Journal of Intelligent Manufacturing
https://doi.org/10.1007/s10845-025-02570-z

Data management in industry: concepts, systematic review and future directions

Nelson Freitas[1] · Andre Dionisio Rocha[1] · Jose Barata[1]

Received: 22 May 2024 / Accepted: 10 January 2025

Abstract

Data management, particularly in industrial environments, is increasingly vital due to the necessity of handling ever-growing volumes of information, commonly referred to as big data. This survey delves into various papers to comprehend the practices employed within industrial settings concerning data management, by searching for relevant keywords in Q1 Journals related to data management in manufacturing in the databases of WebOfScience, Scopus and IEEE. Additionally, a contextual overview of core concepts and methods related to different aspects of the data management process was conducted. The survey results indicate a deficiency in methodology across implementations of data management, even within the same types of industry or processes. The findings also highlight several key principles essential for constructing an efficient and optimized data management system.

Keywords Data management · Big data · Data pipeline · Manufacturing · Industry

Introduction

With the onset of the Fourth Industrial Revolution, the amount of information produced is increasing substantially each year. Consequently, new techniques and concepts such as big data have emerged, becoming central to data management infrastructure designs, both now and in the future (Ishikiriyama & Gomes, 2019). The primary objective of this paper is twofold. First, it emphasizes the importance of understanding the key characteristics of this data and how the industrial sector is managing it, not only in technical terms but also conceptually and through best practices in data management. To this end, several significant concepts relevant to today's state of the art are discussed, including big data (Singh, 2021), data analytics (Duan & Xu, 2021), data collection/extraction (Achouch, et al., 2022; Munappy et al., 2020), and other essential aspects of data management in industrial and other environments. These main concepts are thoroughly explained with examples and architectures to establish a comprehensive foundation.

This foundation sets the stage for the second objective, the conducting of a survey based on keywords developed after an initial study, with further reasoning detailed in "Survey approach" Sect. The survey aims to explore the data management approaches employed in the industrial sector, identifying common methods across various industries. Additionally, it seeks to understand why certain approaches are underutilized while others are overused. Throughout the survey, several related works were identified, including frameworks (Corallo et al., 2022; Guo et al., 2023; Kozjek et al., 2020; Majeed et al., 2019; Saqlain & Shim et al., 2019; Wang & Luo, 2021; Yu et al., 2022; Zhang et al., 2023), principles (Allian et al., 2021; Corallo et al., 2023a; Gopalakrishnan et al., 2022; Ismail et al., 2019a; Kumar et al., 2021; Lee & Chien, 2022; Liu et al., 2023a; Mitra & Munir, 2019; Raj et al., 2023; Raut et al., 2021; Shukla et al., 2019; Wei et al., 2021), and technologies (Ismail et al., 2019a). However, this survey distinguishes itself by focusing solely on

✉ Nelson Freitas
n.freitas@uninova.pt

Andre Dionisio Rocha
ad.rocha@fct.unl.pt

Jose Barata
jab@uninova.pt

[1] NOVA School of Science and Technology, Center of Technology and Systems (UNINOVA-CTS) and Associated Lab of Intelligent Systems (LASI), NOVA University Lisbon, 2829-516 Lisbon, Portugal

Published online: 15 February 2025

data management, analyzing the methods used and identifying the gaps in the research topic, establishing future avenues of investigation (Corallo et al., 2023b; Tardio et al., 2020).

As the survey delves into data management research, specifically within the industrial context, it aims to address the core challenges and opportunities that arise in this field. The survey provides a thorough analysis of the current practices and technologies used, alongside an exploration of the foundational aspects such as Data Extraction, Preprocessing, Storage, and Processing, which are critical in handling Big Data in industrial environments. These concepts, paired with modern architectures, are explored to understand their practical implementation and relevance across different industrial applications.

The contribution of this work can be divided into several major components. First, a contextual overview of data management applications is conducted, addressing the fundamental aspects of efficiently managing large volumes of data while maintaining its quality. This segment serves to establish key notions and provide the authors' perspectives on various concepts central to data management. It examines applications and architectures, presenting insights into best practices for extracting, storing, and processing industrial data.

Second, the survey section presents an in-depth analysis of data management within various industrial environments. It crosses multiple industries to highlight four critical concepts, focusing on their application in real-world use cases. These concepts—data heterogeneity, real-time data processing, data preprocessing, and cloud computing adoption—serve as benchmarks to assess data management approaches in the sector. This part of the work provides concrete examples derived from the survey to complement the theoretical discussions, showcasing how these concepts are implemented across the industry.

Ultimately, this paper examines the current state of research in data management settings within industrial applications. Despite an abundance of studies in this field, limitations persist, notably the lack of a well-defined methodology for designing data pipelines tailored to industrial environments. Additionally, the paper identifies emerging areas for further exploration, such as leveraging Industry 4.0 reference architectures to advance the systematic construction of data management pipelines.

The paper is structured as follows: Chapter 2 details the survey approach, explaining how the survey was conducted, and the criteria used. In Chapter 3, an overview of core concepts related to data management is presented, including established but limited concepts and state-of-the-art architectures. Chapter 4 focuses on the survey, highlighting four crucial concepts prevalent across various industrial environments and exploring their utilization in industry data management. Chapter 5 discusses the main findings of the survey, presenting conclusions, posing questions, and identifying gaps in the field of data management in industry. Finally, Chapter 6 concludes the paper by summarizing the key points discussed throughout the article.

Survey approach

With the objective of identifying the methodologies used in various industrial fields for data management, not only on the shop floor but also across different sections of the industrial environment, an evidence-based approach was adopted. A Systematic Literature Review (SLR) is a well-established method that spans various research fields and aims to "systematically, replicably, and transparently process to synthesize research results and practices" (Camarinha-Matos, 2016). Conducting an SLR enables a structured, comprehensive approach to reviewing and synthesizing research on a given topic by systematically identifying, evaluating, and summarizing relevant studies according to predefined criteria that help reduce bias. This methodology is particularly useful for addressing specific research questions, as it involves setting clear inclusion and exclusion criteria, analyzing findings, and extracting data and conclusions. Numerous authors across disciplines have employed this approach in their literature reviews, as demonstrated in works like (Camarinha-Matos, 2016; Gholipour & Bastas, 2023). In the manufacturing sector specifically, notable examples of SLRs include studies (Bastas, 2021; Dogan & Birant, 2021).

Therefore, the objective of this chapter is to formulate research questions that guide the survey, and to present the methodology, processes, and criteria used for paper selection.

Research questions

The research questions addressed in this survey are the following:

RQ1: Which data management approaches can be found in industrial applications?

Given that the answer to this question can result in multiple approaches it is important to group these different approaches given their specific characteristics and usefulness for their respective sector. As such, a second research questions arises:

RQ2: Which approaches are more adequate for different types of industrial sectors and processes?

Search process and sources

An initial exploration of the topic was conducted using Google Scholar, aimed at establishing a preliminary direction and gaining a broader understanding of the research landscape. This informal search provided insight into the topic by using various keywords related to 'data' and its utilization or construction in manufacturing environments. Keywords such as "data collection," "data storage," "data extraction," "data analytics," and "big data" were employed, all within the context of manufacturing processes. The initial search yielded a substantial number of meaningful results, which laid a robust foundation for the subsequent survey process.

However, the goal of the survey was not to focus solely on isolated aspects of data management or the data pipeline. Instead, it sought to capture the broader picture, exploring how the industrial sector as a whole is handling and utilizing data. Given the relevance of Industry 4.0, the topic generated a large volume of results. Even when keywords were combined with restrictions, the number of search results remained considerable. This abundance of information highlighted the importance of refining the scope to focus on the most relevant data management practices and trends in industrial settings.

For the search and selection of research papers, the databases of Web of Science, SCOPUS, and IEEE Xplore were utilized. While this approach may limit the number of papers eligible for review, these databases host papers with high impact metrics and are associated with the most relevant journals in the field of manufacturing. Considering these factors, the use of these databases is deemed an acceptable tradeoff. The search was then conducted using the search terms outlined in Table 1.

Inclusion and exclusion criteria

To gather research material for the survey, decisions regarding inclusion and exclusion criteria were made to increase the relevance of the selected papers. The first criteria were to set a time frame for the gathered papers, choosing a span of 5 years to maintain the relevant state of the art within the search scope. Papers from 2019 to October 2023 were considered. Language was restricted to English, and the paper type was limited to journals. Given the expansive nature of the topic, the study initially focused on journals and was later narrowed down to only Q1 journals according to SCImago ("SCImago"., 2023) as of 2023. While this narrowed the scope significantly, it ensured the inclusion of the most relevant works and journals, making the trade-off deemed acceptable and enabling the review.

The chosen research areas were 'engineering and computer science' aligning with the primary focus of the scope. Search terms were applied to the title, abstract, and keywords. With the initial criteria defined, the iterations can be observed in Fig. 1. The total number of papers obtained with the search terms was 11,626. The first step in filtering the papers involved removing duplicates and processing only Q1 journal papers, resulting in a reduction to 2786 papers. To further narrow down the selection, only papers with 'data,' 'pipeline,' or 'management' in the title were considered, bringing the focus closer to the authors' objectives. After this query, the titles of the papers were reviewed, and 160 were selected for abstract reading. Following the abstract reading, half of the papers did not fully align with the scope of the research. Consequently, the final set of 78 papers was thoroughly reviewed. However, two were excluded as they fell outside the research scope, resulting in a total of 76 papers included in the survey.

Concept overview

Data is becoming a pillar in every organization with the objective of being at the vanguard of technological advancements. From medical applications to smart cities, from industry to agriculture, all are progressing towards a point where data is becoming increasingly crucial for their objectives and goals (Qi, 2020).

When well-managed, data enables improvements in safety, productivity, efficiency, and innovation, organizations that leverage the power of data analytics gain a competitive edge, allowing them to adapt to changing circumstances, make informed decisions, and achieve their goals more efficiently. However, given the rapid expansion of this topic with a multitude of ideas and concepts emerging, it is crucial to take a step back and analyze the core concepts involved in transforming raw data into useful information. This capability is vital for driving organizations and society forward (Brous et al., 2020; Malik et al., 2021).

The topic of data management is vast, covering areas from data privacy and ethics to integration, storage, quality standards, and beyond. This paper also focus on big data, acknowledging the ever-increasing volume of data collected, which poses a significant challenge to current infrastructure. Building upon this foundation, the following concepts are explored: extraction, storage, preprocessing, and processing of this data. This overview of concepts not only references the work found in the review but also incorporates insights from complementary papers discovered during the research.

In the concluding section of the chapter, architectures and frameworks aimed at harmonizing all these ideas are presented. These architectures primarily focus on the concepts of data warehouse, data lake, and data lakehouse. These

Table 1 Number of papers by search terms and database

Search terms	Web of science	SCOPUS	IEEE Xplore	
"Data" & "management" & "manufacturing"	2261	4624	697	
"Data" & "pipeline" & "manufacturing"	114	275	84	
"Data" & "architecture" & "manufacturing"	1003	1995	583	
Total	3378	6894	1354	11,626

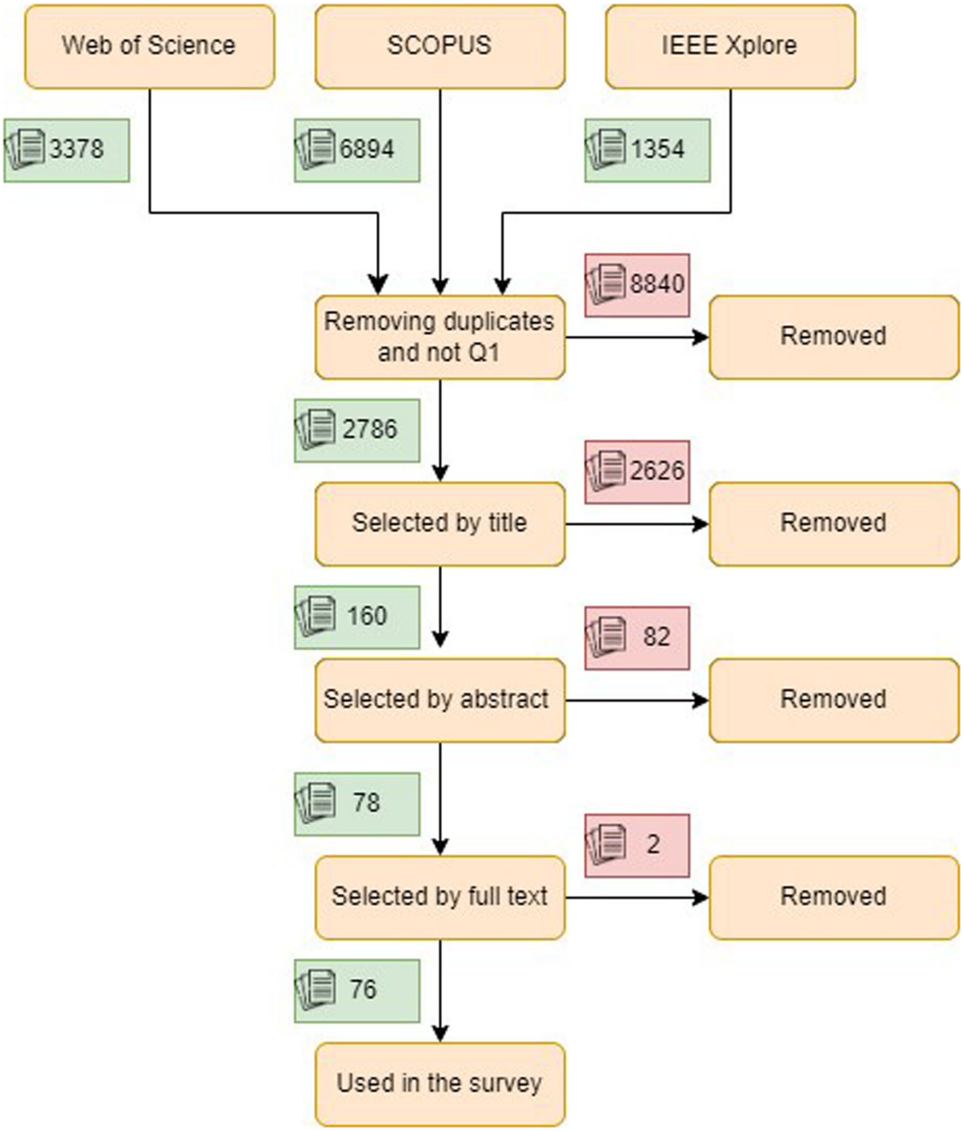

Fig. 1 Papers accepted and rejected in each stage of selection

pivotal architectures in data management are gaining increasing usage and definition, particularly within the industry. It becomes crucial to establish clear definitions for the various concepts entailed in these architectures as they play an increasingly important role in shaping the landscape of data management.

Finally, the objective of the concept overview is also to present the authors' vision of each concept. With many concepts having diverse meanings, visions, and understandings, the overview serves as a foundation for the subsequent survey and for defining present and future definitions of each concept. It is important to note that the authors do not impose a specific vision found in existing literature, but rather aim for alignment between the reader's understanding and the authors' perspective. This alignment ensures clarity in conveying the intended message, particularly when referring to complex and interrelated concepts.

Data management

Data management involves handling data from its collection to its processing. In (Raptis et al., 2019), the authors categorize it into three main groups: data presence, data coordination, and data computation. These groups primarily focus on data collection or gathering, preprocessing or cleaning (such as removing measurement errors or annotating information), storage (determining where data should be stored for optimal use), and data processing or analysis (compiling data, drawing relations or knowledge from it, and integrating it with other tools or representations) (Qi, 2020).

It's essential to recognize that data management is a broad field, and while this paper primarily addresses collection/extraction, preprocessing, storage, and processing/analysis, it extends beyond these aspects. From privacy, security, and ethics to integration, quality, and more, data management encompasses various terminologies that constitute the data landscape (Raptis et al., 2019). Furthermore, there are numerous frameworks focused on data management, each offering unique approaches and perspectives. For example, Huacarpuma et al. (2017) presents a framework for Industrial Internet of Things (IIoT) utilizing distributed data services to receive data from multiple sources and create metadata modules. Another example is (Saqlain & Shim et al., 2019) where a framework divided into five layers (physical, network, middleware, database, and application layer) encompasses not only data management but also resource management and event management.

Data management serves as the fundamental principle for overseeing data from its collection phase through to information processing. Achieving this requires the integration of various software components that facilitate the entire journey from data gathering to analysis. This intricate interconnection of processes is referred to as a data pipeline (Munappy et al., 2020). Data pipelines exhibit diverse architectures and functionalities tailored to specific use cases' requirements. However, the underlying principle governing all pipelines remains consistent: they comprise multiple modules where the output of one module serves as the input for another, thereby advancing through successive processes until reaching the final destination (Munappy et al., 2020).

The data pipeline represents a delicate balance, as its modules must process information at the necessary speed to ensure the overall functionality of the process, particularly when handling large volumes of data, commonly referred to as big data (Sebei et al., 2018).

Big data

One of the central concepts requiring thorough definition is the concept of Big Data. In simple terms, Big Data refers to datasets that are so extensive or intricate that traditional methods of data extraction, processing, storage, and application are inadequate to handle them. On a broader scale, the original definition of Big Data comprised three Vs (Volume, Velocity, and Variety) representing the most notable and challenging aspects of the concept (Majeed et al., 2021; Wang et al., 2022a). However, these characteristics proved insufficient in describing Big Data comprehensively, leading authors to introduce additional Vs for a more precise and robust description (González García & Álvarez-Fernández, 2022).

The first significant expansion beyond the initial 3Vs was the introduction of the 5Vs, which incorporated the characteristics of Veracity and Value into the definition of Big Data (Cui et al., 2020; Hajjaji et al., 2021; Tewari & Dwivedi, 2019; Yu et al., 2020). Despite a consensus around these key characteristics, some authors further elaborate on these points, introducing terms like Volatility and Variability (Belhadi et al., 2019), or even Visualization (González García & Álvarez-Fernández, 2022). Notably, nearly all referenced authors emphasize the importance of cybersecurity or data security when discussing Big Data. This concern is heightened when processing or storing data involves sending information to the Cloud, as it poses a significant increase in security risks (Guo et al., 2023) or the use of open software, some of which may have insufficient security mechanisms as referenced in Hajjaji et al. (2021).

Another critical aspect to mention is that many authors highlight the importance of visualizing Big Data. Finding ways to display key characteristics and relationships clearly and automatically between different metrics can be increasingly challenging. Failing to achieve this can compromise the usefulness of the entire system, especially when the end goal of these data is people, such as in the marketing or sales department.

Delving deeper into the key characteristics of Big Data, a detailed definition of each of the 7 Vs is presented, corresponding to the broader vision of all the authors mentioned previously (González García & Álvarez-Fernández, 2022):

- Volume: This is the initial focus of the concept, given the name "BIG data." It refers to the vast volume of data collected and generated by the system. The volume of data has been exponentially growing since the digital transition (Javed et al., 2018), posing challenges regarding all aspects of data management, such as storage, processing, and analysis.
- Velocity: This characteristic represents how quickly data is generated and needs to be processed or analyzed. With the

growth and increasing necessity and popularity of real-time systems, this exacerbates the speed at which data needs to be dealt with.
- Variety: Refers to the innumerable types of formats that data can take. This can include structured data from relational databases, semi-structured data in formats such as JSON or XML, or unstructured data, including videos or text documents. Platforms for processing or analyzing Big Data need to handle these distinct forms of information efficiently.
- Veracity: This characteristic reflects the reliability, accuracy, and trust in the data. Data can come from various sources, some prone to error, noise, inconsistencies, or biased readings. To ensure data veracity, mechanisms such as data cleansing, quality assurance, and validation processes are crucial.
- Variability: Refers to the volatility and inconsistency of data in terms of structure, format, and arrival rate. Data can be unpredictable and irregular, making it challenging to analyze and process. Big Data systems require flexibility and adaptiveness in data processing and analysis to deal with these changes effectively.
- Visualization: The process of presenting Big Data and its analysis in a meaningful and understandable format. With large and complex data, visualization helps humans comprehend and interpret information more effectively, aiding in decision-making.
- Value: Signifies the goal of the Big Data process—to extract value from data in the form of insights, patterns, and correlations. This characteristic is achieved by using and mastering the other six characteristics, completing the seven Vs referenced by the aforementioned authors.

It's important to note that there is are several definitions on this topic, and these are some of the most agreed-upon or widely used definitions by various authors, as can be seen in González García and Álvarez-Fernández (2022). Despite the concept being relatively old, with some claiming it dates back to 1997 (Saggi & Jain, 2018), it is increasingly explored and idealized, resulting in newfound characteristics that still complement the concept of Big Data. Consequently, the concept is not entirely defined or agreed upon; nevertheless, there is a good definition and an overall accepted view of what Big Data is.

On a related note, a notable manifestation of the surge in big data is the concept of Dark Data. Dark Data refers to data that remains hidden or unused for various reasons, ranging from difficulties in accessing it due to a lack of labels or tools, to high costs, or simply because it is concealed and challenging to discern. Nevertheless, the information embedded in this type of data is the most prevalent and constitutes the largest volume within a big data system (Corallo et al., 2023a).

Practical implementations, challenges, and their resolutions within the realm of big data are illustrated in Shukla et al. (2019); Raut et al., 2021). These sources offer a concise overview of opportunities in utilizing big data, providing brief yet informative descriptions of problems along with resolutions for various challenges encountered by industries. In (Raj et al., 2023) the primary challenges of employing big data in the manufacturing supply chain are highlighted, with the main hurdles being the product safety barrier, limited information sharing, and low managerial commitment. These studies contribute to a better comprehension of big data applications in diverse environments, serving as valuable references for initiating new implementations and constructing roadmaps for their development (Gökalp et al., 2021).

Data storage

In the realm of big data, driven primarily by high volumes, diverse types of data, and varied structures, a critical challenge revolves around the "where" and "how" of storing the massive volume of data. Concerning the "where," there are typically two options for data storage: locally or in a cloud provided by a private entity (Dai et al., 2020).

The cloud environment offers the convenience of infrastructure as a service, enabling the storage of large volumes of data without passing the concerns about space to the provider. The flexibility to scale up or down as needed is a significant advantage, making it adaptable even in cases of disruptions. Additionally, tools for data processing can seamlessly integrate into a cloud environment. However, utilizing infrastructure as a service presents several challenges (Qi & Tao, 2019; Syed et al., 2020), including:

- Overfull bandwidth: Transmitting all data to the cloud requires high bandwidth, incurring increased monetary costs.
- Unavailability: Disruptions in the network can render the use of cloud services impossible, affecting accessibility.
- Latency: In real-time or concurrent scenarios, the time it takes for data to travel to the cloud becomes critical.
- Data validity: Storing a large amount of insignificant raw data in the cloud may occupy significant portions of the database.
- Security and privacy: With the rising number of cyberattacks, ensuring data protection becomes crucial, particularly for personal data.

Moreover, inefficient interactions can occur, especially when communication between manufacturers, users, and machines lacks direct means to connect with a cloud environment.

On the other hand, the local environment presents fewer challenges than the cloud in terms of data transmission. Issues such as bandwidth, latency, security, privacy, and interactions can be mitigated as the storage is directly connected to the internal network of the company and disconnected from the internet. However, implementing data storage locally introduces its own set of problems. For instance, being locally bound requires physical presence on-site to work on the data. Additionally, all hardware costs are directly borne by the user, eliminating the flexibility in memory allocation seen in the cloud environment. The investment must always consider the storage capacity of the local environment. As such, a clear universal solution is not existent, each problem derives a set of constraints and requisites that should be met, allowing the utilization of the most well-suited approach.

Concerning the "how," various methodologies can be explored for storing information (Dai et al., 2020). The article will delve into three methodologies: data warehouse, data lake, and data lakehouse (Mazumdar et al., 2023).

Although not fully within the article's scope, it's crucial to touch upon data modulation and the significance of robust systems capable of classifying and adapting based on information. An intriguing example that adheres to the RAMI 4.0 principles can be found in Nagorny et al. (2020). The paper delves into the modulation and classification of various components in car manufacturing. This study can be complemented by the work of Neubauer et al. (2023), where similar concepts like asset administration shells are explored and expanded upon in a comprehensive scenario.

Data extraction

Data extraction is a critical component responsible for collecting data from various sources, with sensors being the most common source. Ideally, this component should not be limited by any specific data type or format, allowing it to collect any relevant data. The concept of data extraction is of utmost importance, as it involves retrieving data after it has been sensed, enabling the system to analyze the information and take appropriate actions.

The rise of sensor deployment in industrial complexes, not only on the shop floor but also in other areas, is evident, driven primarily by the development of the Internet of Things (IoT). It is estimated that around 14 billion devices were present in 2022 ("Number of Internet of Things (IoT) connected devices worldwide" 2023; "Number of connected IoT devices growing 16% to 16.7% billion globally", 2023) and this number is expected to double by 2030. This growth is also reflected in the increased publication of papers on the term IoT (Dachyar et al., 2019). Therefore, having the infrastructure and methodologies to effectively harvest this data is crucial to prevent system overflow and ensure scalability with the exponential increase in devices.

A common observation in methodologies related to data extraction is the recurring and overall design found in state-of-the-art architectures such as (Cecchinel et al., 2014; Kuzlu et al., 2022; Mocnej et al., 2016; Mussina et al., 2021; Trunzer et al., 2019; Zhang et al., 2022). These architectures predominantly focus on the use of middleware, where all sensors or, more broadly, data sources can connect and transmit their data, either locally or through a cloud platform. As depicted in Fig. 2, these architectures share a similar idea. Given the current technological landscape, this architecture can readily meet the requisites. However, it also suggests that this approach is indeed one of the most useful methods for connecting data sources, particularly from the perspective of data extraction.

Several studies conducted by other authors have delved into the technologies employed at the data extraction or ingestion points of the system. These investigations reveal that the predominant technologies include ingestion platforms, often within cloud environments; message brokers utilized both locally and in the cloud; and ad hoc software, customized solutions that do not rely on specific software (Ismail et al., 2019a; Mussina et al., 2021; Sahal et al., 2020). An insightful article (Ismail et al., 2019a) provides a comprehensive survey of various aspects of data pipelines in the industry, focusing on the ingestion part, as it becomes apparent that the prevalent use of ad hoc applications is driven by the unique characteristics of each industry scenario, encompassing protocols, communication types, and standards. However, there is a need to find solutions that can minimize the reliance on ad hoc systems, thereby streamlining the communication process. Open Platform Communications Unified Architecture (OPC-UA) communication emerges as a fundamental consideration, given its increasing adoption as a typical industrial practice (Pivoto et al., 2021). Consequently, custom applications or servers are often constructed to accommodate this protocol. Recognized as a valuable protocol seamlessly integrated into Programmable Logic Controllers (PLCs), OPC-UA is even recommended and exemplified by RAMI 4.0 (Adolphs & Epple, 2015). This positions the protocol as a robust solution for data extraction, particularly in scenarios involving PLCs.

In instances where ad hoc systems are not employed, a preference for ingestion and message broker software prevails. Technologies like Apache Kafka, RabbitMQ, Amazon Kinesis, Microsoft Event Hubs, and Google Pub/Sub are highlighted (Lu & Xu, 2019; Sahal et al., 2020). Beyond the specific technologies presented in these articles, it is crucial to look deeper and understand the underlying characteristics, often involving ingestion or message brokers/streaming capabilities for event-driven reasoning. These solutions typically possess near real-time capabilities, with some software deployable both locally and in the cloud, while others are exclusive cloud services.

The extraction process should adhere to various parameters that impact data collection. Fundamental concepts like data access control, availability of physical devices, and data

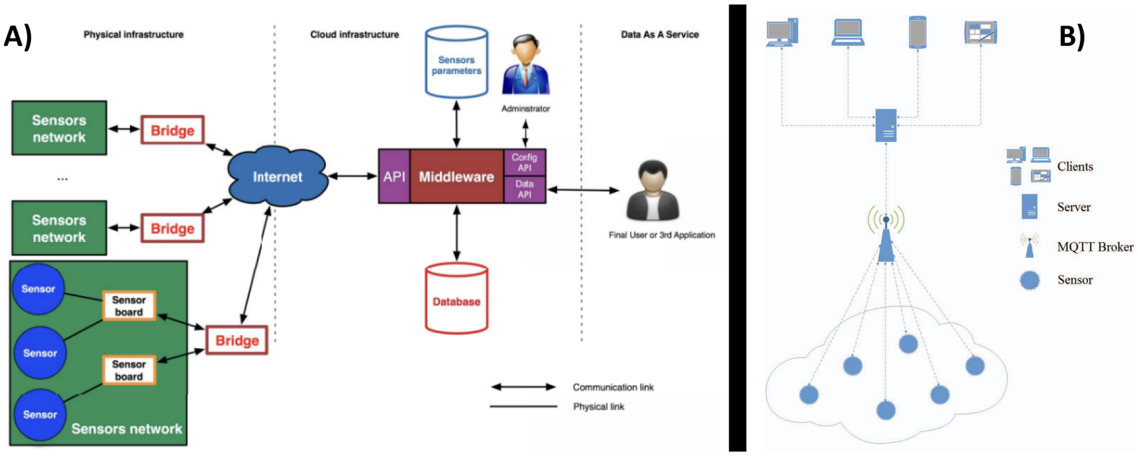

Fig. 2 Example of architecture with middleware. **A** Taken from Cecchinel et al. (2014) and **B** Taken from Mocnej et al. (2016)

availability are crucial for organized and accurate data extraction. In this regard (Allian et al., 2021) proposes a hybrid approach utilizing a combination of the mentioned software solutions. This hybrid approach can be effective, acting as a centralized mechanism for data before transmission to cloud platforms. For instance, employing a message broker to collect data from a shop floor before transmitting it to a cloud platform for analysis represents a feasible and advantageous strategy in numerous cases.

It's crucial to acknowledge that designing such infrastructure is a complex process. The extraction and utilization of data vary, and each case may present specific nuances regarding challenges and their resolutions. There are several pitfalls that need consideration in the design and implementation of these systems. Questions like "How to collect the dataset? How much data do we need? How to address imbalanced data?" (Lee & Chien, 2022) are paramount. Contemplating these questions not only in the immediate future but also in the long term is essential, as the growth of industrial necessities may reveal the need to expand or alter the way data is collected, with scalability and process standardization becoming increasingly relevant with growth (Mitra & Munir, 2019).

Data preprocessing

Data is often collected from heterogeneous sources with the intention of gaining insights into processes or factory operations. However, raw data is extremely challenging to process and often contains information that complicates the extraction of knowledge (Kabugo et al., 2020). Therefore, various methodologies can be employed to prepare this data for information processing. Additionally, concepts such as ETL (Extract, Transform, Load) and ELT (Extract, Load, Transform) are explored more deeply as intermediate steps between final storage and data extraction.

Table 2 Comon operations of data preprocessing

Methodology	Description
Cleansing	This refers to the removal of data that is significantly different from the rest and that is prone to negatively affect the performance of the usage of the data. It can be removal of outliers, noise, among others
Normalization	Data can have different scales and the normalization of this data can be essential for a correct storage or processing of information, a common example is information from two sensors of distance are passed in meters and the other in centimeters
Imputation	Raw data can often have missing values or incomplete reads, it is important to carefully check each of the sensor reads to be able to remove or fill (if possible) the missing or incomplete values
Transformation	The transformation of data can be split in several subtopics as represented in Farooqui et al. (2020). However, the general concept is to add (simple information like data or ID part), aggregate (such as compress several messages in one) or rearrange the messages, to better suit the needs for storage or processing

Based on Para et al. (2019)

Comon operations on data pre-processing

In the data pre-processing environment, various operations are performed to clean and prepare data for orderly storage or further processing. Numerous methods exist for preprocessing raw data; however, despite the various methodologies, the principles remain the same. Therefore, some of the most commonly used and useful methods are presented in Table 2 (Para et al., 2019).

These techniques form the core of data pre-processing and enable the elimination or mitigation of errors associated with the raw extraction of information from a shop floor. It is important, however, to understand that not all data requires pre-processing, and doing so indiscriminately can result in the loss of valuable insights. A classic example of this scenario is the removal of outliers, which can serve as a valuable indicator for scheduling preventive maintenance. Therefore, it is crucial to critically examine the data before pre-processing to ensure that essential information is not discarded, as it may be pivotal to the ultimate objective of data processing.

Data can often undergo pre-processing either before or after being stored initially. These two methodologies, referred to as ETL and ELT, have intrinsic properties and concepts that are important to delve into.

Extract transform load/extract load transform

In data pre-processing, two crucial topics related to data transformation should be discussed: Extract Transform Load (ETL) and Extract Load Transform (ELT). Despite a simple change in the order of letters, this fundamentally alters how information is processed and stored.

ETL is a method that involves extracting data from the source, transforming it to comply with the data storage schema, and loading this transformed data into storage. This is a more established method that has existed since there was a need to store specific values or filter information from sources. However, ETL has characteristics that may not be suitable for all applications. One such characteristic is the requirement for an intermediary system, often called data staging, to transform all the data and make it compliant with the schema, as depicted in Fig. 3 (Singhal et al., 2022). This creates a system unsuitable for time critical applications, as the data staging unit can become a bottleneck. All processing tasks, such as arranging, separating, clearing duplicates, standardizing, interpreting, and checking the consistency of information sources, need to be performed by this unit, resulting in complex and time-consuming systems, especially when dealing with a large volume of information (IBM Cloud Education, 2023; Singhal et al., 2022). Another significant characteristic of ETL is that, due to the transformation before data storage, the stored data is specifically designed to meet the system's needs at the time of its design. If, in the future, the data discarded by the data staging unit is deemed useful, it is lost with no method of recovery (Haryono et al., 2020). Consequently, the ETL method can be considered useful for synchronizing data from multiple sources, providing a unified and coherent final data storage solution, and addressing the need to migrate or update data from legacy systems. Legacy systems often require data transformation to adapt to new structures and formats that can then be stored.

Consequently, the ETL method can be considered useful for synchronizing data from multiple sources, providing a unified and coherent final data storage solution, and addressing the need to migrate or update data from legacy systems. Legacy systems often require data transformation to adapt to new structures and formats that can then be stored (IBM Cloud Education, 2023).

Regarding the ELT method, this approach involves extracting data from the data sources and then loading it into the storage system. After loading, the data can be transformed, or in some instances, it can be directly transformed inside the storage system, as seen in Fig. 4 (Haryono et al., 2020).

This system, unlike ETL, is much more suitable for time critical applications as it avoids the bottleneck associated with the transformation of data (Haryono et al., 2020). Some systems leverage functions within the data storage systems to transform the data, while others use external programs to process and subsequently store the processed data in the data storage unit. The system can employ both approaches and, if possible, operate in parallel to optimize the transformation of data and reduce time consumption. However, implementing this system can be challenging in a relational database when extracting unstructured data. Most relational databases operate on a schema-on-write basis, meaning that there is already a data structure and organization before any data is ingested. Therefore, for unstructured data, the use of non-relational data storage units may be a more suitable approach for some applications (MongoDB & "Unstructured, 2023).

A table comparing ETL and ELT can be found in Table 3. This table is based on the study conducted in Singhal et al. (2022); Haryono et al., 2020).

Data processing

Collecting and storing data by itself has limited use. Ideally, data collected in a manufacturing process, a city, or a social network can yield insights that are not immediately clear. These data can be interconnected, revealing underlying information not easily discerned by a human (Tripathi et al., 2021; Zhang & Han, 2021). In this section, the most common methodologies for processing data, namely batch and streaming processing, will be explored. While divided into sections, the data processing methodology is not necessarily segregated in its implementation, and often a combination of both batch and stream processing can yield the best results. This is particularly evident in energy efficiency, quality inspection or predictive maintenance problems, where a model is frequently created using a batch processing approach. Subsequently, this model is employed in real-time analyses of shopfloor machines or process information. Extensive examples and implementations of these methodologies can be found in Gopalakrishnan et al. (2022).

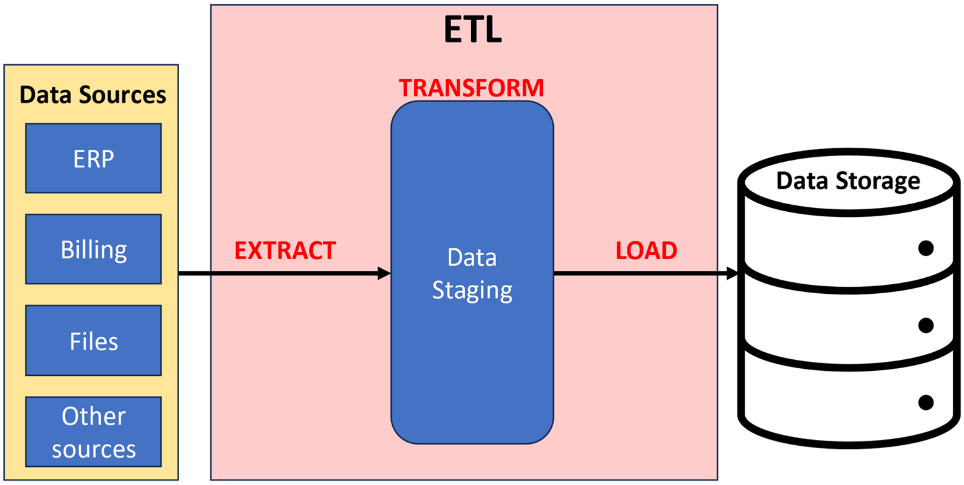

Fig. 3 ETL example architecture adapted from Haryono et al. (2020)

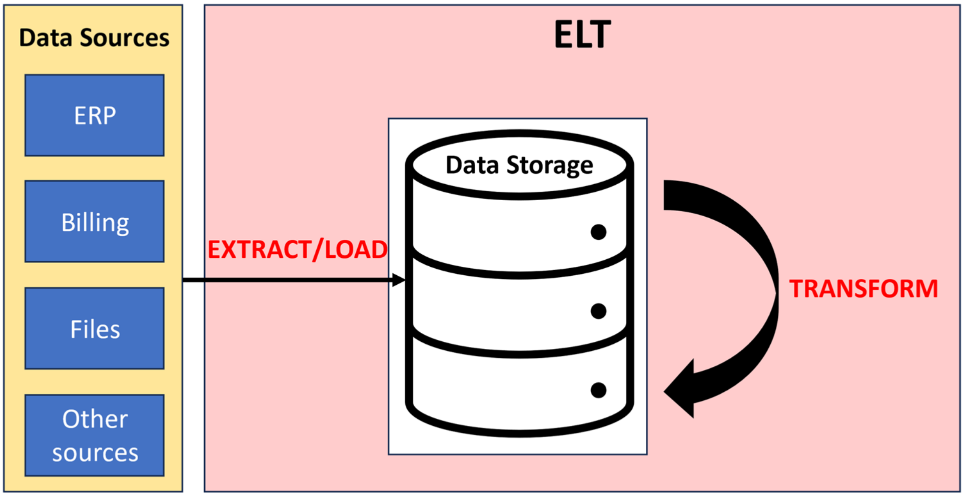

Fig. 4 ELT example architecture adapted from Haryono et al. (2020)

Batch processing

Batch processing is a method of analyzing and processing information or data that involves using a dataset, typically a large one, to extract relationships or insights. When this batch computing process is complete, the results are usually presented as a metric or a new dataset that highlights the insights found during processing (Pfandzelter & Bermbach, 2019). Thus, this approach has an intrinsic characteristic, requiring the full dataset to be present before the batch processing method starts working (Pfandzelter & Bermbach, 2019).

In its elemental form, batch processing is a straightforward process. A large dataset with multiple pieces of information that are equally structured is presented to an algorithm that produces an output aligned with the dataset. This can range from discovering the best trajectory around a town (Zhang & Han, 2021) to dynamic scheduling (Fowler & Mönch, 2022).

However, this system has some inherent flaws. The first and most obvious one is the quality of the data. As this method relies on a large amount of accurate data, ensuring data quality becomes increasingly challenging when dealing with large volumes of data. Several methods exist to try to minimize this risk, and novel methods are continuously researched to address this point. For example, the work in Peng and ChunHao (2022) uses a fuzzy learning system to check the quality of input data in multi-domain batch processing. Another important aspect to keep in mind is that this system is not suitable for real-time operations or events, as it relies on existing data and is optimized for large datasets. Additionally, batch processing is not flexible enough to deal with events in a timely manner (Pfandzelter & Bermbach, 2019).

Figure 5 depicts the typical architecture of a batch processing system. This specific architecture is from the Microsoft Azure data architecture for batch processing (Tejada, 2024a), proven to work and satisfy the requirements of numerous users. In this architecture, data from the sources is first stored in a data storage, which does not need to have any organization and can even be non-SQL storage. The only restriction is that it needs to follow the batch processing unit data format

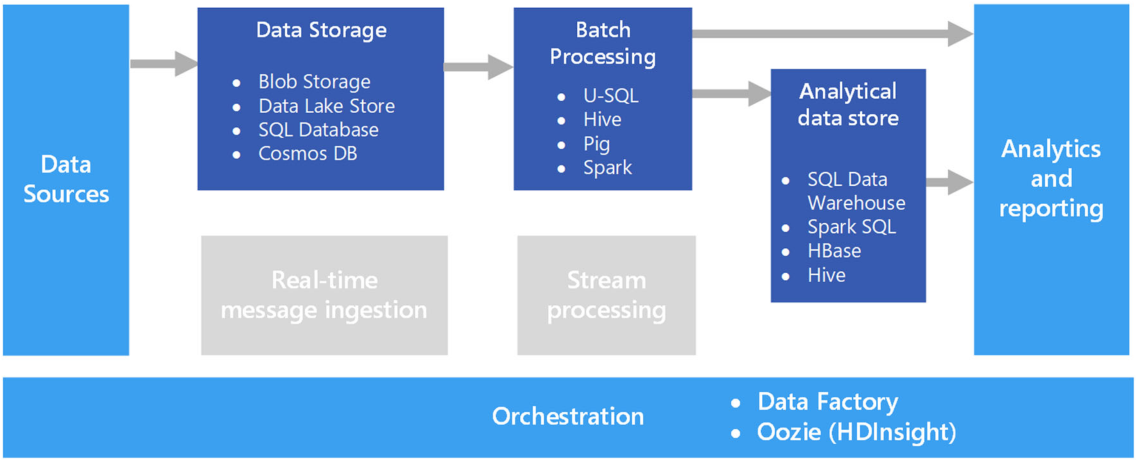

Fig. 5 Batch processing architecture of Azure, from Tejada (2024a)

and encoding, or vice versa. Another challenge is when critical information arrives with a delay; for example, if data from yesterday only arrives today, the units (storage or batch) must be capable of processing this data; otherwise, the information will be discarded (Tejada, 2024a). The batch processing unit can use various technologies or techniques (some represented in the image), but it always has the same objective: to analyze and extract non-obvious information or knowledge from the datasets fed into the unit. After batch processing is completed, the data can be stored directly in another storage center, fed to other systems or processing units, or finalized into a report. Finally, the orchestration unit is responsible for copying the data into the data storage, initializing batch processing, and copying and migrating the data into the analytical data store and reporting layers.

Stream processing

Despite the usefulness of batch processing, as seen previously, for some interactions, this is simply not the most suitable method of analyzing information. A classic example is fraud detection in financial transactions or in medical applications, where the latency of batch processing is enough to make it unsuitable for responding to this kind of analytical problem (Isah et al., 2019). Data streaming processing offers the advantage of ideally processing the information in real-time or near real-time and should be used where the information contained in the collected data degrades rapidly with the passing of time (Wampler, 2016).

As with all methodologies, this one also has limitations. To achieve low latency, stream processing must perform the processing without having resource-intensive computational operations, maintaining real-time or near real-time processing (Isah et al., 2019). This is an important point, clearly distinguishing between batch and stream processing. For simple and quick data processing that doesn't require more data than is currently present in the pipeline, stream processing is the ideal solution. Another challenge that is important to consider is the fact that the data is coming directly from the data sources, meaning that no or little preprocessing is done in the data. As such, the data stream processing should be able to deal with imperfect data, such as delayed, out-of-order, or missing values (Isah et al., 2019).

Figure 6 depicts the typical architecture of a stream processing system. This specific architecture is from the Microsoft Azure data architecture for stream processing, and as such, we can conclude that this architecture is a useful and functional one, as Microsoft Azure has numerous clients that utilize their platform with positive results. In this architecture, it is possible to see that the data coming from the data sources can be stored in a data storage component or be directly fed into an ingestion unit. This ingestion unit is a method of capturing and storing real-time messages to be consumed by the stream processing unit. This ingestion unit can be simply a data storage, but often it is done using a message broker (Tejada, 2024b). The next unit is stream processing; this unit has several data processing functions that can be configured and work in series or parallel to create a flow of data that analyzes, extracts, or compiles results from the incoming data streams (Gualtieri & Yuhanna, 2016). The stream processing units can use a variety of technologies and techniques, often a combination of several is used with the objectives of optimization and reliability. After the stream processing unit, the data can be stored directly in another storage center, fed to other systems or processing units, or finalized into a report.

Architectures

With the concepts previously described, in this chapter, several architectures will be presented, starting with the classic and usually used data warehouse (Chandra & Gupta, 2018),

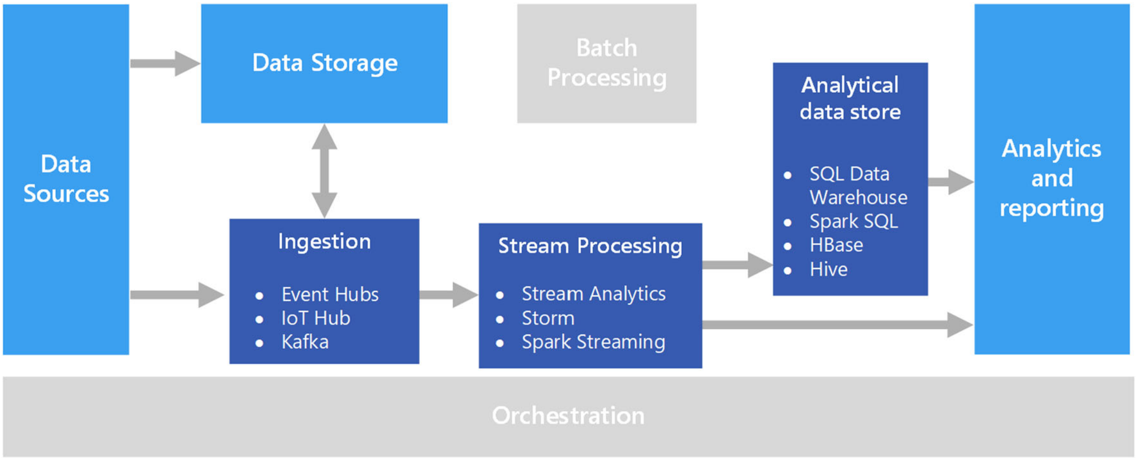

Fig. 6 Stream processing architecture of Azure, from Tejada (2024b)

then the ever-growing in popularity data lake (Hlupic et al., 2022), and finally, the novel junction of the two of them, the data lakehouse (Orescanin et al., 2021). These implementations often have the end goal of analyzing big data; however, it is important to refer to several factors on which the end goal is highly dependent, one of such factors is the quality of the data and data acquisition barriers, being the most important point presented in Kumar et al. (2021) and Moktadir et al. (2019). However, the confidence in the processes in upper management is often one of the most pointed cases according to Raut et al. (2021) and Kumar et al. (2021). This compels the utilization of clear architecture and proven methods to demonstrate the usefulness of the architecture, such as the ones presented in this chapter.

One important notion is that the presented architecture can be complemented with several studies that already exist. In Liu et al. (2023a) and Zhang et al. (2020a), there is a complex framework that approaches several aspects discussed in this paper, as it presents several important notions in each component from extraction, storage, and processing, giving a good complement to the architecture presented in the chapter, as well as a more guideline framework approach to the implementation of such. In Wei et al. (2021) several technologies used to implement the architectures are also referred to, comparing the use of data type and function that the system is meant to develop with the corresponding technology. The study also provides several examples of utilization from smart cities to manufacturing.

Data warehouse

Data warehouse as a concept was introduced around the 1980s by IBM with the objective of developing an architecture that would improve the flow of data in decision support systems. By definition, a data warehouse is a subject-oriented, integrated, non-volatile, and time-variant multitude of data with the objective of supporting system decisions (Chandra & Gupta, 2018; Nambiar & Mundra, 2022). According to Chandra and Gupta (2018), an architecture of the data warehouse should have several building blocks that make a complete system, from receiving to querying data and from storage to the management of such data. A brief overview of the main blocks will be present:

- Source of data: The sources of data may vary greatly, but they are usually incorporated into four main categories: production or operational data, internal data, external data, and archived data.
- Data staging: This step revolves around the preprocessing of the data, including steps such as cleaning, transforming, reducing, and integrating. Here the concept of ETL enters into effect, as in this stage extraction, transformation, and loading become the core objectives.
- Data storage: Data storage is a critical component, allowing the storage of the processed and transformed information to be accessed later. Here, relational databases or multidimensional databases are used to store data.
- Information delivery: Here are the components that provide information to the several recipients that require it. This information can range from simple queries to complex reports.
- Metadata: The metadata represents all the information regarding the description of every component of the data warehouse. It can also include information about end users, processes, and any other information deemed useful for the system or its users.
- Management and control: This block is responsible for the coordination of all the activities and the synchronization between all the other blocks. Also responsible for the synchronization of information between metadata and the other components inside and outside the data warehouse.

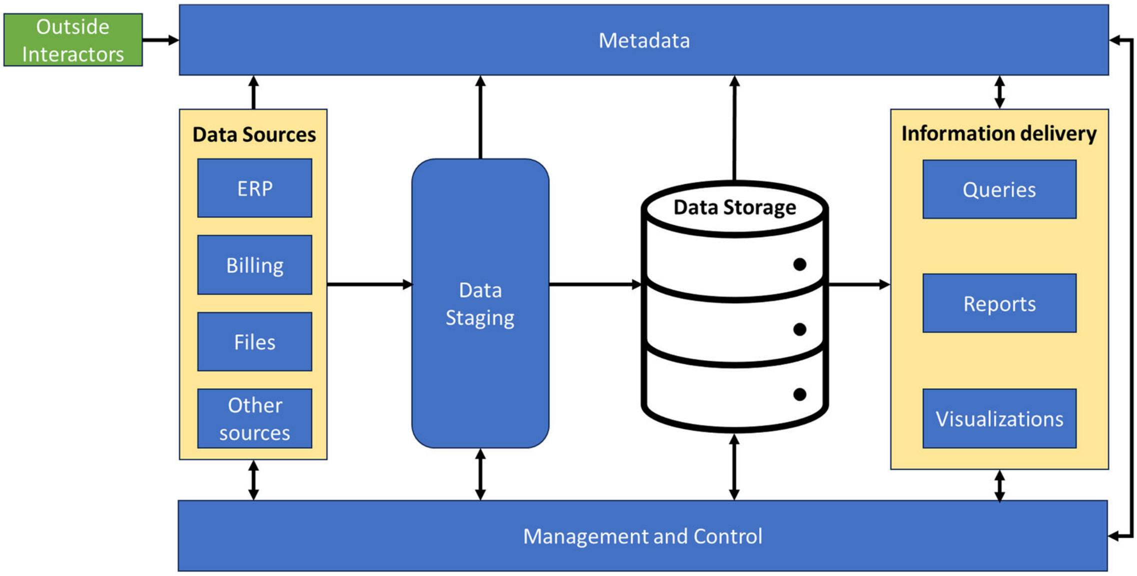

Fig. 7 Data warehouse architecture based on Chandra and Gupta (2018) and Nambiar and Mundra (2022)

Overall, a data warehouse architecture is more than simply a storage system; it is a complex system that provides the reception of data or elements of ETL, allowing the extraction of said data, their transformation, and their storage in a relational or multidimensional database. Beyond this fact, the system does not end here, as it allows the extraction of the storage data. The data warehouse contains a complete description of the metadata of its components and, ideally, of all the entities that interact with the system. Finally, it incorporates management and control as one of its core components, as it is responsible for coordinating all the other components and synchronizing them not only with the contained components but also with outside information when needed. An architectural image of a data warehouse can be found in Fig. 7.

Data lake

With the ever-growing amount of information provided by the integration of different devices, it becomes critical that systems are capable of dealing with this volume of information, also known as big data. Data warehouses have fundamental problems that make them difficult to use in conjunction with big data. The most common issue is the impossibility of storing multimedia-type data, which is often unstructured, as referenced in 3.2. Big data. The other critical problem is that to accommodate this data in data warehouses, ETL is usually used to transform unstructured or extract data in a standardized way from multimedia types of data. However, this process is bound to lose important information and is increasingly complex (Singh et al., 2022).

To combat the problems imposed on data warehouses by big data, the data lake concept was developed. The data lake concept has several key characteristics that allow for increased flexibility in handling data. These characteristics are as follows (Russom, 2017; Sawadogo & Darmont, 2021; Singh et al., 2022):

- **Storage at a Lower Cost:** Data lakes offer more cost-effective storage, including memory and processing power, compared to data warehouses.
- **Multimedia Data Storage without Pre-processing:** Data lakes can store multimedia types of data without the necessity of pre-processing.
- **On-the-Fly Data Transformation:** A data lake should be able to transform and prepare data on the fly, not bound by only ETL transformations but having a broader system capable of data exploration and discovery-oriented data analytics.
- **Schema Flexibility:** The data lake architecture is not bound by a traditional rigid structure of data, allowing it to store different types of data in the same environment without specific rules.

An important point regarding a data lake is to avoid it turning into a data swamp. Data swamps are undocumented and disorganized data storage, difficult to navigate or use. This problem can be managed with good data management practices or automated by a governance zone in the data lake architecture. However, it is crucial to have an understanding of this problem as it can render the data lake useless if not managed from the beginning (Russom, 2017).

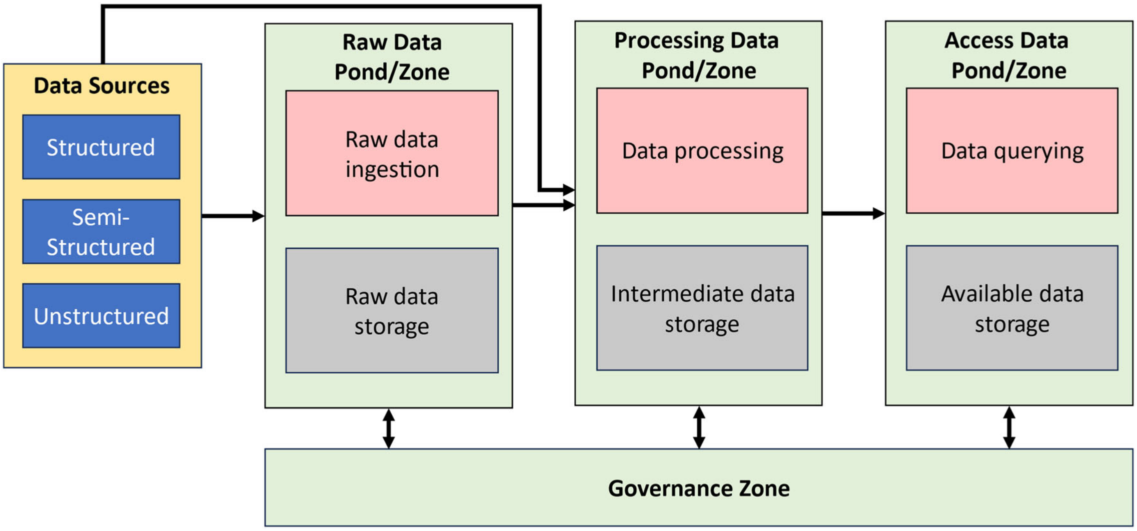

Fig. 8 Data lake architecture. Based on the work of Hlupic et al. (2022) and Sawadogo and Darmont (2021)

Finally, a general architecture of a data lake is presented in Fig. 8. This architecture subdivides the data lake into smaller data ponds, each responsible for specific storage and execution of tasks related to that specific data. In the case of raw data, this pond is responsible for storing and ingesting raw information from heterogeneous sources. The processing data pond is responsible for the storage of pre-processed data, for example, data from streaming or batch processing, and has methods for processing data, either from the other ponds or directly from outside sources. The data access pond is responsible for storing already processed data or conclusions drawn from it, with methods to facilitate viewing or gathering data from the system. Transversal to all the ponds is the governance zone, responsible for metadata, allowing monitoring, managing, and governing several aspects of the ponds and the data itself. This zone is also responsible for data quality, catalog, and security (Russom, 2017; Sawadogo & Darmont, 2021; Singh et al., 2022). One final remark is that when the data persists after being transferred to the next pond, the ponds are often called zones (Sawadogo & Darmont, 2021).

Data lakehouse

Despite the previous architecture, sometimes it is useful to have the best of both worlds, either by upgrading from an existing data warehouse to incorporate some data lake properties without sacrificing previous functionalities, or by introducing some useful structure or organization in a complex data lake (Harby & Zulkernine, 2022).

The data lakehouse architecture aims to harness the best features of both the data lake and the data warehouse. It strives to offer not only the flexibility, data integration, discovery, and management of the data lake architecture but also the structured, clean, and integrated storage of a data warehouse (Harby & Zulkernine, 2022). Therefore, a data lakehouse can be described as low-cost, providing easy and direct access storage, with elements of data analytics and processing, as well as traditional database management system analytical and performance features (Armbrust et al., 2021). A comparative table highlighting the characteristics of the three architectures can be viewed in Table 4, facilitating a comprehensive comparison between them and what they aim to accomplish.

It is evident that the data lakehouse aims to achieve a balance by incorporating the best characteristics of both architectures. This approach not only accommodates the preferences of data scientists, who often favor data lake systems, but also meets the requirements of business-oriented management, which typically leans towards data warehouse systems. The integration of these features creates an environment that fosters seamless collaboration between data scientists and business-oriented management, facilitating the exchange of information, discoveries, and analyses.

To delve into the architecture of a data lakehouse system, it's crucial to recall the concepts discussed in previous chapters, as the presented information and architecture will build upon them. In Fig. 9, a general architecture of a data lakehouse is depicted. The journey begins with information from data sources being delivered to a landing zone. This landing zone, which can function as a data pond, then disseminates the information to both the data warehouse and the data lake. This dual delivery enables structured or business-oriented information to be sent to the data warehouse for processing and availability, while simultaneously allowing the data lake to store all non-structured information, ready for

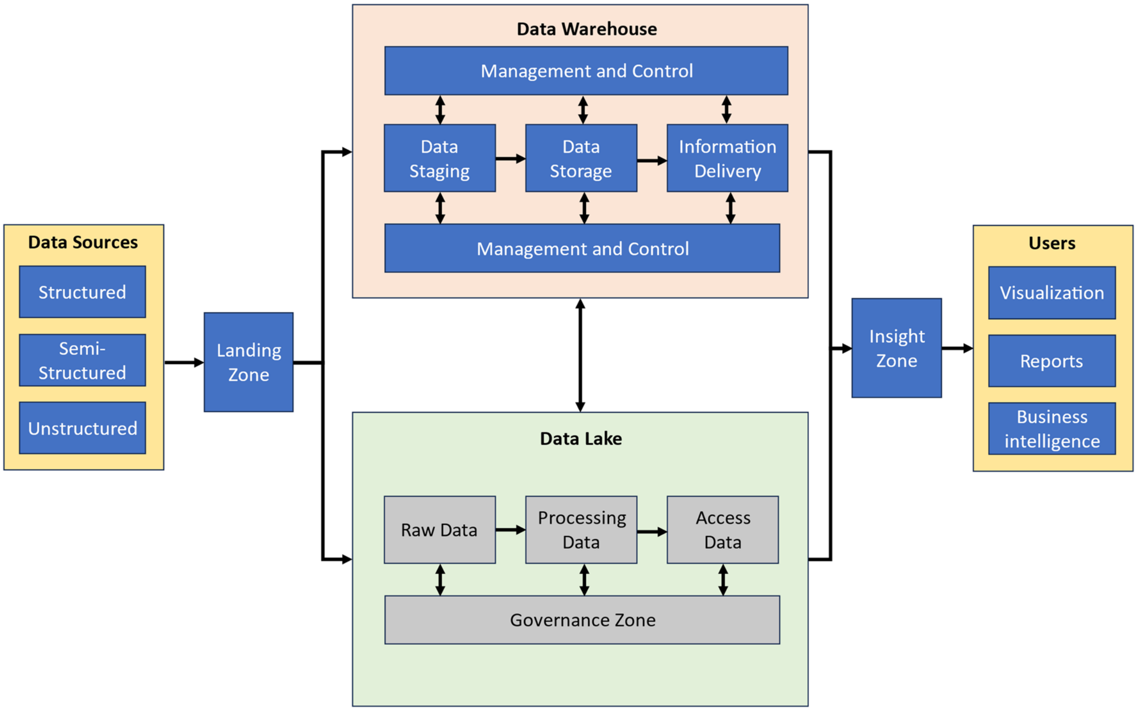

Fig. 9 Data lakehouse architecture based on the work of the previous architectures and Mazumdar et al. (2023) and Orescanin et al. (2021)

archiving or analysis at a later time. This approach involves a compromise, as data storage may not be as efficient as a dedicated data lake or as fixed in nature as a data warehouse (Orescanin et al., 2021). However, it ensures the preservation of a clean and structured portion of information critical to certain processes, a characteristic of the data warehouse. This approach avoids discarding any valuable information by capturing it in the staging zone of the data warehouse as well as having improved query times for data stored in the data warehouse in comparison to data stored in the data lake (Orescanin et al., 2021). Towards the end of the architecture, there is an insight zone responsible for communicating with users and providing necessary information tailored to various user needs, whether it be for a machine, an analyst, a business report, or any other requirement.

Notions and assumptions

Until now, the principal topics and sub-topics related to data management, along with comprehensive examination of its various aspects, has been presented, incorporating survey findings on the subject. The discussion encompasses key and well-defined concepts such as big data, data storage, and data processing. References and perspectives from different papers, as well as the viewpoints of the paper's authors, contribute to the exploration of these concepts. It is crucial to note that certain concepts admit multiple definitions, rendering it impractical to adopt non-complementary definitions for the same theme. While the authors assert that none of the identified definitions are inherently incorrect by their point of view, they acknowledge the diverse perspectives offered by different individuals, which may impact the description of specific technologies or architectures. The objective of this chapter is to convey the authors' viewpoint and the rationale behind their adherence to a particular classification. It is emphasized that this choice does not invalidate other perspectives but aligns more cohesively with the authors' understanding of the subject of data management.

Two crucial concepts warrant emphasis, with the first focusing on the definition of IoT. IoT can be defined in various aspects; for instance, it may represent a device capable of directly transmitting data to the internet, such as a smart sensor (Ning, 2013). However, some devices only transmit data to the specific manufacturer's application, despite technically being on the internet, with limitations on data transmission destinations. Other devices come equipped with built-in mechanisms to transmit Message Queuing Telemetry Transport (MQTT) messages, potentially to a server, but once again constrained by the transmitted protocol. Additionally, certain devices require an intermediate gateway, enabling data transmission to any internet service (Sorri et al., 2022). In the authors' perspective, any of these definitions is valid, depending on the circumstances. The chosen definition aligns with (Wei et al., 2021), which posits that IoT could be perceived as a combination of multiple devices, including other IoT

Table 3 Comparison between ETL and ELT

Parameters	ETL	ELT
Optimal Use	Ideally for structured or semi-structured data, legacy systems and to load or extract from relational DBs	Faster data loads for both structured and unstructured data ideally for large dataset. The transformation can be done as needed
Privacy	Personal Identifiable Information can be eliminated in the Preload transformation step	Major safeguards for privacy are required since data is directly loaded
Transformations	Secondary servers are needed to perform the transformations. Usually resource intensive operations are needed	Higher speed and efficiency are achieved since the database performs load and transform simultaneously. The transformation operations can also be done using in built functions of the data storage unit and in parallel with other systems
Maintenance	High maintenance due to the presence of multiple processing servers. Complexity of the systems can increase exponential with the complexity of data	Reduced maintenance burden because of fewer systems. Simpler as all the data is loaded into the DB
Expenses	Monetary issues due to separate servers	Less Monetary overhead because of simplified data stacks
Compatibility with Data Lake	Output is Structured	Output can be Structured, unstructured or semi-structured
Amount of Data	Ideal for datasets of small and moderate volume	Ideal for datasets of large volume

Based on Haryono et al. (2020) and Singhal et al. (2022)

Table 4 Comparison between data warehouse, data lake and data lakehouse

Features	Data warehouse	Data lake	Data lakehouse
Data	Structured, Processed	Structured, Semi-Structured, Unstructured, Raw, Processed	Structured, Semi-Structured, Unstructured, Raw, Processed
Processing	Schema-on-write	Schema-on-read	Schema-on-write, Schema-on-read
Storage	Expensive for large data volumes	Designed for low-cost storage	Designed for low-cost storage
Agility	Less agile, fixed configuration	Highly agile, adjustable configuration	Highly agile, adjustable configuration
Security	Mature	Maturing	Maturing

Based on the work presented in Armbrust et al. (2021) and Harby and Zulkernine (2022)

devices. Complex devices, as long as they exhibit IoT characteristics, qualify as IoT devices. With this perspective, all previously mentioned examples are considered IoT devices, as they consist of one or several components collaborating to provide services or data to the network.

The second noteworthy concept pertains to the definitions of data lakes/warehouses within the survey context. Some authors argue that data lakes/warehouses constitute a complex environment beyond being merely a storage service, as previously mentioned in chapter Architecture (Harby & Zulkernine, 2022). Nevertheless, several authors use this term loosely, defining it based on the type of storage or a specific technology, such as MongoDB (Sawadogo & Darmont, 2021). In line with this trend, the authors of this paper also adopt a loose usage of the term, particularly for the survey section. This is due to the inherent difficulty in distinguishing which parts are often interconnected or operate in the same context as data lakes/warehouses, given that full implementations are frequently inadequately detailed to comprehend the entire process. While the authors recognize the significance of other aspects in the data lake/warehouse architecture, as highlighted in the chapter Architectures, they consider it a fair tradeoff. This is because the manner in which information is stored remains one of the most crucial aspects in both data lake and data warehouse architectures.

Survey findings

The field of data management is expansive and intricate, encompassing crucial concepts that profoundly impact its

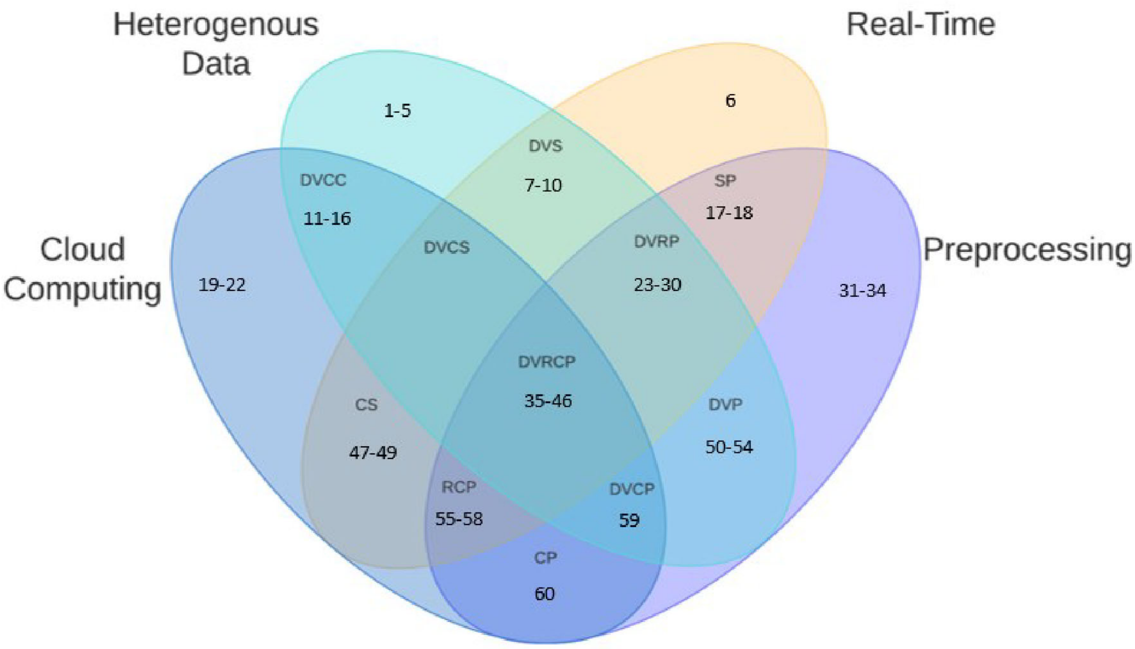

Fig. 10 Paper classification based on the concepts used

effectiveness. The interplay between various subjects and system characteristics often unveils the requirements and nuances essential for optimal system functionality. A classic illustration of this scenario includes the imperative to handle diverse types of data or the need to operate in a real-time or quasi real-time environment. In such settings, prompt decision-making becomes imperative to uphold product quality or ensure personal safety.

However, systems frequently grapple with a combination of challenges. There are instances where a system must concurrently handle various data types in real-time, bringing forth the interconnected nature of different facets within the data management system. According to the authors, it is crucial to acknowledge the diverse aspects of the data management system and comprehend their collaborative dynamics. This understanding provides valuable insights into the intricacies involved in working with data management systems.

Therefore, the review not only categorized papers into four major systems but also explored the intersections among these systems. According to the authors, these systems represent crucial aspects in dealing with data. While acknowledging that more concepts and nuances exist within this domain, the classification of these four systems and their intersections serves as a valuable initial indicator for drawing conclusions and provides a solid starting point for discussing collaboration and integration across various aspects.

In Fig. 10 a visual representation illustrates the classification of articles based on their utilization of the four concepts. Some articles may not comprehensively delve into all the systems due to factors such as article size, complexity, or specific purpose. However, when clear utilization of these concepts is evident, they have been incorporated into one of the intersections. It is essential to note that certain significant articles are not explicitly mentioned in Fig. 10. This omission arises from articles that present challenges, methodologies, or frameworks that may not distinctly emphasize any of the four concepts but are nonetheless highly relevant for establishing and applying data management systems. These articles have been duly incorporated into the chapter Concept overview of this paper and are here explicitly referenced (Allian et al., 2021; Corallo et al., 2023a; Gökalp et al., 2021; Gopalakrishnan et al., 2022; Ismail et al., 2019a; Kumar et al., 2021; Lee & Chien, 2022; Liu et al., 2023a; Mitra & Munir, 2019; Moktadir et al., 2019; Neubauer et al., 2023; Raj et al., 2023; Raut et al., 2021; Shukla et al., 2019; Wei et al., 2021; Zhang et al., 2020a).

In the subsequent sections of this chapter, a thorough exploration of these four concepts is undertaken. The focus is on delving deeply into the utilization and discussion of each concept, supported by relevant articles from the survey. This approach is aimed at enriching and exemplifying the discourse surrounding each subject.

Acronym description of Figure 10

Acronym	Name	Description	Reference	Number
-	Heterogenous Data	It represents a wide variety of data formats that can emerge from different processes that often requires effective strategies to ensure seamless integration and utilization	(Helu et al., 2020; Schmetz et al., 2020; Shin, 2021; Tufano, 2023; R. Wang et al., 2022c)	1-5
-	Real-Time	It represents the need for (quasi) real-time processing of data, where actions must be performed within strict time constraints. This is particularly crucial for system performance and decision-making	(Izagirre et al., 2022)	6
DVS	Data Variety Streaming	This intersection is the junction between the elements of heterogenous data and real-time. This represents a high variety of data that can be transported and processed in (quasi) real-time	(Chen, 2020; Göppert et al., 2023; Majeed et al., 2019; J. Zhang et al., 2023)	7-10
DVCC	Data Variety Cloud Computing	This intersection is the junction between the elements of heterogenous data and cloud computing. This intersection represents that a cloud environment is being used and can process heterogenous data	(Bonnard et al., 2021; Corradi et al., 2022; Jurmu et al., 2023; Raptis et al., 2019; L. Zhang et al., 2020b; Wang et al., 2020)	11-16
DVP	Data Variety Preprocessing	This intersection is the junction between the elements of heterogenous data and preprocessing. This intersection represents the system has the ability to apply preprocessing operations in heterogenous data	(Corallo et al., 2022; Deshmukh et al., 2021; Michalkowski et al., 2023; Park & Huh, 2023; Sorri et al., 2022)	50-54
SP	Streaming Preprocessing	This intersection is the junction between the elements of real-time and preprocessing. This intersection represents the capacity by the system to preprocess the data in (quasi) real-time, ideally in a streaming environment	(Ma et al., 2022; Villalobos et al., 2020)	17-18
-	Cloud Computing	It represents the use of cloud environments for processing collected data, whether for running complex analytical models or facilitating simple data visualization	(Guo et al., 2023; Kozjek et al., 2020; Luo et al., 2022; Y. Wu, 2021)	19-22
DVCS	Data Variety Cloud Streaming	This intersection is the junction between the elements of heterogenous data, real-time and cloud computing. This represents the ability to process heterogenous data in real-time in a cloud environment. Important to notice that in this intersection, preprocessing is not used		
DVRP	Data Variety Real-time Preprocessing	This intersection is the junction between the elements of heterogenous data, real-time and preprocessing. In this intersection heterogenous data is preprocessed in (quasi) real-time, ideally trough a utilization of a streaming environment	(Bi et al., 2023; Cerquitelli et al., 2021; Hinojosa-Palafox et al., 2021; Kahveci et al., 2022; Kammerer et al., 2020; Kim & Lee, 2021; Qiu et al., 2020; W. Wang et al., 2019)	23-30
-	Preprocessing	It represents the use of preprocessing techniques to clean and refine data for optimal utilization. This can include tasks such as adding relevant features, removing outliers, or handling missing values	(Farooqui et al., 2020; Horak et al., 2022; Para et al., 2019; P. Wang & Luo, 2021)	31-34

Acronym	Name	Description	Reference	Number
DVRCP	Data Variety Real-time Cloud Preprocessing	This intersection is the junction between the elements of heterogenous data, real-time, cloud computing and preprocessing. This intersection it utilizes all of the concepts, being a system capable of dealing with heterogenous data, in a (quasi) real-time, while utilizing preprocessing and cloud computing for the processing	(Alabadi et al., 2022; Fahmideh & Beydoun, 2019; Filz et al., 2023; Fortoul-Diaz et al., 2023; Kabugo et al., 2020; Majeed et al., 2021; Rajnoha & Hadac, 2022; Saqlain et al., 2019; Sarker et al., 2023; K. Wang et al., 2022b; H. Wu et al., 2022; Yang et al., 2020)	35-46
CS	Cloud Streaming	This intersection is the junction between the elements of cloud computing and real-time. In this intersection the system utilizes (quasi) real-time and cloud computing to process or store data, ideally in a streaming environment	(B A et al., 2020; Koprov et al., 2022; N. Zhang, 2021)	47-49
RCP	Real-time Cloud Preprocessing	This intersection is the junction between the elements of real-time, cloud computing and preprocessing. In this intersection the preprocessing is made in a cloud environment in a (quasi) real-time period, ideally in a stream environment. In this intersection there is not a variety of data	(Fang et al., 2020; Leang et al., 2019; Yu et al., 2020, 2022)	55-58
DVCP	Data Variety Cloud Preprocessing	This intersection is the junction between the elements of heterogenous data, cloud computing and preprocessing. This intersection represents the systems that deal with heterogenous data making the preprocessing of this data in a cloud environment. In this intersection the data is usually handled in batches, and not in a real-time environment	(Lu & Xu, 2019)	59
DVRCC	Data Variety Real-time Cloud Computing	This intersection is the junction between the elements of heterogenous data, real-time and cloud computing. This intersection is responsible for systems that use heterogenous data in a (quasi) real-time environment while utilizing cloud computing for processing or storage. All this environment does not use any kind of preprocessing		
CP	Cloud Preprocessing	This intersection is the junction between the elements of cloud computing and preprocessing. In this intersection the system uses the cloud computing to deal with the preprocessing of the data	(Liu et al., 2023b)	60

Heterogenous data

In a data management context, a multitude of devices and software regularly generate data that can be harnessed or transformed for application in optimization or problem detection. This scenario is exemplified in environments like a shopfloor, where data from machines and sensors aid in predicting defects or optimizing maintenance procedures. However, the application of data is not confined to specific domains and extends across various fields, encompassing tasks ranging from simple registration and storage to complex transactions, tracking, and safety protocols. As the utilization of data from diverse sources increases, the systems encounter a surge in data with different formats, standards, and types, commonly referred to as heterogeneous data.

The prevalence of heterogeneous data is noteworthy, constituting a significant portion with 40 out of the 76 papers classified falling within this category. This surge can be attributed to several factors. Firstly, the proliferation of IoT devices plays a pivotal role, offering increasing diversity and ubiquity, consequently this is bound to lead to a heterogeneity in the types of data produced by these devices. This diversity of devices enables the selection of the most suitable device for specific data collection challenges, albeit introducing variations in how data is conveyed. Another contributing factor is the ubiquitous sourcing of data—from middleware services and various platforms—introducing a plethora of formats into the system. This trend highlights the growing importance of managing and extracting meaningful insights from heterogeneous data in contemporary data management landscapes.

The application of various technologies and approaches in the extraction of diverse types of data is exemplified in Ismail et al., 2019b, where the authors scrutinize manufacturing process pipelines. In this study, customized tools constitute the predominant method for data extraction. While the intricacies of custom tools can pose challenges in paper analysis, the current study identifies multiple technologies and methods. These range from hardware devices capable of collecting signals from different sources (Wang et al., 2022b) to the integration with PLCs using protocols like MTConnect or OPC-UA (Michalkowski et al., 2023). Additionally, the study explores the use of shared dataspaces and asset administration shells (Neubauer et al., 2023) as well as the adoption of source connectors and middleware (Nagorny et al., 2020; Park & Huh, 2023). This diverse array of technologies underscores the multifaceted nature of data extraction methods employed in managing and analyzing manufacturing processes.

The use of diverse technologies and methods tailored to specific problems and convenience is a reality. Effectively managing heterogeneous data is crucial for extracting valuable insights, making informed decisions, and ensuring compatibility in integrated systems. This challenge often goes hand in hand with another concept: preprocessing.

Preprocessing

In the process of data collection, there are often opportunities for minor adjustments to the extracted data to enhance its suitability for utilization. These adjustments may include straightforward operations such as aggregation, where data from various time frames is consolidated for more effective processing, analysis, or storage (Para et al., 2019; Saqlain & Shim et al., 2019). Additionally, it may involve operations like joining basic information, such as the part number of a manufactured piece (Farooqui et al., 2020), as well as tasks like filling or removing null numbers or outliers, among a myriad of other potential operations.

Understandably, this operation holds immense significance when dealing with heterogeneous data, as the data often requires aggregation, normalization, and cleaning to be efficiently stored. This ensures that processing and analyses of the data can be conducted accurately. This intricate process can be facilitated through a myriad of tools. Notably, the review highlights that ELT, and ETL processes conducted through streaming are the most commonly encountered methods for achieving these data transformation tasks. A clear example of this case is found in Fahmideh and Beydoun (2019), where the authors use ETL processes to filter and in real-time process the information coming from the shopfloor.

Preprocessing does have evident limitations, primarily in the form of introducing delays in the data flow, from the point of reception to the processing stage. This limitation is particularly impactful in real-time applications, such as medical applications, where swift data processing is critical. A common workaround involves the implementation of parallel processing. This allows real-time applications to take precedence, ensuring dedicated and prioritized processing, while other non-time-sensitive data can be handled with less urgency. This example applied to the industry can is described in Majeed et al. (2021) where the authors provide a framework that divides the time-critical and non-time critical applications, allowing a faster response where a larger volume of resources and more careful design data pipeline when time requirements exist.

Lastly, but no less important, the efforts put into preprocessing to handle data gain special significance, particularly when the information is to be analyzed or stored in the cloud. The necessity arises from the fact that data must traverse the internet, and in an era where the challenge of big data is increasingly prevalent, dealing with colossal volumes of information poses a substantial hurdle. Transmitting all this untreated data to the cloud is often unfeasible. Not all collected data holds equal importance for the company or requires analysis. Therefore, culling irrelevant data becomes imperative, serving not only to conserve storage space but also to optimize bandwidth usage (Alabadi et al., 2022). Even in cases where all the information is deemed valuable, it is judicious to compress and prepare the data locally before transmitting it to the cloud.

Preprocessing is a critical process for optimizing the storage and processing/analysis of information harvested within a system. It plays a pivotal role in differentiating between elevated storage costs, where useless and missing values are stored without any present or future utility, and preparing information for tasks such as training models, yielding better results within a more efficient time span. Moreover, this process is essential for streamlining the transmission of information to the cloud or other non-local systems, alleviating the strain on the available bandwidth for the company.

Cloud computing

The third process to be discussed is Cloud Computing. Frequently, companies either lack the resources or prefer to avoid the complexities associated with managing their own data storage, preprocessing, or processing/analysis. In such cases, they can leverage services provided by companies like Amazon and Microsoft to handle all aspects of information processing (Kabugo et al., 2020; Leang et al., 2019).

The cloud computing environment presents numerous advantages but also entails various nuances that should be comprehended before utilization, as there is a risk that this process may not align with the user's requirements. Cloud computing offers seemingly limitless storage and processing capabilities, making it an ideal choice for storing vast amounts of data without concerns about where and how to store it. It is particularly well-suited for training heavy or multiple models without the need for a significant investment in hardware infrastructure. The authors in Bonnard et al. (2021) use the infrastructure already build in several industries, such as PLC or enterprise resource planning (ERP), to send all the information to a representational state transfer (REST) application programming interface (API) connected to the cloud computing, for storing and analyzing the data.

Furthermore, this environment boasts the advantage of accessibility from anywhere, enabling data visualization through dashboards via a simple web browser. Users can also perform configurations, change parameters, or train models seamlessly. Cloud computing facilitates both real-time streaming processing and batch processing, providing users with the flexibility to choose and tailor their usage according to their needs (Tejada, 2024a, 2024b). Additionally, specialized personnel within the cloud computing system can offer support and assistance in case of any issues.

However, this system comes with clear disadvantages. Given that cloud computing is typically a third-party service, users are often required to pay a utilization or monthly fee. Additionally, as mentioned earlier, some of these systems offer preprocessing in the cloud. For substantial volumes of data, this can be impractical, as a significant bandwidth must be allocated for transmitting irrelevant information that will be discarded during cloud-based preprocessing (Alabadi et al., 2022). Furthermore, latency can be a significant issue, particularly with large data volumes, posing a risk for real-time applications that are sensitive to millisecond delays. Adopting an approach without local preprocessing in such cases may be less suitable.

The cloud computing environment stands as an essential tool, serving purposes ranging from visualization and storage to complex model training. However, it is not a one-size-fits-all solution, as there are clear disadvantages that emerge if implemented without a thorough assessment of the specific problems and systems in operation. From an ethical perspective, the storage of potentially sensitive information on property not owned by the company may raise concerns, as there is no certainty that the information is not being accessed or used by third parties. Additionally, as briefly mentioned, time-sensitive applications may encounter real challenges, with the guarantee of timely delivery falling short for various reasons. These can include issues with internet connection, the transmission of large volumes of information on low bandwidth, and the inherently increased latency of the system when compared to local systems.

Real-time

Time-sensitive activities are among the most crucial operations carried out on a shop floor. This significance arises from the fact that shop floors typically adhere to strict schedules and timelines, aiming to minimize idle time and fulfill orders within agreed-upon timeframes. Consequently, when it comes to data analyses, certain processes must be inherently time sensitive. This urgency is essential to ensure the production of the highest-quality products within the shortest possible timeframe. While the focus is often on shopfloors, it is noteworthy that real-time analyses also hold vital importance in various other domains, including medicine, where timely decisions are critical, and in scenarios related to natural disasters, where immediate responses can be pivotal.

The necessity for real-time processing is often pervasive across different types of data, extending beyond a single device or data type. This is exemplified in Wang et al. (2022b) and Chen (2020) where the authors showcase the use of streaming services to process data from heterogeneous sources in real-time. One approach involves no preprocessing, as the collected data is inherently meant for preprocessing, while the other involves preprocessing to tailor the data before processing.

Real-time application use cases vary, as demonstrated by the authors in Fang et al. (2020), who focus on real-time visibility of processes utilizing streaming and cloud computing environments. Another example is found in Majeed et al. (2019), where the emphasis is on real-time optimization of machines in an additive manufacturing shopfloor. These examples underscore the versatility of real-time processing across diverse applications and industries.

The utilization of real-time applications is extensive and varies significantly from one company to another and from sector to sector. In certain cases, the primary objective may be the overall visualization of the factory processes, providing a comprehensive view of operations. On the other hand, in different scenarios, real-time sensitive feedback control of machinery, particularly in precision processes like CNC, becomes imperative to ensure the production of the highest quality products (Kim & Lee, 2021; Lu & Xu, 2019). The diversity in objectives highlights the adaptability of real-time applications to cater to the specific needs and priorities of different industries and companies.

The real-time concept is a crucial consideration for all data management systems, necessitating awareness and preparation to handle time-sensitive data. Starting from the extraction phase, it's essential to ensure precision and speed. In preprocessing, optimized algorithms play a key role in efficiently aggregating and cleaning information. This optimization ensures that analyses and visualizations can be performed swiftly and accurately, providing timely and reliable data for accurate overviews. This capability is fundamental for informed decision-making, whether conducted by humans or machines. The entire data management infrastructure should be equipped to handle time-sensitive data, ensuring responsiveness and effectiveness throughout the entire process.

It is crucial to understand the boundaries of real-time applications. While the discussion often revolves around real-time information and analyses, the reality is more intricate. Firstly, the models used are not trained in real-time; they are typically trained beforehand with data collected previously. The real-time aspect primarily involves the application of the pre-trained models to the newly collected data (Farooqui et al., 2020).

Secondly, despite the use of the term "real-time," nothing is truly instantaneous. Handling large volumes of information and establishing network connections are necessary processes that introduce inherent delays.

Thirdly, the concept of real-time is always constrained by the technology in use. If a sensor cannot provide faster readings or a hard drive has limitations on the speed of storage, the system cannot operate at a faster rate.

Finally, the necessity of real-time analyses should be carefully considered. While it may be crucial for certain machines or processes, it may not be equally important for others. The impact of receiving information within milliseconds versus minutes may be irrelevant for the end goal of some processes. Additionally, the pursuit of real-time applications without a clear necessity can potentially overwhelm the infrastructure without delivering substantial benefits (Wu, 2021).

Discussion

In this section, it will be discussed an overview of the findings from the survey, provide responses to the research questions, and offer final insights and future directions based on the literature review and survey results.

Survey findings discussion

After providing an overview of the main concepts and presenting various papers that delve into the core of each concept, it is crucial to draw attention to areas that did not receive as much focus, particularly the intersection between heterogeneous data, real-time, and cloud computing, as depicted in Fig. 10. Throughout the research and evaluation of the papers, one category consistently remained empty across all 76 papers assessed—DVCS. This category highlights the challenge of dealing with a multitude of different data types that require processing in milliseconds or, at most, seconds, approaching quasi real-time. DVCS is also part of the cloud computing section, signifying that various data types are processed (quasi) in real-time within a cloud environment.

From the initial analysis, it becomes evident that this setup might pose some complications, particularly because the category does not involve preprocessing. In a thought exercise, it is possible to discern potential challenges with this configuration. Processing a myriad of different data types into useful models or applications without specific preprocessing can be exceptionally challenging. Another noteworthy issue arises when the volume of heterogeneous data is substantial. In such a scenario, sending all this information to the cloud without any filter can become problematic, introducing delays and consuming significant portions of bandwidth with information that may not be useful for the processing.

The previously discussed use case is among the few examples illustrating how a well-constructed and optimized data management system can make the difference between a feasible and useful utilization of data and a challenging and inefficient one. Upon reanalyzing Fig. 10, it becomes evident that one of the intersections with a substantial number of papers is DVRCP, where the only difference from the previously discussed category (DVCP) is the inclusion of preprocessing. This understandably places increased emphasis on the concept of preprocessing and suggests that a combination of various concepts working together is a commonly favored choice in the industrial environment. This interplay among different concepts proves beneficial, delivering better, faster, and more comprehensive results to the end user.

Another noteworthy topic worth discussing is that, while nearly all the different concepts presented in the context review are referenced in the survey papers, there is one particular concept not directly mentioned in any of the papers gathered for the survey—data lakehouse. Despite data warehouse and data lake being closely related concepts explicitly referred to and used in the surveyed articles, the term "data lakehouse" is never directly mentioned. However, upon careful analysis of selected papers, some hints suggest that an architecture similar to the data lakehouse concept might be utilized, although authors never explicitly claim it. One notable example is Zhang (2021).

In the authors' viewpoint, there might be several reasons for the absence of a clear mention or application of the concept. One of the most understandable reasons is the novelty of the idea. While the concepts of data lake and data warehouse have been around for some years, with data warehouse being the oldest, the amalgamation and formalization of these two concepts as one might not be as well-established, especially in certain industrial environments. Another reason could be that there might be a separation of both concepts in the data management environment, with some data being fed to the data warehouse architecture and other data to the data lake. Having these two architectures separated might not be considered as a data lakehouse. This perspective could also be shared by fellow researchers. Despite some papers in the survey being well-documented and written, understanding some of these subtle nuances of architecture dynamics becomes more challenging.

Research questions overview

The information gathered from the survey and literature review on data management in the industrial sector provided insights into the prevalent data management approaches in this field. In response to RQ1, "Which data management approaches can be found in industrial applications?", four main settings were identified: data collection, data preprocessing, data storage, and data processing/analysis. These settings represent the primary focus areas in industrial data management. However, these settings are not always clear-cut; for instance, data collection is always present but varies widely in methods and applications. For the survey, key characteristics relevant to constructing a data management pipeline were included and referenced in Fig. 10. These characteristics encompass data heterogeneity, real-time environment depiction, preprocessing usage, and the adoption of cloud computing environments.

Finally, it is worth mentioning the lack of method in the approaches found, corroborating the lack of clear responses to RQ2. In the survey presented in this article, several implementations within the same type of company were observed. For example, in the machining use case, mainly using CNC machines, papers (Chen, 2020; Corallo et al., 2022; Koprov et al., 2022; Lu & Xu, 2019; Luo et al., 2022; Wang et al., 2022b) are responsible for the machining papers in the survey. While these papers are similar in their applications, they employ various methods of data extraction, ranging from external IoT sensors to pub/sub directly from the CNC machines, to the usage of PLC and OPC-UA, or even data acquisition instruments such as NI-DAQ (a data acquisition hardware from National Instruments). Across the several reviewed papers, multiple similar cases arise, making it difficult to gather a specific response to "Which approaches are more adequate for different types of industrial sectors and processes?" However, some information about the most commonly used practices could be acquired, such as the utilization of streaming for real-time environments, the major utilization of ad hoc software for local preprocessing or prebuilt infrastructure (Amazon, Google, or Microsoft Azure) for preprocessing in the cloud, and the utilization of pub/sub approaches or the protocol OPC-UA for data collection.

Final insights, limits and future directions

The limitations of this study primarily stem from the chosen criteria used to filter and select the papers for analysis. Due to the vast number of results generated, it was impossible to conduct a thorough analysis of every single paper manually. Consequently, some relevant results were inevitably excluded. For instance, papers from Q2 journals or prestigious conferences were not included in the survey. This does not imply that the findings from these sources are irrelevant, but rather that deliberate choices were made to focus on the most meaningful and impactful information within the available time frame. This prioritization was necessary to ensure a manageable and thorough analysis.

As for the final insights, there are clear and diverse ways of extracting information, there are also multiple ways for that information to be processed, ranging from batch to streaming, from cloud to local, and all of them were found in these papers. While some frameworks (Corallo et al., 2022; Guo et al., 2023; Kozjek et al., 2020; Majeed et al., 2019; Saqlain & Shim et al., 2019; Wang & Luo, 2021; Yu et al., 2022; Zhang et al., 2023) and principles (Allian et al., 2021; Corallo et al., 2023a; Gopalakrishnan et al., 2022; Ismail et al., 2019a; Kumar et al., 2021; Lee & Chien, 2022; Liu et al., 2023a; Mitra & Munir, 2019; Raj et al., 2023; Raut et al., 2021;

Shukla et al., 2019; Wei et al., 2021) are presented and found during a review, there is still room for improvement. Frameworks are often too specific to be useful on a broad scale or too impractical to be implemented. Principles while excellent at highlighting potential pitfalls and challenges and offering solutions, they often lack clear methods for implementation or guidance on the best approach for specific user cases.

A methodology that captures user input and desired outcomes while providing a clear, step-by-step approach, while highlighting essential features and challenges, remains absent. This survey reveals that across the architectures and implementations reviewed, this methodological gap is particularly significant. A well-defined approach would clarify important factors, such as when preprocessing is advantageous, whether processing should occur locally or in the cloud, and the necessity for real-time capabilities. It could also recommend cloud or local computing based on data characteristics, along with other key considerations for effectively managing data flow to achieve optimal results.

Moreover, the authors believe that the data management processes can be augmented with the use of reference architecture of Industry 4.0, such as Reference Architectural Model Industrie 4.0 (RAMI 4.0) or Industrial Internet Reference Architecture (IIRA), with the objective of harmonizing the integration of the different technologies and methods. This integration with reference architectures would allow a better description and scalability of the components present in the data management system, allowing even easier construction of such system and a better understanding of their functionalities and capabilities.

For future directions, the development of a concrete methodology stands as a promising path to simplify the design and implementation of data management infrastructures and pipelines. Establishing such methodologies would make these processes more straightforward and efficient. Furthermore, the integration of AI can enhance the understanding and facilitate the construction of these infrastructures. From an academic perspective, prioritizing research on methodologies for building data management infrastructures is essential. This research could be closely linked with industry sectors that specialize in system integration, enabling the development of practical, robust methodologies. This would not only ensure that the methodology can be effectively translated into real-world environments but also encourage a smoother exchange of knowledge between academia and industry.

As data is agnostic in its nature, this methodology problem might also be relevant to other fields besides industrial data management. Ideally, this methodology should work for all, or at least for a relatively large number of fields that deal with data, either as their primary objective or as a secondary one. It is important, however, to understand that this work is primarily focused on the industrial case, even if a generalization of conclusions obtained might still be possible.

Conclusion

Data is omnipresent in the processes constructed by humans, and recently, humans have come to appreciate how useful this data can be for optimizing their work. The present survey is conducted to understand the impact of this data in an industrial environment perspective and how data management is conducted in this context.

In this paper, is present a context overview of critical concepts related to the data management, in industrial environments. This context overview serves as a starting point, establishing a baseline for understanding the core subjects of data management. The context overview delves into key concepts, including big data, data extraction, preprocessing, storage, processing, and essential architectures relevant to their utilization.

In the survey section, 76 papers are classified, emphasizing critical concepts at the core of data management in each paper. This classification led to the identification of four crucial concepts that facilitate a better understanding of the important aspects of data management utilization in the industrial sector. These include recognizing the heterogeneity of the data, which involves diverse types and sources of data. Another concept is the adoption, or lack thereof, of a real-time data processing approach, which identifies whether real-time data processing is utilized. Additionally, the utilization of preprocessing during data preparation for analysis or storage is examined to understand the extent to which preprocessing is used to prepare data. Finally, the adoption or avoidance of a cloud environment for various tasks is explored, examining the use of cloud environments for data management tasks.

The discussion delves into the survey findings, focusing on the intersections of concepts with varying paper volumes and providing rationale for these observations. It underscores the significance of the concepts elucidated in the survey and context overview, while noting their presence or absence in certain survey papers. Additionally, the discussion addresses the reasoning behind the observed trends and characteristics during the survey, as well as the deficiencies in methodologies and practices seen in the implementations. It emphasizes the lack of a clear method outlining the data pipeline approach in response to user inputs and needs.

While the paper predominantly concentrates on the industrial setting, it is evident that data's relevance extends to various fields, proving crucial for enhancing both new and established processes. The current paper establishes a significant baseline by articulating clear concepts and offering a critical perspective on the existing state of data management within industrial environments.

Acknowledgements This work was funded in part by Fundação para a Ciência e Tecnologia through the program UIDB/00066/2020 and Center of Technology and Systems (CTS).

Funding Open access funding provided by FCT|FCCN (b-on).

Data availability The data presented in this study are available on request from the corresponding author.

Declarations

Competing interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Achouch, M., Dimitrova, M., Ziane, K., Karganroudi, S. S., Dhouib, R., Ibrahim, H., & Adda, M. (2022). On predictive maintenance in industry 4.0: Overview, models, and challenges. *Applied Sciences, 12*(16), 8081. https://doi.org/10.3390/app12168081

Adolphs, P., & Epple, U. (2015). Status report reference architecture model industrie 4.0 (RAMI4.0). Retrieved from www.vdi.de

Alabadi, M., Habbal, A., & Wei, X. (2022). Industrial Internet of Things: Requirements, architecture, challenges, and future research directions. *IEEE Access, 10*, 66374–66400. https://doi.org/10.1109/ACCESS.2022.3185049

Allian, A. P., Schnicke, F., Antonino, P. O., Rombach, D., & Nakagawa, E. Y. (2021). Architecture drivers for trustworthy interoperability in industry 4.0. *IEEE Systems Journal, 15*(4), 5454–5463. https://doi.org/10.1109/JSYST.2020.3041259

Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021). Lakehouse: A new generation of open platforms that unify data warehousing and advanced analytics. In *Proceedings of CIDR* (vol 8, p. 28)

Bastas, A. (2021). Sustainable manufacturing technologies: A systematic review of latest trends and themes. *Sustainability, 13*(8), 4271. https://doi.org/10.3390/su13084271

Belhadi, A., Zkik, K., Cherrafi, A., Yusof, S. M., & S. El fezazi,. (2019). Understanding big data analytics for manufacturing processes: Insights from literature review and multiple case studies. *Computers & Industrial Engineering, 137*, 106099. https://doi.org/10.1016/j.cie.2019.106099

Bi, Z., Jin, Y., Maropoulos, P., Zhang, W.-J., & Wang, L. (2023). Internet of things (IoT) and big data analytics (BDA) for digital manufacturing (DM). *International Journal of Production Research, 61*(12), 4004–4021. https://doi.org/10.1080/00207543.2021.1953181

Bonnard, R., Arantes, M. D. S., Lorbieski, R., Vieira, K. M. M., & Nunes, M. C. (2021). Big data/analytics platform for Industry 4.0 implementation in advanced manufacturing context. *The International Journal of Advanced Manufacturing Technology, 117*(5–6), 1959–1973. https://doi.org/10.1007/s00170-021-07834-5

Brous, P., Janssen, M., & Krans, R. (2020). Data governance as success factor for data science. *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (pp. 431–442). Springer.

Camarinha-Matos, L. M. (2016). Collaborative smart grids—A survey on trends. *Renewable and Sustainable Energy Reviews, 65*, 283–294. https://doi.org/10.1016/j.rser.2016.06.093

Cecchinel, C., Jimenez, M., Mosser, S., & Riveill, M. (2014). An architecture to support the collection of big data in the Internet of Things. *2014 IEEE world congress on services* (pp. 442–449). IEEE.

Cerquitelli, T., et al. (2021). Manufacturing as a data-driven practice: methodologies, technologies, and tools. *Proceedings of the IEEE, 109*(4), 399–422. https://doi.org/10.1109/JPROC.2021.3056006

Chandra, P., & Gupta, M. K. (2018). Comprehensive survey on data warehousing research. *International Journal of Information Technology (Singapore), 10*(2), 217–224. https://doi.org/10.1007/s41870-017-0067-y

Chen, W. (2020). Intelligent manufacturing production line data monitoring system for industrial internet of things. *Computer Communications, 151*, 31–41. https://doi.org/10.1016/j.comcom.2019.12.035

Corallo, A., Crespino, A. M., Del Vecchio, V., Gervasi, M., Lazoi, M., & Marra, M. (2023b). Evaluating maturity level of big data management and analytics in industrial companies. *Technological Forecasting and Social Change, 196*, 122826. https://doi.org/10.1016/j.techfore.2023.122826

Corallo, A., Crespino, A. M., Del Vecchio, V., Lazoi, M., & Marra, M. (2023a). Understanding and defining dark data for the manufacturing industry. *IEEE Transactions on Engineering Management, 70*(2), 700–712. https://doi.org/10.1109/TEM.2021.3051981

Corallo, A., Crespino, A. M., Lazoi, M., & Lezzi, M. (2022). Model-based big data analytics-as-a-service framework in smart manufacturing: A case study. *Robotics and Computer-Integrated Manufacturing, 76*, 102331. https://doi.org/10.1016/j.rcim.2022.102331

Corradi, A., Di Modica, G., Foschini, L., Patera, L., & Solimando, M. (2022). SIRDAM4.0: A support infrastructure for reliable data acquisition and management in industry 4.0. *IEEE Transactions on Emerging Topics in Computing, 10*(3), 1605–1620. https://doi.org/10.1109/TETC.2021.3111974

Cui, Y., Kara, S., & Chan, K. C. (2020). Manufacturing big data ecosystem: A systematic literature review. *Robotics and Computer-Integrated Manufacturing, 62*, 101861. https://doi.org/10.1016/j.rcim.2019.101861

Dachyar, M., Zagloel, T. Y. M., & Saragih, L. R. (2019). Knowledge growth and development: Internet of Things (IoT) research, 2006–2018. *Heliyon*. https://doi.org/10.1016/j.heliyon.2019.e02264

Dai, H. N., Wang, H., Xu, G., Wan, J., & Imran, M. (2020). Big data analytics for manufacturing internet of things: Opportunities, challenges and enabling technologies. *Enterp Inf Syst, 14*(9–10), 1279–1303. https://doi.org/10.1080/17517575.2019.1633689

Deshmukh, R. A., Jayakody, D., Schneider, A., & Damjanovic-Behrendt, V. (2021). Data spine: A federated interoperability enabler for heterogeneous IoT platform ecosystems. *Sensors, 21*(12), 4010. https://doi.org/10.3390/s21124010

Dogan, A., & Birant, D. (2021). Machine learning and data mining in manufacturing. *Expert Systems with Applications, 166*, 114060. https://doi.org/10.1016/j.eswa.2020.114060

Duan, L., & Da Xu, L. (2021). Data analytics in industry 4.0: A survey. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-021-10190-0

Fahmideh, M., & Beydoun, G. (2019). Big data analytics architecture design—An application in manufacturing systems. *Computers & Industrial Engineering, 128*, 948–963. https://doi.org/10.1016/j.cie.2018.08.004

Fang, P., Yang, J., Zheng, L., Zhong, R. Y., & Jiang, Y. (2020). Data analytics-enable production visibility for cyber-physical production systems. *Journal of Manufacturing Systems, 57*, 242–253. https://doi.org/10.1016/j.jmsy.2020.09.002

Farooqui, A., Bengtsson, K., Falkman, P., & Fabian, M. (2020). Towards data-driven approaches in manufacturing: An architecture to collect sequences of operations. *International Journal of Production Research, 58*(16), 4947–4963. https://doi.org/10.1080/00207543.2020.1735660

Filz, M.-A., Bosse, J. P., & Herrmann, C. (2023). Digitalization platform for data-driven quality management in multi-stage manufacturing systems. *Journal of Intelligent Manufacturing*. https://doi.org/10.1007/s10845-023-02162-9

Fortoul-Diaz, J. A., Carrillo-Martinez, L. A., Centeno-Tellez, A., Cortes-Santacruz, F., Olmos-Pineda, I., & Flores-Quintero, R. R. (2023). A smart factory architecture based on industry 4.0 technologies: Open-source software implementation. *IEEE Access, 11*, 101727–101749. https://doi.org/10.1109/ACCESS.2023.3316116

Fowler, J. W., & Mönch, L. (2022). A survey of scheduling with parallel batch (p-batch) processing. *European Journal of Operational Research, 298*(1), 1–24. https://doi.org/10.1016/j.ejor.2021.06.012

Gholipour, E., & Bastas, A. (2023). State-of-the-art review of neural network applications in pharmaceutical manufacturing: Current state and future directions. *Journal of Intelligent Manufacturing*. https://doi.org/10.1007/s10845-023-02206-0

Gökalp, M. O., Gökalp, E., Kayabay, K., Koçyiğit, A., & Eren, P. E. (2021). Data-driven manufacturing: An assessment model for data science maturity. *Journal of Manufacturing Systems, 60*, 527–546. https://doi.org/10.1016/j.jmsy.2021.07.011

GonzálezGarcía, C., & Álvarez-Fernández, E. (2022). What is (not) big data based on its 7 vs challenges: A survey. *Big Data and Cognitive Computing, 6*(4), 158. https://doi.org/10.3390/bdcc6040158

Gopalakrishnan, M., Subramaniyan, M., & Skoogh, A. (2022). Data-driven machine criticality assessment—Maintenance decision support for increased productivity. *Production Planning & Control, 33*(1), 1–19. https://doi.org/10.1080/09537287.2020.1817601

Göppert, A., Grahn, L., Rachner, J., Grunert, D., Hort, S., & Schmitt, R. H. (2023). Pipeline for ontology-based modeling and automated deployment of digital twins for planning and control of manufacturing systems. *Journal of Intelligent Manufacturing, 34*(5), 2133–2152. https://doi.org/10.1007/s10845-021-01860-6

Gualtieri, M., & Yuhanna, N. (2016). The Forrester Wave™: Big Data Hadoop Distributions; Q1 2016. https://www.forrester.com/report/The-Forrester-Wave-Big-Data-Hadoop-Distributions-Q1-2016/RES121574

Guo, J., Cheng, Y., Wang, D., Tao, F., & Pickl, S. (2023). Industrial dataspace for smart manufacturing: Connotation, key technologies, and framework. *International Journal of Production Research, 61*(12), 3868–3883. https://doi.org/10.1080/00207543.2021.1955996

Hajjaji, Y., Boulila, W., Farah, I. R., Romdhani, I., & Hussain, A. (2021). Big data and IoT-based applications in smart environments: A systematic review. *Computer Science Review*. https://doi.org/10.1016/j.cosrev.2020.100318

Harby, A. A., & Zulkernine, F. (2022). From data warehouse to lakehouse: A comparative review. *2022 IEEE international conference on big data (big data)* (pp. 389–395). IEEE.

Haryono, E. M., Fahmi, Tri W, A. S., Gunawan, I., Hidayanto, A. N., & Rahardja, U. (2020). Comparison of the E-LT vs ETL method in data warehouse implementation: A qualitative study. *Proceedings—2nd international conference on informatics, multimedia, cyber, and information system, ICIMCIS 2020* (pp. 115–120). Institute of Electrical and Electronics Engineers Inc.

Helu, M., Sprock, T., Hartenstine, D., Venketesh, R., & Sobel, W. (2020). Scalable data pipeline architecture to support the industrial internet of things. *CIRP Annals, 69*(1), 385–388. https://doi.org/10.1016/j.cirp.2020.04.006

Hinojosa-Palafox, E. A., Rodríguez-Elías, O. M., Hoyo-Montaño, J. A., Pacheco-Ramírez, J. H., & Nieto-Jalil, J. M. (2021). An analytics environment architecture for industrial cyber-physical systems big data solutions. *Sensors, 21*(13), 4282. https://doi.org/10.3390/s21134282

Hlupic, T., Orescanin, D., Ruzak, D., & Baranovic, M. (2022). An overview of current data lake architecture models. *2022 45th jubilee international convention on information, communication and electronic technology (MIPRO)* (pp. 1082–1087). IEEE.

Horak, T., Strelec, P., Kebisek, M., Tanuska, P., & Vaclavova, A. (2022). Data integration from heterogeneous control levels for the purposes of analysis within industry 4.0 concept. *Sensors, 22*(24), 9860. https://doi.org/10.3390/s22249860

Huacarpuma, R. C., De SousaJunior, R., De Holanda, M., De Oliveira Albuquerque, R., Villalba, L. G., & Kim, T.-H. (2017). Distributed data service for data management in internet of things middleware. *Sensors, 17*(5), 977. https://doi.org/10.3390/s17050977

IBM Cloud Education. (2023). ELT vs. ETL: What's the difference? Retrieved from https://www.ibm.com/blog/elt-vs-etl-whats-the-difference/2/2

Isah, H., Abughofa, T., Mahfuz, S., Ajerla, D., Zulkernine, F., & Khan, S. (2019). A survey of distributed data stream processing frameworks. *IEEE Access, 7*, 154300–154316. https://doi.org/10.1109/ACCESS.2019.2946884

Ishikiriyama, C. S., & Gomes, C. F. S. (2019). *Big data: A global overview* (pp. 35–50). Springer.

Ismail, A., Truong, H.-L., & Kastner, W. (2019a). Manufacturing process data analysis pipelines: A requirements analysis and survey. *J Big Data, 6*(1), 1. https://doi.org/10.1186/s40537-018-0162-3

Ismail, A., Truong, H.-L., & Kastner, W. (2019b). Manufacturing process data analysis pipelines: A requirements analysis and survey. *J Big Data, 6*(1), 1. https://doi.org/10.1186/s40537-018-0162-3

Izagirre, U., Andonegui, I., Landa-Torres, I., & Zurutuza, U. (2022). A practical and synchronized data acquisition network architecture for industrial robot predictive maintenance in manufacturing assembly lines. *Robotics and Computer-Integrated Manufacturing, 74*, 102287. https://doi.org/10.1016/j.rcim.2021.102287

Javed, M., Nagabhushan, P., & Chaudhuri, B. B. (2018). A review on document image analysis techniques directly in the compressed domain. *Artificial Intelligence Review, 50*(4), 539–568. https://doi.org/10.1007/s10462-017-9551-9

Jurmu, M., et al. (2023). Exploring the role of federated data spaces in implementing twin transition within manufacturing ecosystems. *Sensors, 23*(9), 4315. https://doi.org/10.3390/s23094315

Kabugo, J. C., Jämsä-Jounela, S.-L., Schiemann, R., & Binder, C. (2020). Industry 4.0 based process data analytics platform: A waste-to-energy plant case study. *International Journal of Electrical Power & Energy Systems, 115*, 105508. https://doi.org/10.1016/j.ijepes.2019.105508

Kahveci, S., Alkan, B., Ahmad, M. H., Ahmad, B., & Harrison, R. (2022). An end-to-end big data analytics platform for IoT-enabled smart factories: A case study of battery module assembly system for electric vehicles. *Journal of Manufacturing Systems, 63*, 214–223. https://doi.org/10.1016/j.jmsy.2022.03.010

Kammerer, K., Pryss, R., Hoppenstedt, B., Sommer, K., & Reichert, M. (2020). Process-driven and flow-based processing of industrial sensor data. *Sensors, 20*(18), 5245. https://doi.org/10.3390/s20185245

Kim, J., & Lee, J. Y. (2021). Server-edge dualized closed-loop data analytics system for cyber-physical system application. *Robot Comput Integr Manuf, 67*, 102040. https://doi.org/10.1016/j.rcim.2020.102040

Koprov, P., Ramachandran, A., Lee, Y.-S., Cohen, P., & Starly, B. (2022). Streaming machine generated data via the MQTT Sparkplug B protocol for smart factory operations. *Manufacturing Letters, 33*, 66–73. https://doi.org/10.1016/j.mfglet.2022.07.016

Kozjek, D., Vrabič, R., Rihtaršič, B., Lavrač, N., & Butala, P. (2020). Advancing manufacturing systems with big-data analytics: A conceptual framework. *International Journal of Computer Integrated Manufacturing, 33*(2), 169–188. https://doi.org/10.1080/0951192X.2020.1718765

Kumar, N., Kumar, G., & Singh, R. K. (2021). Big data analytics application for sustainable manufacturing operations: Analysis of strategic factors. *Clean Technologies and Environmental Policy, 23*(3), 965–989. https://doi.org/10.1007/s10098-020-02008-5

Kuzlu, M., Kalkavan, H., Gueler, O., Zohrabi, N., Martin, P. J., & Abdelwahed, S. (2022). An end to end data collection architecture for IoT devices in smart cities. *2022 Ieee power and energy society innovative smart grid technologies conference, ISGT 2022* (pp. 1–12). Institute of Electrical and Electronics Engineers Inc.

Leang, B., Ean, S., Ryu, G.-A., & Yoo, K.-H. (2019). Improvement of Kafka streaming using partition and multi-threading in big data environment. *Sensors, 19*(1), 134. https://doi.org/10.3390/s19010134

Lee, C.-Y., & Chien, C.-F. (2022). Pitfalls and protocols of data science in manufacturing practice. *Journal of Intelligent Manufacturing, 33*(5), 1189–1207. https://doi.org/10.1007/s10845-020-01711-w

Liu, J.-C., Hsu, C.-H., Zhang, J.-H., Kristiani, E., & Yang, C.-T. (2023b). An event-based data processing system using Kafka container cluster on Kubernetes environment. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-023-08326-1

Liu, Y., Yu, W., Rahayu, W., & Dillon, T. (2023a). An evaluative study on IoT ecosystem for smart predictive maintenance (IoT-SPM) in manufacturing: Multiview requirements and data quality. *IEEE Internet of Things Journal, 10*(13), 11160–11184. https://doi.org/10.1109/JIOT.2023.3246100

Lu, Y., & Xu, X. (2019). Cloud-based manufacturing equipment and big data analytics to enable on-demand manufacturing services. *Robotics and Computer-Integrated Manufacturing, 57*, 92–102. https://doi.org/10.1016/j.rcim.2018.11.006

Luo, D., Guan, Z., He, C., Gong, Y., & Yue, L. (2022). Data-driven cloud simulation architecture for automated flexible production lines: Application in real smart factories. *International Journal of Production Research, 60*(12), 3751–3773. https://doi.org/10.1080/00207543.2021.1931977

Ma, S., Ding, W., Liu, Y., Ren, S., & Yang, H. (2022). Digital twin and big data-driven sustainable smart manufacturing based on information management systems for energy-intensive industries. *Applied Energy, 326*, 119986. https://doi.org/10.1016/j.apenergy.2022.119986

Majeed, A., et al. (2021). A big data-driven framework for sustainable and smart additive manufacturing. *Robotics and Computer-Integrated Manufacturing, 67*, 102026. https://doi.org/10.1016/j.rcim.2020.102026

Majeed, A., Lv, J., & Peng, T. (2019). A framework for big data driven process analysis and optimization for additive manufacturing. *Rapid Prototyping Journal, 25*(2), 308–321. https://doi.org/10.1108/RPJ-04-2017-0075

Malik, P. K., et al. (2021). Industrial Internet of Things and its applications in industry 4.0: State of the art. *Computer Communications, 166*, 125–139. https://doi.org/10.1016/j.comcom.2020.11.016

Mazumdar, D., Hughes, J., & Onofre, J. (2023). The data lakehouse: Data warehousing and more. Retrieved from http://arxiv.org/abs/2310.08697

Michalkowski, C., Janhsen, J., & Springer, P. (2023). Concept for a generic modular software architecture for the integration of quality relevant data and sample implementation for a laser sintering system. *Progress in Additive Manufacturing, 8*(1), 67–73. https://doi.org/10.1007/s40964-022-00390-8

Mitra, A., & Munir, K. (2019). Influence of Big Data in managing cyber assets. *Built Environment Project and Asset Management, 9*(4), 503–514. https://doi.org/10.1108/BEPAM-07-2018-0098

Mocnej, J., Lojka, T., & Zolotova, I. (2016). Using information entropy in smart sensors for decentralized data acquisition architecture. *2016 IEEE 14th international symposium on applied machine intelligence and informatics (SAMI)* (pp. 47–50). IEEE.

Moktadir, Md. A., Ali, S. M., Paul, S. K., & Shukla, N. (2019). Barriers to big data analytics in manufacturing supply chains: A case study from Bangladesh. *Computers & Industrial Engineering, 128*, 1063–1075. https://doi.org/10.1016/j.cie.2018.04.013

MongoDB. (2023). Unstructured data storage. Retrieved 23 Aug 2023, from https://www.mongodb.com/unstructured-data/storage

Munappy, A. R., Bosch, J., & Olsson, H. H. (2020). Data pipeline management in practice: Challenges and opportunities. In *Proceedings* (pp. 168–184). https://doi.org/10.1007/978-3-030-64148-1_11.

Mussina, A. B., Aubakirov, S. S., & Trigo, P. (2021). An architecture for real-time massive data extraction from social media (pp. 138–145). https://doi.org/10.1007/978-3-030-78759-2_11

Nagorny, K., Scholze, S., Colombo, A. W., & Oliveira, J. B. (2020). A DIN Spec 91345 RAMI 4.0 compliant data pipelining model: An approach to support data understanding and data acquisition in smart manufacturing environments. *IEEE Access, 8*, 223114–223129. https://doi.org/10.1109/ACCESS.2020.3045111

Nambiar, A., & Mundra, D. (2022). An overview of data warehouse and data lake in modern enterprise data management. *Big Data and Cognitive Computing, 6*(4), 132. https://doi.org/10.3390/bdcc6040132

Neubauer, M., et al. (2023). Architecture for manufacturing-X: Bringing asset administration shell, eclipse dataspace connector and OPC UA together. *Manufacturing Letters, 37*, 1–6. https://doi.org/10.1016/j.mfglet.2023.05.002

Ning, H. (2013). *Unit and ubiquitous internet of things*. CRC Press.

Number of connected IoT devices growing 16% to 16.7 billion globally. (2023). Retrieved 8 Sep 2023, from https://iot-analytics.com/number-connected-iot-devices/

Number of Internet of Things (IoT) connected devices worldwide. (2023). Retrieved from https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide

Orescanin, D., & Hlupic, T. (2021). Data lakehouse—A novel step in analytics architecture. *2021 44th international convention on information, communication and electronic technology (MIPRO)* (pp. 1242–1246). IEEE.

Para, J., Del Ser, J., Nebro, A. J., Zurutuza, U., & Herrera, F. (2019). Analyze, sense, preprocess, predict, implement, and deploy (ASPPID): An incremental methodology based on data analytics for cost-efficiently monitoring the industry 4.0. *Engineering Applications of Artificial Intelligence, 82*, 30–43. https://doi.org/10.1016/j.engappai.2019.03.022

Park, S., & Huh, J. H. (2023). A study on big data collecting and utilizing smart factory based grid networking big data using Apache Kafka. *IEEE Access*. https://doi.org/10.1109/ACCESS.2023.3305586

Peng, C., & ChunHao, D. (2022). Monitoring multi-domain batch process state based on fuzzy broad learning system. *Expert Systems with Applications, 187*, 115851. https://doi.org/10.1016/j.eswa.2021.115851

Pfandzelter, T., & Bermbach, D. (2019). IoT data processing in the fog: Functions, streams, or batch processing?. https://doi.org/10.1109/ICFC.2019.00033

Pivoto, D. G. S., de Almeida, L. F. F., da Rosa Righi, R., Rodrigues, J. J. P. C., Lugli, A. B., & Alberti, A. M. (2021). Cyber-physical systems architectures for industrial internet of things applications in Industry 4.0: A literature review". *Journal of Manufacturing Systems, 58*, 176–192. https://doi.org/10.1016/j.jmsy.2020.11.017

Prathima, B. A., Sudha, P. N., & Suresh, P. M. (2020). Shop floor to cloud connect for live monitoring the production data of CNC machines. *International Journal of Computer Integrated Manufacturing, 33*(2), 142–158. https://doi.org/10.1080/0951192X.2020.1718762

Qi, C. (2020). Big data management in the mining industry. *International Journal of Minerals, Metallurgy and Materials, 27*(2), 131–139. https://doi.org/10.1007/s12613-019-1937-z

Qi, Q., & Tao, F. (2019). A smart manufacturing service system based on edge computing, fog computing, and cloud computing. *IEEE Access, 7*, 86769–86777. https://doi.org/10.1109/ACCESS.2019.2923610

Qiu, T., Chi, J., Zhou, X., Ning, Z., Atiquzzaman, M., & Wu, D. O. (2020). Edge computing in industrial Internet of Things: Architecture, advances and challenges. *IEEE Communications Surveys & Tutorials, 22*(4), 2462–2488. https://doi.org/10.1109/COMST.2020.3009103

Raj, R., Kumar, V., & Verma, P. (2023). Big data analytics in mitigating challenges of sustainable manufacturing supply chain. *Operations Management Research*. https://doi.org/10.1007/s12063-023-00408-6

Rajnoha, R., & Hadac, J. (2022). Strategic key elements in big data analytics as driving forces of IoT manufacturing value creation: A challenge for research framework. *IEEE Transactions on Engineering Management, 71*, 1–16. https://doi.org/10.1109/TEM.2021.3113502

Raptis, T. P., Passarella, A., & Conti, M. (2019). Data management in industry 4.0: State of the art and open challenges. *IEEE Access, 7*, 97052–97093. https://doi.org/10.1109/ACCESS.2019.2929296

Raut, R. D., Yadav, V. S., Cheikhrouhou, N., Narwane, V. S., & Narkhede, B. E. (2021). Big data analytics: Implementation challenges in Indian manufacturing supply chains. *Computers in Industry, 125*, 103368. https://doi.org/10.1016/j.compind.2020.103368

Russom, P. (2017). Data lakes purposes, practices, patterns, and platforms; Q1 2017. https://info.talend.com/rs/talend/images/WP_EN_BD_TDWI_DataLakes.pdf

Saggi, M. K., & Jain, S. (2018). A survey towards an integration of big data analytics to big insights for value-creation. *Information Processing and Management, 54*(5), 758–790. https://doi.org/10.1016/j.ipm.2018.01.010

Sahal, R., Breslin, J. G., & Ali, M. I. (2020). Big data and stream processing platforms for Industry 4.0 requirements mapping for a predictive maintenance use case. *Journal of Manufacturing Systems, 54*, 138–151. https://doi.org/10.1016/j.jmsy.2019.11.004

Saqlain, M., Piao, M., Shim, Y., & Lee, J. Y. (2019). Framework of an IoT-based industrial data management for smart manufacturing. *Journal of Sensor and Actuator Networks, 8*(2), 25. https://doi.org/10.3390/jsan8020025

Sarker, S., Arefin, M. S., Kowsher, M., Bhuiyan, T., Dhar, P. K., & Kwon, O. J. (2023). A comprehensive review on big data for industries: challenges and opportunities. *IEEE Access, 11*, 744–769. https://doi.org/10.1109/ACCESS.2022.3232526

Sawadogo, P., & Darmont, J. (2021). On data lake architectures and metadata management. *Journal of Intelligent Information System, 56*(1), 97–120. https://doi.org/10.1007/s10844-020-00608-7

Schmetz, A., et al. (2020). Evaluation of industry 4.0 data formats for digital twin of optical components. *International Journal of Precision Engineering and Manufacturing-Green Technology, 7*(3), 573–584. https://doi.org/10.1007/s40684-020-00196-5

SCImago. Retrieved 31 Oct 2023, from https://www.scimagojr.com/

Sebei, H., HadjTaieb, M. A., & BenAouicha, M. (2018). Review of social media analytics process and big data pipeline. *Social Network Analysis and Mining*. https://doi.org/10.1007/s13278-018-0507-0

Shin, S.-J. (2021). An OPC UA-compliant interface of data analytics models for interoperable manufacturing intelligence. *IEEE Transactions on Industrial Informatics, 17*(5), 3588–3598. https://doi.org/10.1109/TII.2020.3024628

Shukla, N., Tiwari, M. K., & Beydoun, G. (2019). Next generation smart manufacturing and service systems using big data analytics. *Computers & Industrial Engineering, 128*, 905–910. https://doi.org/10.1016/j.cie.2018.12.026

Singh, H. (2021). Big data, industry 4.0 and cyber-physical systems integration: A smart industry context. *Materials Today: Proceedings, 46*, 157–162. https://doi.org/10.1016/j.matpr.2020.07.170

Singh, J., Singh, G., & Bhati, B. S. (2022). The implication of data lake in enterprises: a deeper analytics. *8th international conference on advanced computing and communication systems, ICACCS 2022* (pp. 530–534). Institute of Electrical and Electronics Engineers Inc.

Singhal, B., & Aggarwal, A. (2022). ETL, ELT and reverse ETL: A business case Study. *2nd IEEE international conference on advanced technologies in intelligent control, environment, computing and communication engineering, ICATIECE 2022*. Institute of Electrical and Electronics Engineers Inc.

Sorri, K., Mustafee, N., & Seppänen, M. (2022). Revisiting IoT definitions: A framework towards comprehensive use. *Technol Forecast Soc Change, 179*, 121623. https://doi.org/10.1016/j.techfore.2022.121623

Syed, A., Purushotham, K., & Shidaganti, G. (2020). Cloud storage security risks, practices and measures: A review. *2020 IEEE international conference for innovation in technology (INOCON)* (pp. 1–4). IEEE.

Tardio, R., Mate, A., & Trujillo, J. (2020). An iterative methodology for defining big data analytics architectures. *IEEE Access, 8*, 210597–210616. https://doi.org/10.1109/ACCESS.2020.3039455

Tejada, Z. (2024b). Real-time processing. Retrieved from https://learn.microsoft.com/en-us/azure/architecture/data-guide/big-data/real-time-processing

Tejada, Z. (2024a). Batch processing Azure. Retrieved 28 May 2024, from https://learn.microsoft.com/en-us/azure/architecture/data-guide/big-data/batch-processing

Tewari, S., & Dwivedi, U. D. (2019). Ensemble-based big data analytics of lithofacies for automatic development of petroleum reservoirs. *Computers & Industrial Engineering, 128*, 937–947. https://doi.org/10.1016/j.cie.2018.08.018

Tripathi, S., Muhr, D., Brunner, M., Jodlbauer, H., Dehmer, M., & Emmert-Streib, F. (2021). Ensuring the robustness and reliability of data-driven knowledge discovery models in production and manufacturing. *Frontiers in Artificial Intelligence, 4*, 14. https://doi.org/10.3389/frai.2021.576892

Trunzer, E., Prata, P., Vieira, S., & Vogel-Heuser, B. (2019). Concept and evaluation of a technology-independent data collection architecture for industrial automation. *IECON 2019–45th annual conference of the IEEE industrial electronics society* (pp. 2830–2836). IEEE.

Tufano, A. (2023). Data governance in smart factories: Consistency rules for improved data quality in logistics & operations. *Manuf Lett, 37*, 57–60. https://doi.org/10.1016/j.mfglet.2023.07.019

Villalobos, K., Ramírez-Durán, V. J., Diez, B., Blanco, J. M., Goñi, A., & Illarramendi, A. (2020). A three level hierarchical architecture for an efficient storage of industry 4.0 data. *Computers in Industry, 121*, 103257. https://doi.org/10.1016/j.compind.2020.103257

Wampler, D. (2016). Fast data architectures for streaming applications getting answers now from data sets that never end. O'Reilly Media, 2016, Sebastopol.

Wang, J., Xu, C., Zhang, J., Bao, J., & Zhong, R. (2020). A collaborative architecture of the industrial internet platform for manufacturing systems. *Robotics and Computer-Integrated Manufacturing, 61*, 101854. https://doi.org/10.1016/j.rcim.2019.101854

Wang, J., Xu, C., Zhang, J., & Zhong, R. (2022a). Big data analytics for intelligent manufacturing systems: A review. *Journal of Manufacturing Systems, 62*, 738–752. https://doi.org/10.1016/j.jmsy.2021.03.005

Wang, K., Dave, P., Hanchate, A., Sagapuram, D., Natarajan, G., & Bukkapatnam, S. T. S. (2022b). Implementing an open-source sensor data ingestion, fusion, and analysis capabilities for smart manufacturing. *Manufacturing Letters, 33*, 893–901. https://doi.org/10.1016/j.mfglet.2022.07.109

Wang, P., & Luo, M. (2021). A digital twin-based big data virtual and real fusion learning reference framework supported by industrial internet towards smart manufacturing. *Journal of Manufacturing Systems, 58*, 16–32. https://doi.org/10.1016/j.jmsy.2020.11.012

Wang, R., Gu, C., He, S., Shi, Z., & Meng, W. (2022c). An interoperable and flat Industrial Internet of Things architecture for low latency data collection in manufacturing systems. *Journal of Systems Architecture, 129*, 102631. https://doi.org/10.1016/j.sysarc.2022.102631

Wang, W., Fan, L., Huang, P., & Li, H. (2019). A new data processing architecture for multi-scenario applications in aviation manufacturing. *IEEE Access, 7*, 83637–83650. https://doi.org/10.1109/ACCESS.2019.2925114

Wei, D., et al. (2021). Dataflow management in the internet of things: Sensing, control, and security. *Tsinghua Sci Technol, 26*(6), 918–930. https://doi.org/10.26599/TST.2021.9010029

Wu, H., Yan, Y., Chen, B., Hou, F., & Sun, D. (2022). FADA: A cloud-fog-edge architecture and ontology for data acquisition. *IEEE Transactions on Cloud Computing, 10*(3), 1792–1805. https://doi.org/10.1109/TCC.2020.3014110

Wu, Y. (2021). Cloud-edge orchestration for the internet of things: Architecture and AI-powered data processing. *IEEE Internet of Things Journal, 8*(16), 12792–12805. https://doi.org/10.1109/JIOT.2020.3014845

Yang, C., Lan, S., Wang, L., Shen, W., & Huang, G. G. Q. (2020). Big data driven edge-cloud collaboration architecture for cloud manufacturing: A software defined perspective. *IEEE Access, 8*, 45938–45950. https://doi.org/10.1109/ACCESS.2020.2977846

Yu, W., Dillon, T., Mostafa, F., Rahayu, W., & Liu, Y. (2020). A global manufacturing big data ecosystem for fault detection in predictive maintenance. *IEEE Transactions on Industrial Informatics, 16*(1), 183–192. https://doi.org/10.1109/TII.2019.2915846

Yu, W., Liu, Y., Dillon, T., Rahayu, W., & Mostafa, F. (2022). An integrated framework for health state monitoring in a smart factory employing IoT and big data techniques. *IEEE Internet of Things Journal, 9*(3), 2443–2454. https://doi.org/10.1109/JIOT.2021.3096637

Zhang, C., & Han, J. (2021). Data mining and knowledge discovery (pp. 797–814). https://doi.org/10.1007/978-981-15-8983-6_42

Zhang, J., Liu, J., Zhuang, C., Guo, H., & Ma, H. (2023). A data-driven smart management and control framework for a digital twin shop floor with multi-variety multi-batch production. *The International Journal of Advanced Manufacturing Technology*. https://doi.org/10.1007/s00170-023-10815-5

Zhang, L., Li, F., Wang, P., Su, R., & Chi, Z. (2022). A Blockchain-assisted massive IoT data collection intelligent framework. *IEEE Internet of Things Journal, 9*(16), 14708–14722. https://doi.org/10.1109/JIOT.2021.3049674

Zhang, L., Yuan, H., Chang, S.-H., & Lam, A. (2020b). Research on the overall architecture of Internet of Things middleware for intelligent industrial parks. *The International Journal of Advanced Manufacturing Technology, 107*(3–4), 1081–1089. https://doi.org/10.1007/s00170-019-04310-z

Zhang, N. (2021). A cloud-based platform for big data-driven CPS modeling of robots. *IEEE Access, 9*, 34667–34680. https://doi.org/10.1109/ACCESS.2021.3061477

Zhang, X., Ming, X., & Yin, D. (2020a). Application of industrial big data for smart manufacturing in product service system based on system engineering using fuzzy DEMATEL. *Journal of Cleaner Production, 265*, 121863. https://doi.org/10.1016/j.jclepro.2020.121863

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.