digital twin support for smart cities facility management. In: **Proc. of the 2021 Conference on Information Technology for Social Good**. [S.l.]: Association for Computing Machinery, Inc, 2021. p. 49–54. ISBN 9781450384780.

CARDOSO, B. B. et al. Data lake architecture for distribution system operator. In: **Proc. of the 2021 IEEE Power and Energy Society Innovative Smart Grid Technologies Conference (ISGT)**. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2021. p. 1–5. ISBN 9781728188973.

CHAPMAN, A. et al. Dataset search: a survey. **VLDB J.**, v. 29, n. 1, p. 251–272, 2020. Available from Internet: <https://doi.org/10.1007/s00778-019-00564-x>.

CHEN, S. et al. Top-Down Human-Cyber-Physical Data Fusion Based on Reinforcement Learning. **IEEE Access**, Institute of Electrical and Electronics Engineers Inc., v. 8, p. 134233–134245, 2020. ISSN 21693536.

CHEVALLIER, Z.; FINANCE, B.; BOULAKIA, B. C. A reference architecture for smart building digital twin. In: **CEUR Workshop Proceedings**. [S.l.: s.n.], 2020. v. 2615. ISSN 16130073.

CHU, X. et al. Data cleaning: Overview and emerging challenges. In: **Proceedings of the 2016 international conference on management of data**. [S.l.: s.n.], 2016. p. 2201–2206.

CLOUTIER, R. et al. The concept of reference architectures. **Systems Engineering**, v. 13, n. 1, p. 14–27, 2010.

CONTI, M. et al. Looking ahead in pervasive computing: Challenges and opportunities in the era of cyber–physical convergence. **Pervasive and Mobile Computing**, v. 8, n. 1, p. 2–21, 2012. ISSN 1574-1192.

CORREIA, J. B. et al. Data management in digital twins for the oil and gas industry: beyond the osdu data platform. **Journal of Information and Data Management**, v. 13, n. 3, 2022.

COUTO, J. et al. A mapping study about data lakes: An improved definition and possible architectures. In: PERKUSICH, A. (Ed.). **Proc. of the 31st International Conference on Software Engineering and Knowledge Engineering, SEKE 2019, Hotel Tivoli, Lisbon, Portugal, July 10-12, 2019**. [S.l.]: KSI Research Inc. and Knowledge Systems Institute Graduate School, 2019. p. 453–578.

CURRY, E. The big data value chain: Definitions, concepts, and theoretical approaches. In: ____. **New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe**. Cham: Springer International Publishing, 2016. p. 29–37. ISBN 978-3-319-21569-3.

DAI, J. et al. Cyber physical power system modeling and simulation based on graph computing. In: **Proc. of the 2017 IEEE Conference on Energy Internet and Energy System Integration (EI2)**. [S.l.: s.n.], 2017. v. 2018-January, p. 1–6. ISBN 9781538614273.

DAO, M. S. et al. A real-time complex event discovery platform for cyber-physical-social systems. In: **ICMR 2014 - Proceedings of the ACM International Conference on Multimedia Retrieval 2014**. [S.l.: s.n.], 2014. p. 201–208. ISBN 1595930361.

DENG, T.; ZHANG, K.; SHEN, Z.-J. M. A systematic review of a digital twin city: A new pattern of urban governance toward smart cities. **Journal of Management Science and Engineering**, v. 6, n. 2, p. 125–134, 2021. ISSN 2096-2320.

DENNEY, A. S.; TEWKSBURY, R. How to write a literature review. **Journal of criminal justice education**, Taylor & Francis, v. 24, n. 2, p. 218–234, 2013.

DEREN, L.; WENBO, Y.; ZHENFENG, S. Smart city based on digital twins. **Computational Urban Science**, Springer, v. 1, n. 1, p. 1–11, 2021.

DERMEVAL, D. et al. Applications of ontologies in requirements engineering: a systematic review of the literature. **Requirements Engineering**, Springer, v. 21, n. 4, p. 405–437, 2016.

DOAN, A.; HALEVY, A.; IVES, Z. **Principles of data integration**. [S.l.]: Elsevier, 2012.

DONG, X. L.; SRIVASTAVA, D. **Big Data Integration**. [S.l.]: Morgan & Claypool Publishers, 2015.

ELAYAN, H.; ALOQAILY, M.; GUIZANI, M. Digital twin for intelligent context-aware iot healthcare systems. **IEEE Internet of Things Journal**, v. 8, n. 23, p. 16749–16757, 2021.

FAN, C. et al. Disaster City Digital Twin: A vision for integrating artificial and human intelligence for disaster management. **International Journal of Information Management**, Elsevier Ltd, v. 56, feb 2021. ISSN 02684012.

FINK, A. **Conducting research literature reviews: From the internet to paper**. [S.l.]: Sage publications, 2019.

FULLER, A. et al. Digital Twin: Enabling Technologies, Challenges and Open Research. **IEEE Access**, Institute of Electrical and Electronics Engineers Inc., v. 8, p. 108952–108971, 2020. ISSN 21693536.

GERACI. Ieee standard computer dictionary: A compilation of ieee standard computer glossaries. **IEEE Std 610**, p. 1–217, 1991.

GIFTY, R.; BHARATHI, R.; KRISHNAKUMAR, P. Faulty-data detection and data quality measure in cyber–physical systems through Weibull distribution. **Computer Communications**, Elsevier B.V., v. 150, p. 262–268, jan 2020. ISSN 1873703X.

GLASZIOU, P. et al. How to review the evidence: systematic identification and review of the scientific literature. National Health & Medical Research Council, 2000.

GÓMEZ-BERBÍS, J. M.; AMESCUA-SECO, A. de. Sedit: Semantic digital twin based on industrial iot data management and knowledge graphs. Springer, v. 1124 CCIS, p. 178–188, 2019. ISSN 18650937.

GROVER, P.; KAR, A. K. Big data analytics: A review on theoretical contributions and tools used in literature. **Global Journal of Flexible Systems Management**, Springer, v. 18, n. 3, p. 203–229, 2017.

GUARINO, N. Formal ontology and information systems. In: **Proc. of the International Conference on Formal Ontology and Information Systems (FOIS'98)**. Amsterdam, Netherlands: IOS Press, 1998. p. 3–15.

GüRDüR, D.; ASPLUND, F. A systematic review to merge discourses: Interoperability, integration and cyber-physical systems. **Journal of Industrial Information Integration**, v. 9, p. 14 – 23, 2018.

HÄNEL, A. et al. Impact of Cyber-physically enhanced manufacturing on the product requirement documentation in high-tech applications. Elsevier B.V., v. 102, p. 210–215, 2021. ISSN 22128271.

HERSCHEL, M.; DIESTELKÄMPER, R.; LAHMAR, H. B. A survey on provenance: What for? what form? what from? **The VLDB Journal**, Springer, v. 26, n. 6, p. 881–906, 2017.

HINOJOSA-PALAFOX, E. A. et al. Towards an Architectural Design Framework for Data Management in Industry 4.0. In: **Proc. of the 2019 7th International Conference in Software Engineering Research and Innovation (CONISOFT)**. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2019. p. 191–200. ISBN 9781728125244.

HOOS, E.; HIRMER, P.; MITSCHANG, B. Context-aware decision information packages: An approach to human-centric smart factories. In: KIRIKOVA, M.; NØRVÅG, K.; PAPADOPOULOS, G. A. (Ed.). **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. [S.l.]: Springer International Publishing, 2017. (Lecture Notes in Computer Science, v. 10509 LNCS), p. 42–56. ISBN 9783319669168. ISSN 16113349.

HUANG, W.; DAI, W. Knowledge storage and acquisition for industrial cyber-physical systems based on non-relational database. In: **Proc. of the 43rd Annual Conference of the IEEE Industrial Electronics Society (IECON)**. [S.l.: s.n.], 2017. v. 2017-January, p. 6671–6676. ISBN 9781538611272.

HUSSAIN, I.; PARK, S. J. Big-ECG: Cardiographic Predictive Cyber-Physical System for Stroke Management. **IEEE Access**, Institute of Electrical and Electronics Engineers Inc., v. 9, p. 123146–123164, 2021. ISSN 21693536.

JIANG, Y.; CHEN, C.; LIU, X. Assembly Process Knowledge Graph for Digital Twin. In: **Proc. of the IEEE International Conference on Automation Science and Engineering**. [S.l.]: IEEE Computer Society, 2021. v. 2021-Augus, p. 758–763. ISBN 9781665418737. ISSN 21618089.

JIRKOVSKY, V.; OBITKO, M.; MARIK, V. Understanding data heterogeneity in the context of cyber-physical systems integration. **IEEE Transactions on Industrial Informatics**, IEEE Computer Society, v. 13, n. 2, p. 660–667, apr 2017. ISSN 15513203.

JONES, D. et al. Characterising the Digital Twin: A systematic literature review. **CIRP Journal of Manufacturing Science and Technology**, Elsevier, v. 29, p. 36–52, may 2020. ISSN 17555817.

JØRGENSEN, M. Estimation of software development work effort: Evidence on expert judgement and formal models. int. **Journal of Forecasting**, 2007.

JOUAN, P.; HALLOT, P. Digital twin: Research framework to support preventive conservation policies. **ISPRS International Journal of Geo-Information**, MDPI AG, v. 9, n. 4, apr 2020. ISSN 22209964.

KADADI, A. et al. Challenges of data integration and interoperability in big data. In: IEEE. **2014 IEEE international conference on big data (big data)**. [S.l.], 2014. p. 38–40.

KASRIN, N. et al. Data-sharing markets for integrating IoT data processing functionalities. **CCF Transactions on Pervasive Computing and Interaction**, Springer, v. 3, n. 1, p. 76–93, mar 2021. ISSN 25245228.

KHAN, K. S. et al. **Undertaking systematic reviews of research on effectiveness: CRD's guidance for carrying out or commissioning reviews**. [S.l.]: NHS Centre for Reviews and Dissemination, 2001.

KIOURTIS, A. et al. Exploring the complete data path for data interoperability in cyber-physical systems. v. 12, n. 4, p. 339–349, 2018. ISSN 17400570.

KIRCHEN, I. et al. Metrics for the evaluation of data quality of signal data in industrial processes. In: **Procof the 2017 IEEE 15th International Conference on Industrial Informatics (INDIN)**. [S.l.: s.n.], 2017. p. 819–826. ISBN 9781538608371.

KITCHENHAM, B.; CHARTERS, S. **Guidelines for performing Systematic Literature Reviews in Software Engineering**. [S.l.], 2007.

KITCHENHAM, B.; MENDES, E.; TRAVASSOS, G. H. A systematic review of cross-vs. within-company cost estimation studies. In: **10th International Conference on Evaluation and Assessment in Software Engineering (EASE) 10**. [S.l.: s.n.], 2006. p. 1–10.

KONG, T. et al. Data Construction Method for the Applications of Workshop Digital Twin System. Elsevier B.V., v. 58, p. 323–328, jan 2021. ISSN 02786125.

KOUSI, N. et al. Digital twin for adaptation of robots' behavior in flexible robotic assembly lines. **Procedia Manufacturing**, Elsevier B.V., v. 28, p. 121–126, 2019. ISSN 23519789.

KRONBERGER, P. et al. The Digitalization Journey of the Brage Digital Twin. In: . [S.l.: s.n.], 2020. (Proc. of the SPE Norway Subsurface Conference, Day 2 Tue, November 03, 2020).

LANDOLFI, G. et al. Intelligent value chain management framework for customized assistive healthcare devices. In: **Procedia CIRP**. [S.l.]: Elsevier B.V., 2018. v. 67, p. 583–588. ISSN 22128271.

LEAL, G. d. S. S.; GUÉDRIA, W.; PANETTO, H. Interoperability assessment: A systematic literature review. **Computers in Industry**, Elsevier, v. 106, p. 111–132, 2019.

LIU, C. et al. Digital Twin-enabled Collaborative Data Management for Metal Additive Manufacturing Systems. **Journal of Manufacturing Systems**, Elsevier B.V., v. 62, p. 857–874, 2022. ISSN 02786125.

LIU, J. et al. A digital twin-driven approach towards traceability and dynamic control for processing quality. **Advanced Engineering Informatics**, Elsevier Ltd, v. 50, oct 2021. ISSN 14740346.

LIU, J. et al. The Research of Ontology-based Digital Twin Machine Tool Modeling. In: **Proc. of the IEEE 6th International Conference on Computer and Communications (ICCC)**. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2020. p. 2130–2134. ISBN 9781728186351.

LU, H. et al. Oil and Gas 4.0 era: A systematic review and outlook. **Computers in Industry**, Elsevier B.V., v. 111, p. 68–90, 2019. ISSN 01663615.

LU, Q. et al. Developing a Digital Twin at Building and City Levels: Case Study of West Cambridge Campus. **Journal of Management in Engineering**, v. 36, n. 3, 2020. ISSN 0742-597X.

LU, Q. et al. Digital twin-enabled anomaly detection for built asset monitoring in operation and maintenance. **Automation in Construction**, Elsevier B.V., v. 118, oct 2020. ISSN 09265805.

LV, Q. et al. A digital twin-driven human-robot collaborative assembly approach in the wake of COVID-19. **Journal of Manufacturing Systems**, Elsevier B.V., v. 60, p. 837–851, jul 2021. ISSN 02786125.

MACDONELL, S. et al. How reliable are systematic reviews in empirical software engineering? **IEEE Transactions on Software Engineering**, IEEE, v. 36, n. 5, p. 676–687, 2010.

MOYNE, J. et al. A Requirements Driven Digital Twin Framework: Specification and Opportunities. **IEEE Access**, Institute of Electrical and Electronics Engineers Inc., v. 8, p. 107781–107801, 2020. ISSN 21693536.

MUNIR, K.; ANJUM, M. S. The use of ontologies for effective knowledge modelling and information retrieval. **Applied Computing and Informatics**, Elsevier, v. 14, n. 2, p. 116–126, 2018.

MYKKÄNEN, J. A.; TUOMAINEN, M. P. An evaluation and selection framework for interoperability standards. **Information and Software Technology**, Elsevier, v. 50, n. 3, p. 176–197, 2008.

NAPOLEÃO, B. et al. Practical similarities and differences between systematic literature reviews and systematic mappings: a tertiary study. In: **SEKE**. [S.l.: s.n.], 2017. p. 85–90.

NÚÑEZ-VALDEZ, E. et al. Incremental Hierarchical Clustering driven Automatic Annotations for Unifying IoT Streaming Data. **International Journal of Interactive Multimedia and Artificial Intelligence**, v. 6, n. 2, p. 15, 2020. ISSN 1989-1660.

OAKES, B. et al. Structuring and accessing knowledge for historical and streaming digital twins. In: **CEUR Workshop Proceedings**. [S.l.: s.n.], 2021. v. 2941, p. 1–13. ISSN 16130073.

OBRST, L. Ontologies for semantically interoperable systems. In: **Proceedings of the twelfth international conference on Information and knowledge management**. [S.l.: s.n.], 2003. p. 366–369.

OKOLI, C. A guide to conducting a standalone systematic literature review. **Commun. Assoc. Inf. Syst.**, v. 37, p. 43, 2015.

PAPADAKIS, G. et al. Blocking and filtering techniques for entity resolution: A survey. **ACM Computing Surveys (CSUR)**, ACM New York, NY, USA, v. 53, n. 2, p. 1–42, 2020.

PERNICI, B. et al. AgileChains: Agile supply chains through smart digital twins. In: **Proc. of the 30th European Safety and Reliability Conference, ESREL 2020 and 15th Probabilistic Safety Assessment and Management Conference, PSAM 2020**. [S.l.: s.n.], 2020. p. 2678–2684. ISBN 9789811485930.

PETERSEN, K.; VAKKALANKA, S.; KUZNIARZ, L. Guidelines for conducting systematic mapping studies in software engineering: An update. **Information and software technology**, Elsevier, v. 64, p. 1–18, 2015.

PLATENIUS-MOHR, M. et al. Interoperable digital twins in IIoT systems by transformation of information models: A case study with asset administration shell. In: **ACM International Conference Proceeding Series**. [S.l.]: ICST, 2019. ISBN 9781450372077. ISSN 21531633.

PLATENIUS-MOHR, M. et al. File- and api-based interoperability of digital twins by model transformation: An iiot case study using asset administration shell. **Future Generation Computer Systems**, v. 113, p. 94–105, 2020. ISSN 0167-739X. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0167739X20302600>.

PROPER, H. A.; BORK, D.; POELS, G. Towards an ontology-driven approach for digital twin enabled governed IT management. In: **CEUR Workshop Proceedings**. [S.l.: s.n.], 2021. v. 2941, p. 14. ISSN 16130073.

QI, Q.; TAO, F. Digital twin and big data towards smart manufacturing and industry 4.0: 360 degree comparison. **Ieee Access**, IEEE, v. 6, p. 3585–3593, 2018.

RAHM, E. The case for holistic data integration. In: **Proc. of the 20th Advances in Databases and Information Systems (ADBIS)**. Springer, 2016. (Lecture Notes in Computer Science, v. 9809), p. 11–27. Available from Internet: <https://doi.org/10.1007/978-3-319-44039-2_2>.

RAHM, E. The case for holistic data integration. In: POKORNÝ, J. et al. (Ed.). **Advances in Databases and Information Systems**. Cham: Springer International Publishing, 2016. p. 11–27. ISBN 978-3-319-44039-2.

RAO, T. R. et al. The big data system, components, tools, and technologies: a survey. **Knowledge and Information Systems**, v. 60, n. 3, p. 1165–1245, 2019. ISSN 0219-3116. Available from Internet: <https://doi.org/10.1007/s10115-018-1248-0>.

RAPTIS, T. P.; PASSARELLA, A.; CONTI, M. Data management in industry 4.0: State of the art and open challenges. **IEEE Access**, Institute of Electrical and Electronics Engineers Inc., v. 7, p. 97052–97093, 2019. ISSN 21693536.

RENNER, S. A community of interest approach to data interoperability. In: CITESEERX SAN DIEGO, CA. **Federal database colloquium**. [S.l.], 2001. v. 1, p. 2.

ROWLEY, J.; SLACK, F. Conducting a literature review. **Management research news**, Emerald Group Publishing Limited, 2004.

RYBNYTSKA, O. et al. PGFMU: Integrating data management with physical system modelling. In: **Proc. of the Conference on Advances in Database Technology (EDBT)**. [S.l.]: APA, 2020. v. 2020-March, p. 109–120. ISBN 9783893180837. ISSN 23672005.

SAHLAB, N. et al. Knowledge graphs as enhancers of intelligent digital twins. In: **Proc. of the 2021 4th IEEE International Conference on Industrial Cyber-Physical Systems (ICPS)**. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2021. p. 19–24. ISBN 9781728162072.

SAWADOGO, P. N.; DARMONT, J. On data lake architectures and metadata management. **J. Intell. Inf. Syst.**, v. 56, n. 1, p. 97–120, 2021.

SEMERARO, C. et al. Digital twin paradigm: A systematic literature review. **Computers in Industry**, v. 130, 2021. ISSN 01663615.

SHA, K.; ZEADALLY, S. Data quality challenges in cyber-physical systems. **Journal of Data and Information Quality**, Association for Computing Machinery, v. 6, n. 2, jun 2015. ISSN 19361963.

SIDI, F. et al. Data quality: A survey of data quality dimensions. In: IEEE. **Proc. of the 2012 International Conference on Information Retrieval & Knowledge Management**. [S.l.], 2012. p. 300–304.

SIMMHAN, Y. L.; PLALE, B.; GANNON, D. A survey of data provenance in e-science. **SIGMOD Rec.**, Association for Computing Machinery, New York, NY, USA, v. 34, n. 3, p. 31–36, sep. 2005. ISSN 0163-5808.

SINGH, S. et al. Data management for developing digital twin ontology model. **Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture**, SAGE Publications Ltd, v. 235, n. 14, p. 2323–2337, dec 2021. ISSN 20412975.

SIRCAR, A. et al. Digital twin in hydrocarbon industry. **Petroleum Research**, Elsevier, 2022.

SIVALINGAM, K. et al. A review and methodology development for remaining useful life prediction of offshore fixed and floating wind turbine power converter with digital twin technology perspective. In: **Proc. of the 2nd International Conference on Green Energy and Applications (ICGEA)**. [S.l.: s.n.], 2018. p. 197–204.

STAPLES, M.; NIAZI, M. Experiences using systematic review guidelines. **Journal of Systems and Software**, Elsevier, v. 80, n. 9, p. 1425–1437, 2007.

STUDER, R.; BENJAMINS, V. R.; FENSEL, D. Knowledge engineering: principles and methods. **Data & knowledge engineering**, Elsevier, v. 25, n. 1-2, p. 161–197, 1998.

SUHAIL, S. et al. Trustworthy Digital Twins in the Industrial Internet of Things with Blockchain. **IEEE Internet Computing**, Institute of Electrical and Electronics Engineers Inc., p. 1–8, 2021. ISSN 19410131.

SUN, S. et al. Data handling in industry 4.0: Interoperability based on distributed ledger technology. **Sensors (Switzerland)**, MDPI AG, v. 20, n. 11, jun 2020. ISSN 14248220.

TAO, F. et al. Digital Twin in Industry: State-of-the-Art. **IEEE Transactions on Industrial Informatics**, IEEE, v. 15, n. 4, p. 2405–2415, 2019. ISSN 15513203.

TAO, F.; ZHANG, M. Digital twin shop-floor: A new shop-floor paradigm towards smart manufacturing. **IEEE Access**, v. 5, p. 20418–20427, 2017.

TIWARI, S.; GUPTA, A. A systematic literature review of use case specifications research. **Information and Software Technology**, Elsevier, v. 67, p. 128–158, 2015.

VENKATESH, K. et al. Challenges and research disputes and tools in big data analytics. **International Journal of Engineering and Advanced Technology**, v. 6, p. 1949–1952, 2019.

VIVI, Q. L. et al. Developing a dynamic digital twin at a building level: Using Cambridge campus as case study. In: **Proc. of the International Conference on Smart Infrastructure and Construction (ICSIC): Driving Data-Informed Decision-Making**. [S.l.]: ICE Publishing, 2019. p. 67–75. ISBN 9780727764669.

WANASINGHE, T. R. et al. Digital Twin for the Oil and Gas Industry: Overview, Research Trends, Opportunities, and Challenges. **IEEE Access**, v. 8, p. 104175–104197, 2020. ISSN 21693536.

WANG, R. Y.; STRONG, D. M. Beyond accuracy: What data quality means to data consumers. **Journal of management information systems**, Taylor & Francis, v. 12, n. 4, p. 5–33, 1996.

WANG, T.; CHENG, L. Large-scale semantic knowledge acquisition and application for cyber-physical-social systems. In: **Proc. of the 2021 IEEE 1st International Conference on Digital Twins and Parallel Intelligence (DTPI)**. [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2021. p. 282–285. ISBN 9781665433372.

WANG, Y.; ZHOU, X. Spatio-temporal semantic enhancements for event model of cyber-physical systems. In: **Proc. of the 2014 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)**. [S.l.: s.n.], 2014. p. 813–818. ISBN 9781479952748.

WU, D.; WANG, H.; SEIDU, R. Toward A Sustainable Cyber-Physical System Architecture for Urban Water Supply System. In: **Proc. of the IEEE Congress on Cybermatics: 2020 IEEE International Conferences on Internet of Things (iThings), IEEE Green Computing and Communications (GreenCom), IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartD)**. [S.l.]: IEEE, 2020. p. 482–489. ISBN 9781728176475.

XU, L. D.; XU, E. L.; LI, L. Industry 4.0: state of the art and future trends. **International journal of production research**, Taylor & Francis, v. 56, n. 8, p. 2941–2962, 2018.

YU, C.; JAGADISH, H. Querying complex structured databases. In: **Proceedings of the 33rd international conference on Very large data bases**. [S.l.: s.n.], 2007. p. 1010–1021.

YU, W. et al. Implementation of industrial cyber physical system: Challenges and solutions. In: **Proc. of the 2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS)**. [S.l.: s.n.], 2019. p. 173–178. ISBN 9781538685006.

ZHANG, C.; JI, W. Digital twin-driven carbon emission prediction and low-carbon control of intelligent manufacturing job-shop. Elsevier B.V., v. 83, p. 624–629, 2019. ISSN 22128271.

ZHANG, Q. et al. Three-dimensional visualization interactive system for digital twin workshop. **Journal of Southeast University (English Edition)**, v. 37, n. 2, p. 137–152, 2021. ISSN 10037985.

ZHANG, Y. et al. Health-CPS: Healthcare cyber-physical system assisted by cloud and big data. **IEEE Systems Journal**, Institute of Electrical and Electronics Engineers Inc., v. 11, n. 1, p. 88–95, mar 2017. ISSN 19379234.

ZHENG, Z. et al. Blockchain challenges and opportunities: a survey. **Int. J. Web Grid Serv.**, v. 14, p. 352–375, 2018.

ZHUANG, C.; GONG, J.; LIU, J. Digital twin-based assembly data management and process traceability for complex products. **Journal of Manufacturing Systems**, Elsevier B.V., v. 58, p. 118–131, jan 2021. ISSN 02786125.

ZONZINI, F. et al. Structural Health Monitoring and Prognostic of Industrial Plants and Civil Structures: A Sensor to Cloud Architecture. **IEEE Instrumentation & Measurement Magazine**, v. 29, n. 9, p. 21–27, 2020.

APPENDIX A — QUALITY ASSESSMENT TABLE OF SELECTED STUDIES

Table A.1 – Quality assessment score of selected studies

ID Study	QQ1	QQ2	QQ3	QQ4	QQ5	QQ6	QQ7	QQ8	QQ9	Total Score
1	1	1	0	1	1	0.5	0	0.5	1	6
2	1	1	0.5	0.5	0.5	0	0	1	0.5	5
3	1	1	1	1	1	1	0	1	1	8
4	1	1	1	1	1	1	0	1	1	8
5	1	1	0.5	1	1	1	0	1	1	7.5
6	1	1	0.5	1	1	1	0.5	1	1	8
7	1	1	0.5	1	1	1	0.5	1	1	8
8	1	1	0	1	1	0.5	0	1	0.5	6
9	1	1	0	1	1	1	1	1	1	8
10	1	1	0.5	1	1	1	0	1	1	7.5
11	1	1	0	1	1	0.5	0	1	1	6.5
12	1	1	0	1	1	0.5	0	1	1	6.5
13	1	1	0.5	1	1	1	0.5	1	1	8
14	1	1	0.5	1	0.5	0.5	0.5	0.5	1	6.5
15	1	1	0	1	1	1	0.5	1	1	7.5
16	1	1	0.5	1	1	1	1	0	1	7.5
17	1	1	0	1	1	0.5	0	0.5	1	6
18	1	1	0	1	1	0.5	0	0.5	1	6
19	1	1	1	1	1	1	0.5	1	1	8.5
20	1	1	0.5	1	1	1	1	1	1	8.5
21	1	1	0	1	1	0.5	0.5	0.5	0.5	6
22	1	1	0	1	1	0.5	0.5	0.5	1	6.5
23	1	1	0	1	1	0.5	0.5	0.5	1	6.5
24	1	1	1	1	1	1	0.5	1	1	8.5
25	1	1	0.5	1	0.5	1	0	1	1	7
26	1	1	0.5	1	1	1	1	1	1	8.5
27	1	1	0.5	1	1	0.5	0	0	1	6
28	1	1	0	1	1	1	0	1	1	7

29	1	1	0.5	1	0.5	0.5	0.5	0.5	1	6.5
30	1	1	0.5	1	1	1	1	1	1	8.5
31	1	0.5	1	1	1	1	1	0	1	7.5
32	1	1	0.5	1	1	1	0.5	1	0.5	7.5
33	1	1	1	1	1	1	1	1	1	9
34	1	1	0.5	1	1	1	1	1	1	8.5
35	1	1	0.5	1	1	1	1	1	1	8.5
36	1	1	1	1	1	1	0.5	1	0.5	8
37	1	1	1	1	1	0.5	0	1	0.5	7
38	1	1	0.5	1	1	1	0	1	1	7.5
39	1	1	0.5	1	1	1	1	1	1	8.5
40	1	1	0.5	1	1	1	1	1	1	8.5
41	1	1	0.5	1	1	0.5	0	0	1	6
42	1	1	0.5	1	1	1	1	1	1	8.5
43	1	1	0.5	1	1	0	0.5	0.5	1	6.5
44	1	1	1	1	1	1	1	0.5	0.5	8
45	1	1	0.5	1	1	1	0.5	1	1	8
46	1	1	1	1	1	0.5	0	0.5	0.5	6.5
47	1	1	1	1	1	1	1	1	1	9
48	1	1	0.5	1	1	0.5	1	1	1	8
49	1	1	0.5	1	1	1	0	0	1	6.5
50	1	1	0.5	1	1	1	1	1	1	8.5
51	1	1	1	1	1	1	0.5	1	1	8.5
52	1	1	1	1	1	1	0.5	1	1	8.5
53	1	1	0.5	1	0.5	1	1	0.5	0.5	7
54	1	1	1	0.5	0.5	0.5	0	0.5	1	6
55	1	1	1	1	1	1	0	0.5	1	7.5
56	1	1	0.5	1	1	0.5	0	0.5	1	6.5
57	1	1	0	1	1	1	0	1	1	7
Average	**1**	**0.99**	**0.52**	**0.98**	**0.95**	**0.81**	**0.45**	**0.78**	**0.92**	**7.39**

Source: The author

APPENDIX B — RESUMO EXPANDIDO

This chapter presents a summary of this master thesis in the Portuguese language, as required by the PPGC Graduate Program in Computing.

Este capítulo apresenta um resumo desta dissertação de mestrado em língua portuguesa, conforme exigido pelo Programa de Pós-Graduação em Computação.

B.1 Introdução e contribuições da dissertação

A crescente popularidade da Internet das Coisas, o advento dos dispositivos vestíveis inteligentes e os avanços contínuos nas técnicas de coleta de dados aumentaram significativamente a quantidade de dados relevantes que podem ser aproveitados para aplicativos inovadores orientados a dados em tempo real. Ao explorar esses dispositivos e várias tecnologias, as informações sobre o mundo físico são transferidas perfeitamente para o mundo virtual, onde são elaboradas para adaptar aplicações e serviços virtuais ao contexto físico, possivelmente modificando/adaptando o próprio mundo físico por meio de atuadores (CONTI et al., 2012). Os gêmeos digitais (GDs) são o próximo passo dessa convergência ciber-física. Os GDs são representações virtuais de objetos físicos, que são totalmente integrados e nos quais a troca automática de dados ocorre de maneira bidirecional (FULLER et al., 2020).

Os GDs estão no centro de inovações disruptivas em diversas áreas (RAPTIS; PASSARELLA; CONTI, 2019). Na manufatura inteligente (Indústria 4.0 - I4.0), o GDs podem cobrir todas as fases do ciclo de vida do produto, incluindo projeto, planejamento, montagem e otimização da fábrica (TAO et al., 2019; FULLER et al., 2020). As empresas do setor de petróleo e gás (O&G) aproveitam essa a inovação para aumentar a produção e maximizar o lucro e ter experiências bem-sucedidas nos campos de petróleo e *pipelining* inteligente, manutenção preditiva e avaliação de riscos (LU et al., 2019; WANASINGHE et al., 2020). Os GDs também podem mudar o conceito de assistência médica digital, onde uma réplica virtual de um paciente pode melhorar a promoção e o controle da saúde, prever tendências futuras usando histórico médico e otimizar as operações de saúde (ELAYAN; ALOQAILY; GUIZANI, 2021). Os GDs de cidades inteligentes visam melhorar a eficiência e a sustentabilidade da logística, consumo de energia, planejamento urbano, gerenciamento de desastres, entre outros (DENG; ZHANG; SHEN, 2021).

Big data e GDs são tecnologias que se reforçam mutuamente (RAO et al., 2019), uma vez que grandes volumes de dados que representam os mundos físicos/virtuais são coletados, transformados e gerados por meio de modelos (por exemplo, simulação, aprendizado de máquina) para agregar valor ao negócio (TAO et al., 2019; JONES et al., 2020). Essas oportunidades exigem lidar com dados em um volume, velocidade e variedade que excedem os recursos dos sistemas tradicionais de gerenciamento de dados, oferecendo valor e veracidade. Nesse contexto, os dados são um recurso fundamental que precisa ser considerado na cadeia de valor de Big Data (CURRY, 2016), que inclui atividades para aquisição de dados, análise, armazenamento, curadoria e uso. *Data Lakes* são um tópico de tendência para resolver problemas de *Big Data* (SAWADOGO; DARMONT, 2021).

Diferentes arquiteturas de GD são propostas na literatura (WANASINGHE et al., 2020). Os GDs anteriores seguem uma arquitetura de três componentes que conecta um sistema físico a um virtual espelhado. Enquanto o espaço físico representa os ativos físicos (e.g., sensores, atuadores), o espaço virtual visa imitar/replicar o ambiente físico com alta fidelidade. GDs após essa arquitetura adotam soluções*ad hoc* para problemas de gerenciamento de dados, como extração de dados e integração de fontes heterogêneas, sanidade de dados, transformação e enriquecimento de dados e consumo de dados pelo ambiente virtual. A existência de silos de dados, o volume de dados e problemas relacionados ao manuseio de múltiplas fontes de dados heterogêneas, formatos e tipos de dados são frequentemente mencionados como desafios significativos (SUN et al., 2020; VIVI et al., 2019; SINGH et al., 2021; SAHLAB et al., 2021).

A arquitetura GD de cinco componentes (TAO; ZHANG, 2017) é uma evolução que inclui explicitamente um componente de gestão de dados. O componente de gestão de dados atua como uma ponte entre todos os subsistemas, servindo como um ponto de ingestão dos dados originais e retorno no momento certo para direcionar o processo de otimização interativa resultante de sua interação. Os trabalhos existentes fornecem a funcionalidade para gerenciar diferentes aspectos dos dados, como limpeza de dados, avaliação de qualidade, transformação, integração, pesquisa, entre outros. Essas funcionalidades de gerenciamento de dados são explicitamente compreendidas em um componente gestão de dados dedicado, conforme proposto por (TAO; ZHANG, 2017) ou distribuída em outros componentes do GD.

Este trabalho apresenta uma Revisão Sistemática da Literatura (RSL) sobre as soluções propostas para questões de gestão de dados no escopo de GDs, na qual está implícito no GD ou explícito como parte de um componente de gestão de dados. Essa RSL foi motivada pela ausência de uma RSL ou revisão da literatura focada em soluções de gerenciamento de dados para GDs e a falta de entendimento do papel central e da funcionalidade do componente gestão de dados. As pesquisas existentes contribuem para a compreensão geral dos conceitos, propriedades e casos de uso primário e aplicações de GDs (SEMERARO et al., 2021; BARRICELLI; CASIRAGHI; FOGLI, 2019; FULLER et al., 2020; JONES et al., 2020). Em relação ao gerenciamento de dados, (RAPTIS; PASSARELLA; CONTI, 2019) apresenta uma revisão da literatura a partir da perspectiva de automação industrial e computação em rede, descrevendo projetos arquitetônicos com base em fatores relacionados a dados (presença, coordenação e computação). Embora (TAO et al., 2019) destaque um componente explícito na arquitetura de um GD para lidar com a gestão dos dados, ele não detalha suas funcionalidades. Argumentamos que o componente gestão de dados pode ser aproximado da funcionalidade de dados e gerenciamento de conhecimento dos *Data Lakes* (KRONBERGER et al., 2020; CORREIA et al., 2022) e visam alcançar uma melhor compreensão das soluções do estado da arte em uma análise de granularidade fina.

Nossa RSL apresenta uma nova perspectiva, considerando e extraindo problemas de gestão de dados da cadeia de valor de Big Data (CURRY, 2016). Uma RSL é definida em (OKOLI, 2015) como “ um método sistemático, explícito, abrangente e reproduzível para identificar, avaliar e sintetizar o corpo existente do trabalho concluído e registrado produzido por pesquisadores, estudiosos e praticantes". Pesquisamos e selecionamos estudos primários através de uma abordagem sistemática e imparcial para lançar luz sobre as soluções gestão de dados propostas para lidar com heterogeneidade de dados, interoperabilidade, integração, pesquisa/descoberta de dados e qualidade no contexto de GDs. Definimos as seguintes perguntas de pesquisa: PP1) *Para quais domínios as soluções de GD foram propostas?*; PP2) *Sob a perspectiva do uso de dados, para quais funções os GDs foram propostos?*; PP3) *Quais tipos de dados as soluções propostas consideram?*; PP4) *Quais soluções foram propostas para os problemas de gestão de dados abordados?*; PP5) *Que tipo de infraestrutura tecnológica é considerada?*.

Nossa RSL complementa e inova o cenário das revisões de literatura existentes sobre o GDs, investigando aspectos de gestão de dados ainda não analisados, expressos pelas questões de pesquisa. As principais contribuições deste trabalho são:

- Uma análise de granularidade fina sob as atividades da cadeia de valor de *Big Data*, destacando os principais problemas a serem abordados pelo componente de gestão de dados em um GD.

- Uma RSL examinando soluções existentes para lidar com a heterogeneidade de dados, interoperabilidade, integração, pesquisa/descoberta de dados e qualidade no GDs. Contextualizamos essas soluções no domínio e na função para as quais o GD foi proposto, o tipo de dados tratados e a infraestrutura tecnológica alavancada. A compilação dessas soluções lança luz sobre a funcionalidade a ser fornecida por um componente gestão de dados de um GD, tendências atuais e oportunidades.

B.2 Metodologia

Uma RSL deve resumir todo a informação atual sobre algum fenômeno de maneira completa e imparpacial. Para isso, neste trabalho foi adotada a metodologia bem definida de revisão sistemática proposta por (KITCHENHAM; CHARTERS, 2007). A metodologia adotada teve três fases, a fase de planejamento do protocolo de revisão, execução do protocolo e escrita do relatório.

Resumidamente o primeiro passo foi identificar a necessidade de uma RSL, a definição das perguntas de pesquisa, a estratégia de busca dos artigos primários, definição dos critérios de inclusão e exclusão para a seleção dos artigos, avaliação da qualidade individual de cada artigo e finalmente a seleção final para a inclusão no estudo.

B.3 Resultados e conclusão

Este trabalho teve como objetivo investigar o estado da arte em termos de gestão de dados no contexto do GDs através de um processo sistemático de revisão de literatura. Com base no conhecimento produzido pelos estudos primários selecionados, respondemos às cinco perguntas de pesquisa definidas, o que nos permitiu sintetizar o conhecimento sobre os domínios da aplicação GDs, suas funções e usos, os tipos de dados envolvidos, quais soluções foram propostas relacionadas ao gerenciamento de dados e o tipo de infraestrutura tecnológica. Além disso, conseguimos identificar os desafios e oportunidades de pesquisa.

Essa RSL foi motivada pela ausência de uma RSL sobre gerenciamento de dados em GDs. A maioria das pesquisas se concentra na compreensão dos conceitos, propriedades e principais casos de uso/aplicações (SEMERARO et al., 2021; BARRICELLI; CASIRAGHI; FOGLI, 2019; FULLER et al., 2020; JONES et al., 2020). Em termos de gerenciamento de dados, (TAO et al., 2019) destaca um componente explícito na arquitetura de um GD para lidar com o gerenciamento de dados e (RAPTIS; PASSARELLA; CONTI, 2019) descreve os projetos arquitetônicos que consideram fatores relacionados a dados (presença, coordenação e computação). Portanto, nossa RSL apresenta uma nova perspectiva, considerando problemas como: integração de dados, interoperabilidade, qualidade dos dados e pesquisa de dados próximos à camada de aplicativos no contexto do GDs.

Em poucas palavras, resumimos as respostas para as perguntas de pesquisa definidas da seguinte forma:

- *PP1: Para quais domínios as soluções de GD foram propostas?* As três principais áreas/domínios de aplicacão de GDs são I4.0, cidades inteligentes e saúde. A maioria dos estudos primários selecionados são da I4.0 (59,64 %), o que indica que os GDs são um conceito mais maduro nessa área, mesmo do ponto de vista de gestão de dados.

- *PP2: Sob a perspectiva do uso de dados, para quais funções os GDs foram propostos?* Os principais usos e funções relatados pelos estudos foram apoio à decisão (38,59%), monitoramento de ativos (24,56%) e otimização de Produção/Processo/Logística (21,05%). As duas últimas funções mostram que o GD é frequentemente usado ou projetado para monitorar ou otimizar funções ao longo do tempo, que lidam com o tempo real ou em tempo correto. Portanto, isso também reflete sobre os requisitos de gestão de dados, e os requisitos do componente específico de gestão de dados, que precisa ser capaz de atender adequadamente a essas necessidades.
  Além disso, percebemos que a maioria dos sistemas são sombras digitais, considerando a distinção conforme definido em (FULLER et al., 2020), o que significa que a maioria das soluções propostas consomem os dados produzidos pelos sistemas físicos sem um *loop* fechado e explícito do sistema virtual para o físico.

- *PP3: Quais tipos de dados as soluções propostas consideram?* Identificamos que 40,35% dos estudos lidam com dados em tempo real e histórico, enquanto 24,56% relataram estritamente *streams* de dados. Embora poucos estudos tenham explicitado os tipos de dados, identificamos a presença de dados heterogêneos e tecnologia de sensoriamento heterogênea. Portanto, dados e tipos de origem são um reflexo das funções e uso do GD. Se o GD for projetado para monitoramento de ativos, lidar com *strems* de dados é um requisito crucial.

- *PP4: Quais soluções foram propostas para os problemas de gestão de dados abordados?* Entre os estudos relacionados à interoperabilidade, encontramos soluções focadas na interoperabilidade de dados, semântica e entre GDs. Estudos que propuseram soluções de integração foram categorizados como repositório centralizado, modelagem *ad hoc*, modelagem com padrões, modelagem com uma camada ontológica, método de integração e enriquecimento semântico. Para a questão da busca de dados, encontramos soluções com base na busca estrutural e busca semântica. Por fim, encontramos soluções para a qualidade dos dados usando métodos estatísticos, métodos baseados em regras para detectar desvios e operações de *pipeline* para pré-processar e melhorar a qualidade dos dados.

- *PP5: Que tipo de infraestrutura tecnológica é considerada?* Os tipos de infraestrutura tecnológica considerados pelos estudos foram categorizados em computação em nuvem, computação híbrida (nuvem, borda, neblina) e sistemas de gerenciamento de banco de dados (SGBDs). Essa categorização foi baseada na infraestrutura para processamento e armazenamento de dados.

A avaliação da qualidade nos permitiu refinar a seleção dos estudos e explicar os resultados pelas diferenças de qualidade. Os estudos selecionados têm uma pontuação média de (7,39), o que é razoável. No entanto, nossa expectativa era encontrar soluções mais robustas em relação aos aspectos de gerenciamento de dados. Concluímos que detalhando explicitamente as funções de gestão de dados nos GDs é um assunto relativamente novo de estudo, principalmente porque a comparação com os trabalhos relacionados e a discussão das limitações das proposições dos estudos foram as questões/critérios de qualidade com as pontuações médias mais baixas. Portanto, são necessárias mais pesquisas em termos teóricos e práticos para promover o conhecimento sobre o gerenciamento de dados para alavancar a criação de soluções práticas e úteis para implementar GDs mais funcionais e eficazes.

O processo de seleção foi um grande desafio para este estudo. Tentamos ser exaustivos e coletar o maior número possível de artigos relevantes, mas encontrar todos os trabalho relevantes já publicados é impossível (BROCKE et al., 2015). Para lidar com o enorme volume de artigos encontrados nas bibliotecas digitais (1.838 estudos em nosso caso), propusemos um processo de três etapas para rastrear e selecionar os estudos para lidar com este volume. Dessa maneira, a seleção foi refinada iterativamente com base no resumo (título, resumo, palavras-chave) e, em seguida, na visão geral do artigo (principalmente introdução, figuras e conclusão) e, finalmente, lendo os artigos integralmente. Esse processo permitiu minimizar o viés na triagem dos trabalhos por meio de dois leitores independentes e consolidando os critérios para seleção/exclusão do estudo posteriormente, com base na leitura completa. Portanto, estamos confiantes de que nossa estratégia e processo metodológico adotados foram bem definidos e executados, o que nos permitiu maximizar o escopo e a qualidade de nossa revisão.

Como trabalho futuro, nossa RSL pode ser complementada considerando a literatura mais recente. No entanto, outros aspectos também podem ser investigados, como arquitetura de dados, armazenamento de dados e operações, segurança de dados e gerenciamento de metadados, gerenciamento de conteúdo e assim por diante. A área de gerenciamento de dados é ampla e inclui várias atividades fundamentais para o sucesso de um GD. Portanto, é importante ter uma visão geral de todos os seus aspectos.