

# RE4DY

MANUFACTURING DATA NETWORKS

Title	D2.2: Digital 4.0 Continuum Reference Framework - v1
-------	--

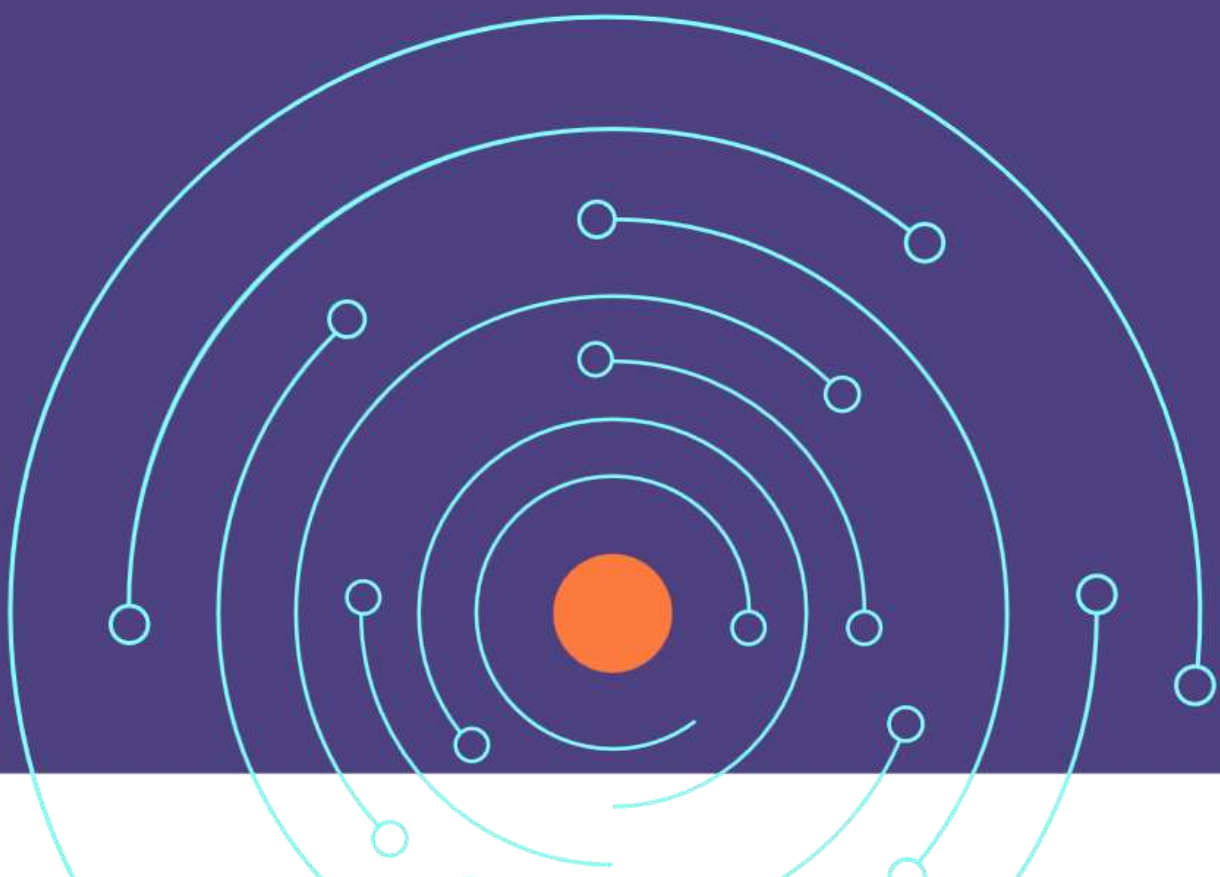
Document Owners	ENG
-----------------	-----

Contributors	ENG, INNO, ATOS, CHAL, KUL, CERTH, UPV, INTRA, SIE
--------------	--

Dissemination	Public
---------------	--------

Date	04/07/2023
------	------------

Version	V01
---------	-----



# Table of contents

Executive Summary .....	8
1 Introduction .....	9
1.1 Context and scope of this document .....	9
1.2 Relationships among WP2 tasks .....	9
2 Methodology .....	10
2.1 Methodological approach .....	10
2.2 Knowledgeable on SoTA.....	11
2.2.1 RAMI 4.0 .....	11
2.2.2 NIST Big Data Reference Architecture.....	13
2.2.3 IIRA Industrial Internet RA.....	15
2.2.4 IDS RAM .....	18
2.2.5 BDV RM.....	19
2.3 Main motivations for a new architecture .....	21
2.4 Built upon key results from the past .....	23
2.5 Bottom-up synergies.....	25
2.6 RE4DY RA v1 design .....	25
2.7 Designed to uptake and exploitation.....	26
3 RE4DY Reference Architecture .....	28
3.1 Business layer.....	29
3.1.1 Ecosystem-based value chain.....	30
3.1.2 Logistics for the future .....	30
3.1.3 Megafactory & E-battery design.....	31
3.1.4 Circular Machining .....	32
3.1.5 Distributed, Green, Zero-X Manufacturing.....	33
3.2 Digital continuity service layer .....	34
3.2.1 Integrated trusted Industrial AI and eExecutable CDT Platform .....	35
3.2.2 Federated Learning Services .....	37
3.2.3 Self-Service analytics and AI marketplace .....	39
3.2.4 Sovereign data clearing house .....	40
3.3 DaaP-toolkits layer.....	42
3.3.1 Data sharing .....	42
3.3.2 Data sovereignty .....	44
3.3.3 Data discovery .....	46
3.3.4 Data storage.....	47



3.3.5	Data cleansing.....	47
3.3.6	Data quality.....	48
3.3.7	Semantic model.....	49
3.3.8	Data container.....	52
3.3.9	Compliance check and assurance.....	54
3.4	Integration layer.....	56
3.4.1	Data Ingestion and ETL Services.....	57
3.4.2	Data sources integration.....	57
3.4.3	Transient data stores.....	59
3.4.4	Data Connection Profile.....	59
3.4.5	Data Space Adapters.....	60
3.4.6	System adapters.....	62
3.5	Computing/Networking continuum.....	62
3.6	Distributed, trustworthy and secure computing toolkit.....	65
3.7	xCDTOps framework.....	66
3.8	Federated Learning and Analytics framework.....	71
3.9	Resiliency & Legal Frameworks: resiliency.....	74
3.10	Resiliency & Legal Frameworks: legal.....	76
3.10.1	Protection of IPR and Data Control Rights.....	76
3.10.2	Compliance with existing data regulations.....	77
4	Conclusions.....	79
5	References.....	80
5.1	EU Legislation and International Treaties.....	80
5.2	Guidelines and Case Law.....	80
5.3	Bibliography.....	80



# Table of figures

*Figure 1: WP2 internal dependencies*..... 9

*Figure 2: The methodological approach in six-stages*..... 10

*Figure 3: RAMI 4.0*..... 11

*Figure 4: NIST Big Data Reference Architecture (NSBDRA)*..... 13

*Figure 5: IIRA viewpoints and relationships with Lifecycle Process and Application Scope* ..... 15

*Figure 6: Functional domains, Crosscutting Functions and System Characteristics* ..... 18

*Figure 7: IDS General structure of RA Model*..... 18

*Figure 8: BDV Reference Model*..... 19

*Figure 9: BOOST 4.0 mapping to RE4DY main needs* ..... 24

*Figure 10: Example of ALIDA asset features collection*..... 25

*Figure 11: RE4DY Reference Architecture (RA)*..... 25

*Figure 12: RE4DY RA mapping to DFA-Digital Service Integration Reference Architecture*...27

*Figure 13: RE4DY RA building blocks* ..... 28

*Figure 14: Business layer building blocks* ..... 29

*Figure 15: Digital continuity service layer building blocks*..... 34

*Figure 16: Cognitive digital twin and industrial AI platform*..... 35

*Figure 17: Product manufacturer perspective*..... 36

*Figure 18: Line machine builder perspective*..... 36

*Figure 19: RA interactions among modules and layers involved in/by “FL services”* ..... 38

*Figure 20: GAIA-X Data Product Self-Description Conceptual Model* ..... 41

*Figure 21: DaaP-toolkits layer building blocks* ..... 42

*Figure 22: Key Idea of Data Sovereignty*..... 44

*Figure 23: Roles, Interactions and Core Components of IDSA RAM*..... 45

*Figure 24: List of Ontologies and Data Models related to Industry 4.0 domain* ..... 50

*Figure 25: OntoCommons Hierarchical Architecture for building Semantic Models* ..... 51

*Figure 26: Process to extract Context Models based on IMF* ..... 52

*Figure 27: Data Container submodules - example of pipelines for historical data which include calls to other RE4DY components*..... 53

*Figure 28: Integration layer building blocks*..... 56

*Figure 29: Data Lifecycle from Raw Data to Transformed Data* ..... 57

*Figure 30: Apache NiFi Flow example regarding Data Transformation* ..... 58

*Figure 31 Data Exchange Services realized by a data connector* ..... 61

*Figure 32: Computing/Networking continuum dimension*..... 62

*Figure 33: Edge-to-cloud computing continuum approach and its advantages (C. Avasalcai, 2020)*..... 64

*Figure 34: Seamless automation, deployment and monitoring from cloud to shopfloor*... 67

*Figure 35: Industrial AI lifecycle: building and operating AI models on scale*..... 67

*Figure 36: Lifecycle of AI solutions* ..... 69

*Figure 37: The RE4DY active resilience framework*..... 75



## Document Status

Leader	ENG
Internal Reviewer 1	POLIMI
Internal Reviewer 2	ATOS
Work Package	WP2: Digital 4.0 Continuum Reference Architecture for Active Resiliency
Deliverable	D2.2: Digital 4.0 continuum reference Framework 1st version
Due Date	M12
Delivery Date	04/07/2023
Version	V01

## Version History

18/04/2023	Deliverable structure, first content drafts
02/05/2023	Updated RA images, added content to §4.8
12/06/2023	Version for review
18/06/2023	Internally reviewed version
23/06/2023	Final version

## Further Information

More information about the project can be found on project website: <https://re4dy.eu/>

## Disclaimer

The views represented in this document only reflect the views of the authors and not the views of the European Union. The European Union is not liable for any use that may be made of the information contained in this document.

Furthermore, the information is provided “as is” and no guarantee or warranty is given that the information is fit for any particular purpose. The user of the information uses it at its sole risk and liability.



## Project Partners

Num	Participant organisation name	Acronym
1	ASOCIACIÓN DE EMPRESAS TECNOLÓGICAS INNOVALIA	INNO
2	CHALMERS TEKNISKA HOGSKOLA AB	Chalmers
3	INTERNATIONAL DATA SPACES EV	IDSA
4	VOLKSWAGEN AUTOEUROPA, LDA	VWAE
5	ASSECO CEIT AS	CEIT
6	UNINOVA-INSTITUTO DE DESENVOLVIMENTO DE NOVAS TECNOLOGIAS-ASSOSIACAO	UNI
7	FILL GESELLSCHAFT MBH	FILL
8	AVL LIST GMBH	AVL
9	VISUAL COMPONENTS OY	VIS
10	UNIVERSIDAD MIGUEL HERNANDEZ DE ELCHE	UMH
11	ATLANTIS ENGINEERING AE	ATLANTIS
12	DATAPIXEL SL	DATA
13	CORE KENTRO KAINOTOMIAS AMKE	CORE
14	UNIVERSITETE I OSLO	UiO
15	GE AVIO	AVIO
16	ENGINEERING-INGENIERIA INFORMATICA SPA	ENG
17	POLITECNICO DI MILANO	POLIMI
18	ATOS IT SOLUTIONS AND SERVICES IBERIA SL	ATOS IT
18.1	ATOS SPAIN SA	ATOS ES
19	KATHOLIEKE UNIVERSITEIT LEUVEN	KU
20	NETCOMPANY-INTRASOFT SA	INTRA
21	NOVA ID FCT - ASSOCIACAO PARA A INOVACAO E DESENVOLVIMENTO DA FCT	NOVA
22	INDUSTRY COMMONS FOUNDATION (INSAMLINGSSTIFTELSE)	ICF
23	ETHNIKO KENTRO EREVNAS KAI TECHNOLOGIKIS ANAPTYXIS	CERTH
24	GRUPO S 21SEC GESTION SA	S21SEC
25	UNIVERSITAT POLITECNICA DE VALENCIA	UPV
26	CONSIGLIO NAZIONALE DELLE RICERCHE	CNR
27	SOCIEDAD ANDALUZA PARA EL DESARROLLO DE LAS TELECOMUNICACIONES SA	SANDETEL
28	SWITZERLAND INNOVATION PARK BIEL/BIENNE AG	SSF
29	GF MACHINING SOLUTIONS AG	GFMS ADVMAN
30	FRAISA SA	Fraisa SA
31	SIEMENS SCHWEIZ AG	SIE



## List of Acronyms/Abbreviations

Acronym / Abbreviation	Description
CI	Continuous Integration
CD	Continuous Delivery
CT	Continuous Testing
CPPS	Cyber-Physical Production System
CPS	Cyber-Physical System
CRM	Customer Relationship Management
DaaP	Data as a Product
DC	Data Container
DCAT	Data Catalog Vocabulary
DCP	Data Connection Profile
DoA	Description of Action
DT	Digital Twin
ERP	Enterprise Resource Planning
FAIR	Findable, Accessible, Interoperable, Reusable
FML	Federated
GDPR	General Data Protection Regulation
HRM	Human Resource Management
I4.0	Industry 4.0
ICT	Information and Communication Technologies
IIoT	Industrial Internet of Things
IoT	Internet of Things
IP	Intellectual Property
IPR	Intellectual Property Rights
IT	Information Technology
KPI	Key Performance Indicator
MDN	Manufacturing Data Networks
MES	Original Equipment Manufacturer
OEM	Manufacturing Execution System
PLM	Product Lifecycle Management
PoC	Proof of Concept
RA	Reference Architecture
SGAM	Smart Grid Architecture Model
SME	Small and Medium Enterprise
TEF	Testing and Experimentation Facility
WP	Work Package



# Executive Summary

This document is the second deliverable of the Work Package 2 which reports on the first version of the Digital 4.0 continuum reference framework. The framework is provided through the design of a Digital Continuum Reference Architecture (RA) that aims to facilitate the implementation of digital continuity across Digital Threads, Data Spaces, Digital Twin workflows and AI/ML/Data pipelines and thus enabling the exploitation of the Data as a Product (DaaP) business model. This model revolves around treating data as a distinct product that can be packaged, marketed, and monetized. Under the DaaP business model, organizations focus on extracting, refining, and delivering data in a structured and meaningful way. Data is thus no longer seen as a by-product of operations but as a valuable offering in itself. This requires identifying valuable data sources, curating and enhancing data quality, and establishing mechanisms for data delivery and access. Additionally, businesses need to consider the legal aspects of data usage, ensuring compliance with relevant regulations and safeguarding data privacy and security.

The Reference Architecture is the result of the converging work carried out by all the tasks of this work package -whose objectives and relationships are depicted within the introduction paragraph- as well as by the synergy established with the technology development team of WP3. Therefore, it serves as an initial endeavour to align technological needs with processes and governance considerations in the form of comprehensive guidelines. These guidelines aim to ensure business resiliency and building trust for the successful adoption of these technologies by the final users. For example, the establishment of a secure and trusted data space is essential for enabling the seamless flow of information along the digital thread. By treating data as a product and ensuring data sovereignty, manufacturers can foster a sense of trust among stakeholders and promote data federation, ultimately leading to improved collaboration and innovation across the industries.

This report also outlines the six-stages approach employed in the design of this architecture, along with the key motivations that led to the development of a novel design. By adopting this approach, the architecture undergoes a continuous enhancement process, incorporating valuable feedback gathered throughout the RE4DY project.





# 1 Introduction

## 1.1 Context and scope of this document

The main objective of this document is to introduce the RE4DY Reference Architecture (RA) with its different dimensions, layers and building blocks, describing the approach adopted and main motivations that led to designing it.

In more detail, the document is structured as follows:

- §1 Introduction: It is the introduction to the document, meant to provide the reader a guided tour of the document sections.
- §2 Methodology: The methodological approach adopted for the design of the first version of the RE4DY Reference Architecture is explained in this section, along with the consideration of background knowledge and the outlining of initial ways of exploitation. It should be noted that this section presents a summary of the main takeaways from existing reference architectures and is not intended to provide an exhaustive dissertation on the topic.
- §3 RE4DY RA: This is the core section of the deliverable as it depicts the RE4DY Reference Architecture (RA) providing details about all the building blocks that form its layers as well as about its vertical dimensions.
- §4 Conclusions: This is the closing section that outlines the conclusions and next steps.
- §5 References: Section with the specific references to citations made in the document.

## 1.2 Relationships among WP2 tasks

The figure below illustrates how the task T2.3, which focuses on technologies, plays a central role in this work package as it develops the RE4DY Digital Continuum Reference Architecture. In fact, the design of this architecture is built on synergy with the objectives pursued by all the WP2 tasks. Firstly, with the ones managing the Trial Handbook (T2.1, T2.5), then with those dedicated to defining the frameworks aimed at ensuring resiliency (T2.2) and finally, with those aimed at building trust (T2.4) between the final users of the solutions being implemented adopting this architecture, through proper rules and methods.

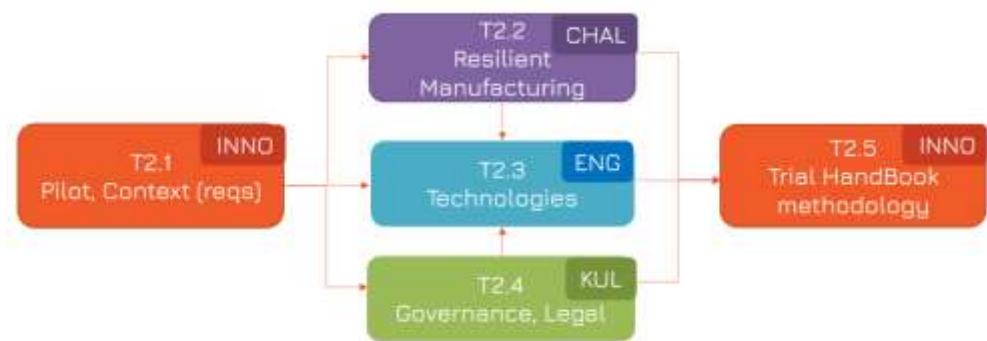


Figure 1: WP2 internal dependencies



## 2 Methodology

### 2.1 Methodological approach



*Figure 2: The methodological approach in six-stages*

The RE4DY RA was developed using an approach in six stages, as shown in the above picture. Initially, a comprehensive investigation was conducted to examine existing architectures that focus on big data technologies and the manufacturing sector. By analyzing various blueprints, as showed in §2.2, it became evident that there was a need for a new architecture to bridge the gap between these established starting points and the specific requirements and aspirations of the project. Furthermore, valuable insights and key results from the previous project BOOST4.0, which involved successful implementations of a related architecture in industrial settings, were also considered as a foundation for the new architecture.

At the same time, efforts were initiated in synergy with WP3 to gather a range of software assets that already incorporate certain essential functionalities. These existing assets, that serve as a starting point for developing the required functionalities in the WP3 tasks, were as well instrumental in determining the specific functional building blocks to be included in the new architecture.

As part of an ongoing iterative process following agile best practices, an initial iteration for consolidating the RE4DY Reference Architecture was started. This first version is subject to continuous improvement along the project and can be updated based on feedback from the development team and requirements identified within the project trials, which serve as a testing ground for the architectural decisions made.



In order to ensure proper uptake and exploitation of this Reference Architecture, it has been finally mapped to an established model from the Digital Factory Alliance (DFA)<sup>1</sup>, as a first step on leveraging its initiatives and long-lasting impact creation potential.

## 2.2 Knowledgeable on SoTA

The investigation conducted to examine existing architectures that focus on big data technologies and the manufacturing sector resulted in identifying five blueprints whose main characteristics are summarized in the following dedicated subsections.

### 2.2.1 RAMI 4.0

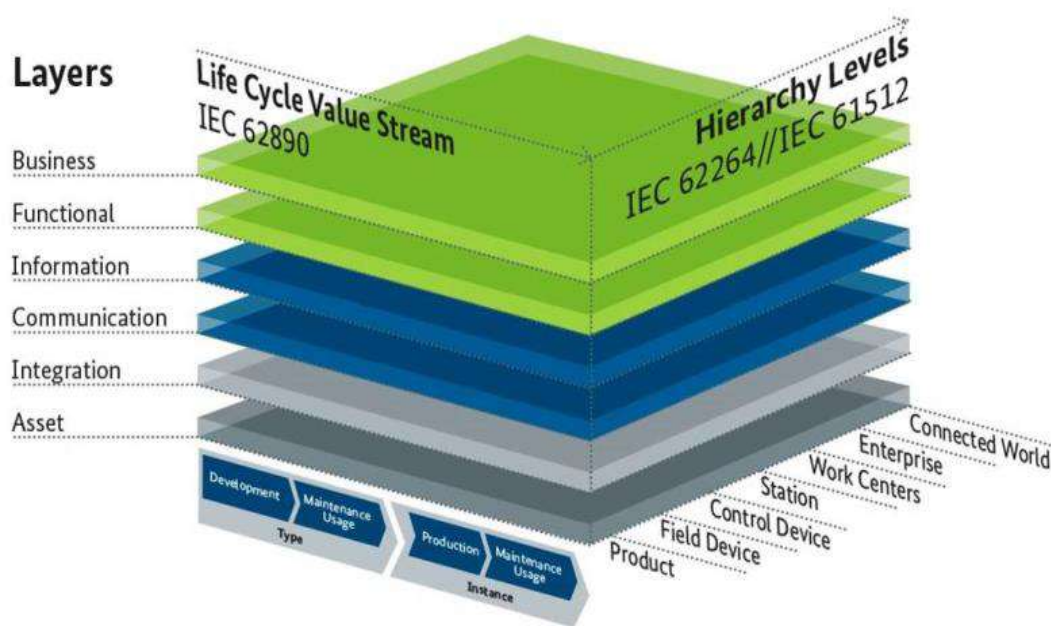


Figure 3: RAMI 4.0

The RAMI 4.0 (Reference Architecture Model Industry 4.0) is a three-dimensional model (ZVEI, 2019) used to represent the Industry 4.0 space. It is based on the Smart Grid Architecture Model (SGAM) and adapted to meet Industry 4.0 requirements. The first dimension (in the vertical axis) of the model consists of six layers representing different perspectives, including data maps, functional descriptions, communication behavior, hardware/assets, and business processes.

The layers of the reference architecture model are as follows:

<sup>1</sup> <https://digitalfactoryalliance.eu> DFA, the Digital Factory Alliance has been established as part of innovative European Commission projects that focus on revitalizing and digitizing future factory assets. Its core belief is that these initiatives will significantly impact the operational and managerial aspects of factories in the coming years. The alliance aims to drive the adoption of Artificial Intelligence Technologies and Data Intelligence to achieve Zero X Manufacturing Environments, emphasizing the importance of advanced technologies and data-driven decision-making in manufacturing.



1. Business Layer: ensures the integrity of functions in the value stream, maps business models and overall processes, and handles legal and regulatory frameworks.
2. Functional Layer: provides a formal description of functions, supports business processes, and executes applications and technical functionality.
3. Information Layer: processes events, ensures data integrity, integrates different data sources, and provides structured data through service interfaces.
4. Communication Layer: standardizes communication and data transmission formats between layers and provides control services for the Integration Layer.
5. Integration Layer: provides information on assets in a computer-readable form, controls the technical process, and interacts with humans through interfaces like the Human Machine Interface (HMI).
6. Asset Layer: represents physical components, documents, and human beings connected to the virtual world through the Integration Layer.

The second dimension (along the left-hand horizontal axis) addresses the Life Cycle and the associated Value Streams that it contains, as aspects of Industry 4.0.

With regards to Life Cycle, the reference architecture model incorporates the life cycle of products, machines, and factories. It distinguishes between the creation of a type (during development and testing) and instances (manufactured products based on the type).

About Value Streams, the digitization and linking of value streams provide significant improvement potential in Industry 4.0. It involves integrating logistics data, order planning, assembly, logistics, maintenance, and customer and supplier information to optimize processes.

The third dimension (right-hand horizontal) contains the hierarchy levels in the reference architecture model, which follow the IEC 62264<sup>2</sup> and IEC 61512<sup>3</sup> standards for classification within a factory. These levels include Enterprise, Work Unit, Station, Control Device, Field Device, and Product. The model also considers observations beyond the enterprise level, including collaboration with external engineering firms, suppliers, and customers, represented as the "Connected World."

<sup>2</sup> ISO/IEC 62264-1:2013 Enterprise-control system integration — Part 1: Models and terminology

<sup>3</sup> IEC 61512-1:1997 - Batch control - Part 1: Models and terminology



### 2.2.2 NIST Big Data Reference Architecture

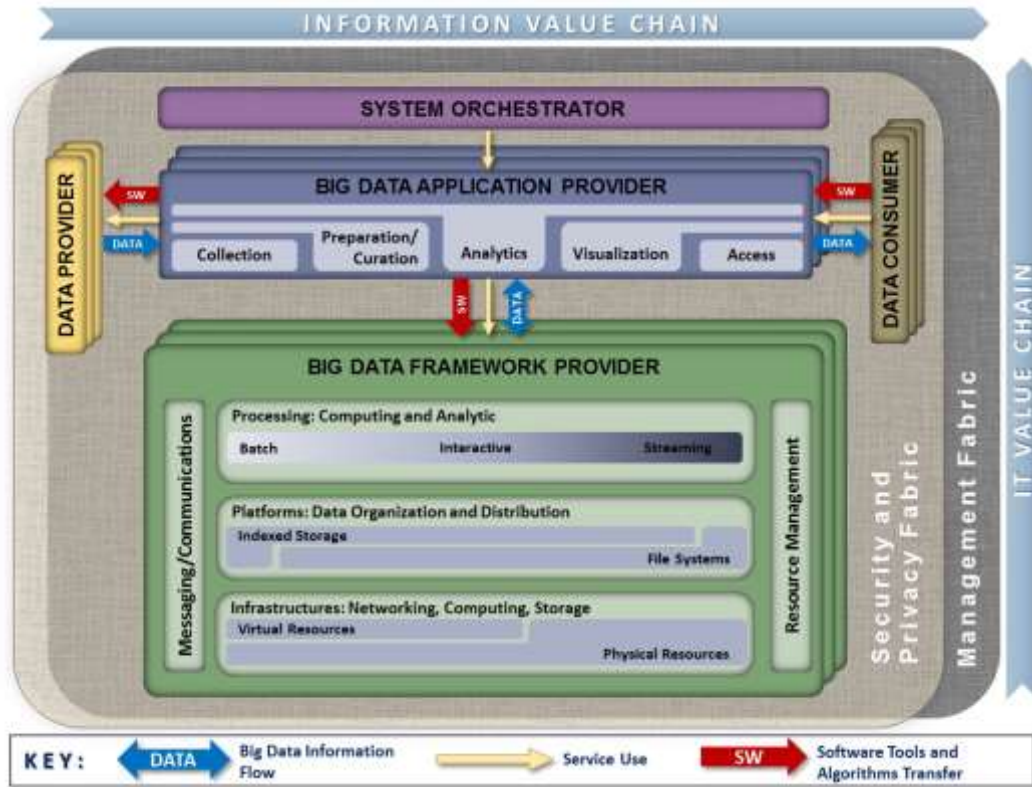


Figure 4: NIST Big Data Reference Architecture (NSBDRA)

The NBD-PWG Reference Architecture Subgroup developed a vendor-neutral conceptual model called the NBDRA (NIST Big Data Reference Architecture) (NIST, 2019) which enables system engineers, data scientists, and decision-makers to develop solutions for diverse Big Data issues within an interoperable ecosystem. It provides a framework applicable to various business environments, from tightly integrated enterprise systems to loosely coupled industries.

The NBDRA represents a Big Data system with five logical functional components connected by interoperability interfaces; besides these components, two fabric roles, the Management and Security and Privacy, envelop the components to highlight their interwoven nature. The architecture is thus organized around five major roles: System Orchestrator, Data Provider, Big Data Application Provider, Big Data Framework Provider, and Data Consumer. Management and Security and Privacy fabrics provide crucial services and functionality to the main roles.

Multiple instances of roles can exist in a Big Data implementation, and there can be different frameworks for different applications within the same architecture.

Data can flow between the roles physically or by reference, and software tools are transferred for processing Big Data.

The architecture supports stacking or chaining of Big Data systems, where a Data Consumer of one system can serve as a Data Provider to another system.



Overall, this architectural model serves as a framework to understand and develop Big Data solutions by emphasizing interoperability, role-based components, and the importance of management, security, and privacy.

In the followings are briefly described roles and fabrics of this architectural schema represented by the picture above:

1. System Orchestrator: is responsible for configuring and managing the components of the Big Data architecture to execute specific workloads. It may involve assigning framework components to nodes or creating workflows for multiple applications. The System Orchestrator also monitors workloads and system performance, adjusting on workload requirement changes.
2. Data Provider: introduces new data feeds into the Big Data system for discovery, access, and transformation. It can collect, persist, and scrub data, create metadata, enforce access rights, and provide interfaces for data access. Data Provider interfaces allow applications to locate and access data sources.
3. Big Data Application Provider: executes operations along the data life cycle, including collection, preparation, analytics, visualization, and access. It encapsulates the business logic and functionality required by the System Orchestrator and follows security and privacy requirements. The activities of the Big Data Application Provider can be implemented by different stakeholders as stand-alone services.
4. Big Data Framework Provider: consists of infrastructure frameworks, data platform frameworks, and processing frameworks. These components form the underlying technology stack that supports the Big Data Application Provider. They can be hierarchically organized and may involve different technologies depending on the system requirements.
5. Data Consumer: can be an end user or another system that interacts with the Big Data system. Data Consumers search, retrieve, download, analyse, report, visualize, and use the data for their own processes. They utilize interfaces provided by the Big Data Application Provider to access the information of interest.

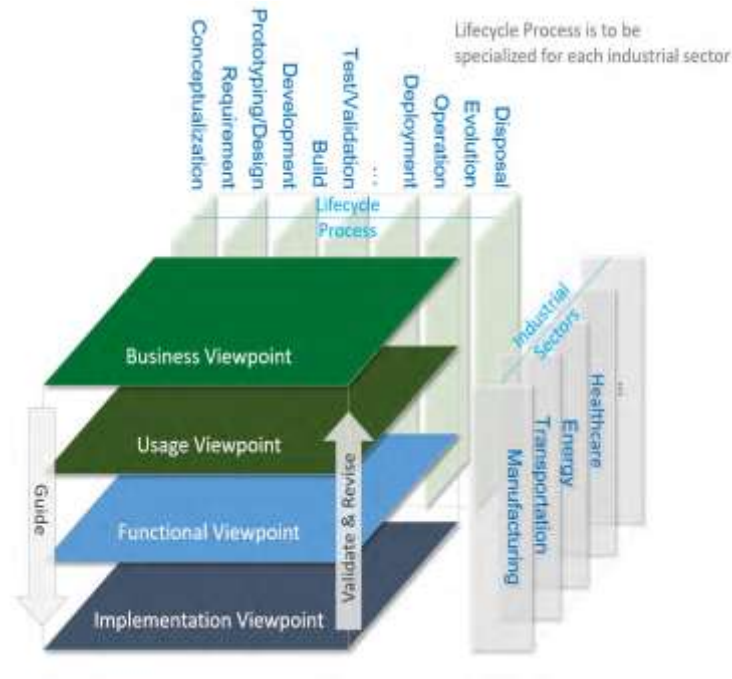
The Management Fabric handles system management and Big Data Lifecycle Management (BDLM). It includes activities such as provisioning, configuration, package management, backup management, and performance management. BDLM encompasses the data life cycle, including collection, preparation, analytics, visualization, and access.

The Security and Privacy Fabric is a fundamental aspect of the Big Data system and affects all components. It ensures the security and privacy of the data and the system as a whole. The Security and Privacy Fabric includes mechanisms, protocols, and agreements to protect data and establish secure interactions between components.





### 2.2.3 IIRA Industrial Internet RA



*Figure 5: IIRA viewpoints and relationships with Lifecycle Process and Application Scope*

The Industrial Internet Reference Architecture (IIRA), published (IIC, 2022) by the Industrial Internet Consortium (IIC), is an open architecture designed to guide the entire lifecycle of an IIoT system, from conception to design and implementation. It incorporates architectural concepts, vocabulary, structures, and patterns, drawing from the ISO/IEC/IEEE 42010-2011 standard. This standard emphasizes stakeholder perspectives and system properties, considering aspects such as implementation, risks, and maintainability. The IIRA aims to address common concerns in IIoT across industries and offers an architectural template and methodology for engineers to analyse and resolve design issues. It aims as well to drive interoperability, guide technology development, and maximize its value across various Industrial Sectors. While it does not describe a specific system lifecycle process, it serves as a framework and methodology for system conceptualization, emphasizing key concerns that impact the overall lifecycle process. By using viewpoints and considering stakeholders, it helps identify important system functions and design gaps. It employs four viewpoints to capture interactions and focuses on software capabilities and business processes, which are:

1. Business viewpoint, which focuses on stakeholders, their vision, and objectives within the business and regulatory context. It addresses how the system achieves its objectives by mapping them to fundamental capabilities. This viewpoint is essential for business decision-makers, product managers, and system engineers involved in IIoT implementation.
2. Usage viewpoint, which focuses on expected system usage and functionality. It involves sequences of activities performed by human or logical users to achieve the system fundamental capabilities. This viewpoint is important for system engineers, product managers, and other stakeholders involved in specifying and representing the users of the IIoT system.



3. Functional viewpoint, which examines the components, interfaces, and interactions within the system, as well as its connections to external elements. It supports the system activities and usages. This viewpoint is relevant to system architects, developers, and integrators involved in designing and integrating the system's functional components.
4. Implementation viewpoint, which focuses on the technologies, communication schemes, and lifecycle procedures required to implement the functional components identified in the functional viewpoint. It supports the system capabilities outlined in the business viewpoint and is of interest to system architects, developers, integrators, and operators involved in the system's design and operation.

The functional viewpoint holds the utmost relevance when outlining a conceptual architecture with the appropriate functional building blocks. Therefore, it is valuable to summarize here its key concepts as follows. This viewpoint focuses on the functional capabilities and structure of the system and its components. Functional domains are introduced as distinct functionalities that decompose the overall system, providing building blocks applicable across various industries. The system is decomposed into five functional domains: Control, Operations, Information, Application, and Business.

Data and control flows occur within and between these domains, as depicted by the arrows in the picture below, with interactions increasing in scope and granularity as they move up the functional hierarchy. As the information becomes broader and richer, the need of new intelligence arises.

In a nutshell, each of these five domains is responsible for:

1. The Control and monitor domain is responsible for implementing industrial control systems and consists of functions such as sensing, actuation, and communication. Components in the control domain may be geographically distributed and may require special security considerations. The control domain functions are abstract and aim to establish connectivity, gather data, and provide optimization feedback to control systems.
2. The Information domain manages and processes data. As such, it gathers data from various domains, transforms and analyses the data to acquire system-wide intelligence. Functions in the information domain aid decision-making, optimize system-wide operations, and improve system models. Components in the information domain may or may not be co-located with control domain components.
3. The Application domain implements the application logic; hence it applies high-level application logic, rules, and models for optimization in a global scope. Functions in this domain do not maintain low-level operations and delegate them to the control domain. The application domain includes functions for logics and rules, as well as APIs and user interfaces.
4. The Business domain implements business functional logic and integrates business functions such as ERP, CRM, PLM, MES, HRM, asset management, and more. It also enables end-to-end operations of IIoT systems.
5. System management domain manages IIoT systems which are inherently complex, consisting of numerous loosely coupled and distributed functional components. This type of management can be further divided into two categories: Orchestration and Lifecycle functions.



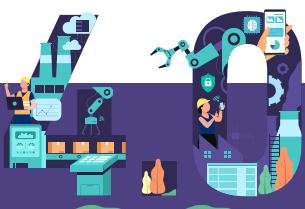


- o The Lifecycle functions encompass the essential tasks associated with the complete software-system lifecycle, including deployment, configuration, monitoring, diagnosis, and updates of the IIoT system and its sub-systems.
- o Orchestration functions are responsible for coordinating the operation of the various subcomponents within an IIoT system. These components are typically loosely coupled and distributed, and the Orchestration functions facilitate their interactions with external systems.

Furthermore,

- The Crosscutting functions are necessary for enabling the major system functions and are shared across multiple system components. For example, connectivity and data management are identified as crosscutting functions in an IIoT system. Connectivity allows system functions to interact with each other, while data management facilitates analytics on the gathered data from industrial assets.
- System characteristics are emergent behaviours or properties that result from the interactions of the system constituent parts; they include security, safety, resilience, reliability, and privacy. The trustworthiness of a system is determined by the implementation of these characteristics and how securely the functional components are integrated and interact with each other. The realization of system characteristics can sometimes conflict with each other or impose constraints on one another. For example, ensuring safety requires ensuring security as well, and inadequate security measures can compromise safety; therefore, it is necessary to balance these system characteristics to achieve the desired levels for each. The IIRA places equal importance on both the functions required to support the system business purpose and the system characteristics necessary for proper performance and integrity. This balanced approach ensures that the functions are performed correctly, and the overall business purpose of the system is not compromised.

The relationship between the functional domains, and crosscutting functions and key system characteristics are summarized in the following picture.



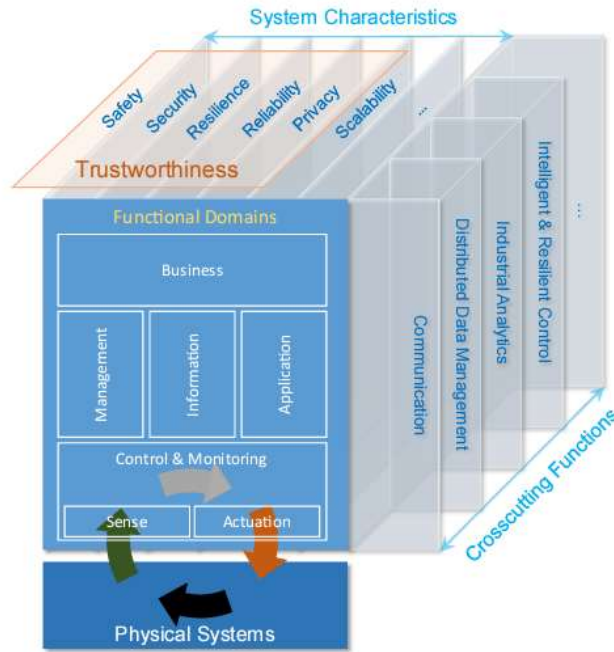


Figure 6: Functional domains, Crosscutting Functions and System Characteristics

The IIRA will be continuously refined based on feedback from testbeds and real-world deployments. It also identifies technology gaps, which encourages the development of new technologies by the industrial internet community.

### 2.2.4 IDS RAM

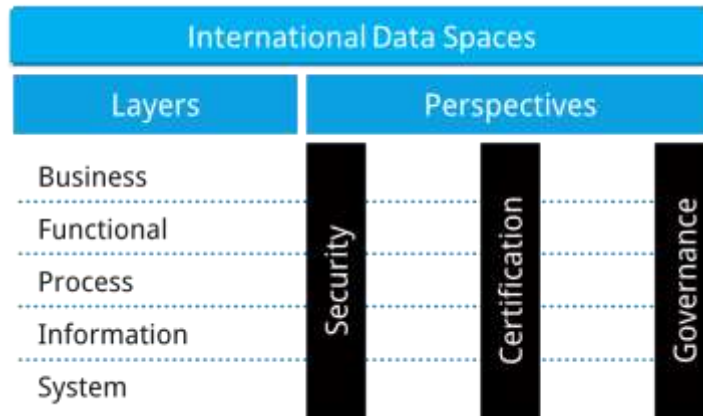


Figure 7: IDS General structure of RA Model

The IDS<sup>4</sup> Reference Architecture Model (IDS-RAM) by (IDSA, IDS-RAM 4.0) is structured using multiple layers, namely the business layer, functional layer, process layer, information layer, and system layer. These layers are interconnected by transversal functionalities that promote security, certification, and governance, as depicted in Figure 7.

<sup>4</sup> <https://internationaldataspaces.org/> The International Data Spaces are meant as a secure, sovereign system of data sharing in which all participants can realize the full value of their data.



- The Business layer classifies and categorizes the various roles that participants within IDS can undertake. It also outlines the key activities and interactions associated with each role.
- The Functional layer defines the functional requirements of IDS and specifies the specific features that emerge from these requirements.
- The Process layer describes the interactions occurring between different components within IDS. Utilizing the BPMN notation, it offers a dynamic view of the Reference Architecture Model.
- The Information layer establishes a conceptual model that utilizes linked-data principles to describe both the static and dynamic aspects of IDS' constituents.
- Lastly, the System layer addresses the decomposition of logical software components, encompassing integration, configuration, deployment, and extensibility considerations for these components.

Expanding the scope beyond individual use cases and embracing interoperability and a platform-oriented perspective, the IDS Reference Architecture Model serves as a framework that connects diverse cloud platforms by employing policies and mechanisms to ensure secure data exchange and trusted data sharing, all while upholding the principle of data sovereignty.

Through the IDS Connector, various entities such as industrial data clouds, enterprise clouds, on-premise applications, and interconnected devices can seamlessly join the International Data Space ecosystem as shown in several sections dedicated to RA building blocks such as the one about data sovereignty (§3.3.2), data cleansing (§3.3.5) as well as by the vertical dimension focused on Federated Learning and Analytics (§3.8).

### 2.2.5 BDV RM

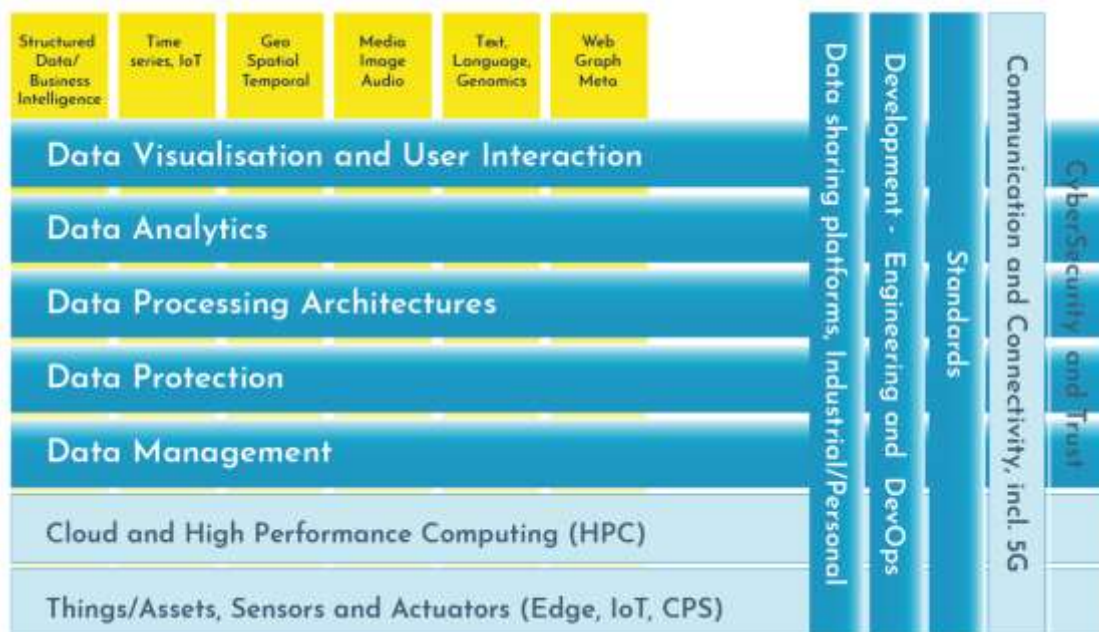


Figure 8: BDV Reference Model



The BDV<sup>5</sup> Reference Model (BDVA, 2017), developed by the BDVA with input from experts and stakeholders in the Big Data Value chain, serves as a common framework to locate Big Data technologies within the IT stack. It addresses key concerns and considerations for Big Data Value systems through the horizontal and vertical concerns of its structure.

Horizontal concerns encompass specific aspects of the data processing chain, including data collection, ingestion, and visualization. It is important to note that the horizontal concerns do not imply a layered architecture. For example, data visualization can be directly applied to collected data without the need for data processing and analytics.

Vertical concerns deal with cross-cutting issues that impact all horizontal concerns and may involve non-technical aspects as well. The BDV Reference Model does not aim to be a technical reference structure but is compatible with other reference architectures, such as the ISO JTC1 WG9 Big Data Reference Architecture.

The BDVA reference model offers a detailed and inclusive depiction of the intersecting concerns between Big Data and cloud platforms. BDVA examines existing gaps and challenges related to dynamic data and presents a list of essential advancements that need to be made. However, since the intended audience includes non-technical experts, technical definitions are kept at minimum, focusing instead on highlighting relevant initiatives and organizations. Additionally, discussions around interoperability, security, and composition are only briefly touched upon.

---

<sup>5</sup> <https://www.bdva.eu/> The Big Data Value Association (BDVA) is an industry-driven organisation with a mission to develop an innovation ecosystem that enables the data-driven digital transformation of the economy and society in Europe.



## 2.3 Main motivations for a new architecture

Once analyzed the state of the art, it became apparent that a novel architectural framework was indispensable to bridge the disparity existing between these existing foundations and the distinct project requirements and ambitions. This disparity can be argued by the following 4 motivations:

- Need to reinforce decentralization on infrastructures, services, and data planes.

In the context of federation and distributed systems, there is a recognized necessity to strengthen decentralization across infrastructures, services, and data planes. Decentralization refers to the distribution of control, authority, and decision-making to multiple entities or nodes in a network, rather than relying on a centralized authority; basically, it is a principle that promotes resilience, scalability, and autonomy within a federated environment.

Reinforcing decentralization on infrastructures, services, and data planes not only enhances fault tolerance and robustness, but also facilitates collaboration and interoperability among federated entities, e.g., exchanging data as a valuable product.

In terms of infrastructures, decentralization involves the deployment and management of resources across multiple physical or virtual locations. This distribution of infrastructure components enables better resource utilization, load balancing, and fault isolation, thereby increasing the overall efficiency and reliability of the system.

Similarly, decentralizing services entails breaking down monolithic applications into smaller, more modular services that can be deployed and scaled independently. This approach allows for greater flexibility, adaptability, and resilience, as each service can be developed and maintained by separate teams or organizations.

Furthermore, decentralization of the data plane involves distributing data storage and processing capabilities across multiple nodes or edge devices. This not only reduces the reliance on a central data repository but also improves data availability, responsiveness as it enables data to be processed and analyzed closer to the source, namely at edge level, minimizing latency.

This need is in line with the objectives pursued by initiatives like Gaia-X<sup>6</sup> and Manufacturing-X<sup>7</sup>, that this novel architectural framework aims to incorporate to ensure preparedness and alignment with the emerging trends. Regarding this matter, the objective of Gaia-X is indeed not to create a singular cloud infrastructure, but rather to establish a federated system that seamlessly connects multiple cloud service providers and users with which to fuel the growth of the European data economy.

Similarly, with a specific focus on the manufacturing sector, Manufacturing-X aims to be a cross-sectoral industrial policy initiative that fosters the development and establishment of a decentralized data economy for both German and European industries.

<sup>6</sup> <https://gaia-x.eu/what-is-gaia-x/about-gaia-x/>

<sup>7</sup> <https://www.plattform-i40.de/IP/Navigation/EN/Manufacturing-X/Manufacturing-X.html>



- Introduce Digital Continuity for computing, networking, and deployment for a seamless exploitation of the digital thread, regardless of where data and applications are.

Digital continuity is the uninterrupted and consistent availability, accessibility, and integrity of digital information and processes throughout the entire lifecycle of products, services, or operations. It involves the seamless integration of digital technologies, systems, and data across different stages, such as design, manufacturing, distribution, and maintenance, to ensure the reliable and efficient flow of information, knowledge, and resources.

By enabling seamless connectivity and integration across various stages of the manufacturing process, from design and prototyping to production and quality control, the real-time data exchange is facilitated, but not only. In fact, the collection and analysis of vast amounts of data generated throughout the manufacturing lifecycle is also enabled. By leveraging advanced analytics and machine learning algorithms on these data, transparently about where they are located, manufacturers can gain valuable insights into their operations, identify patterns, and make data-driven decisions to optimize processes, reduce defects, and enhance overall product quality.

Digital continuity allows different stakeholders, including designers, engineers, suppliers, and customers, to access and share relevant information seamlessly. This fosters closer collaboration, enables concurrent engineering practices, and facilitates faster decision-making, ultimately leading to shorter product development cycles and increased innovation.

- Not only data-driven, but fully exploiting the concept of DaaP, as a marketable digitisation of the value chain (data space).

Through the Data as a Product (DaaP) business model data itself becomes a valuable asset that can be packaged, marketed, and sold to generate revenue. In fact, as organisations increasingly collect and analyse vast amounts of data throughout their operations, they start to recognize the potential value that this data holds beyond its immediate usage. This potential value can be, for example, data sets, insights, predictive analytics models as well as data-driven consulting services that they can provide to other companies within their industry or even to external parties. The digitization of the value chain plays a crucial role in enabling the DaaP business model since as more processes, transactions, and interactions become digitized, they generate valuable data that can be captured, analysed, and monetized.

Implementing this business model requires organizations, besides the proper data analytics technologies, to have robust data governance and privacy frameworks put in place. In fact, they must ensure that data is collected, stored, and shared in compliance with applicable regulations and industry standards. In brief, they need toolkits and framework capable to build and assure trust in data ownership and sharing. In this regard, it is worth noting the conceptual model for a common vision on data spaces recently published by (DSBA, 2023) which depicted a converged view of this concept with the aim of achieving interoperability and portability of solutions across data spaces, by harmonizing technology components and other elements. According to this view, a data product is created by combining services and/or resources that are provided and activated for a specific customer; resources are usually necessary for executing the services. Providers of data products within data spaces should have the capability to offer data services at



clearly defined endpoints, considering that customers, who are initially unknown to them, will be able to consume their data services through those endpoints. On the other hand, customers of data products need to understand how to consume data services available at discovered endpoints. This implies that all participants in data spaces should have a common understanding or "speak the same language". A concrete Data Product gets created (provisioned and activated) when a Data Product Offering made available by a provider is acquired by a customer through the issuance of Product Order.

- Natively creating convergence for Manufacturing and IT operations, integrating toolkits covering the whole lifecycle of an Industrial Data Platform from the realization (design and development) to the commissioning (integration and validation) and to the operation (management and maintenance) of all the SW artefacts. This convergence would enable streamlined processes and data flow between manufacturing and IT operations that enhances efficiency by eliminating data silos, reducing manual interventions, and optimizing the overall workflow. Ensuring that data from manufacturing processes can be effectively processed, shared, and utilized by IT systems is a key challenge. It involves addressing data format compatibility, data mapping, and data transformation issues to enable smooth integration and analysis across the entire lifecycle. Furthermore, many manufacturing environments still rely on legacy systems that may lack of standardization or modern connectivity options.

## 2.4 Built upon key results from the past

BOOST 4.0 has been considered as solid starting point, as part of its layers and vertical dimensions match with the first needs expressed at very early stage of the project and depicted through the onion shape shown in *Figure 9*. On the other hand, main motivations can be summarized as follows:

- It already extends several standard RAs such as RAMI4.0, NIST and IIRA.
- It is aligned with further models such as the IDS and BDV ones.
- It bridges Big Data domain with manufacturing.
- Several instantiations of this RA were tested in at least ten large industry trials.





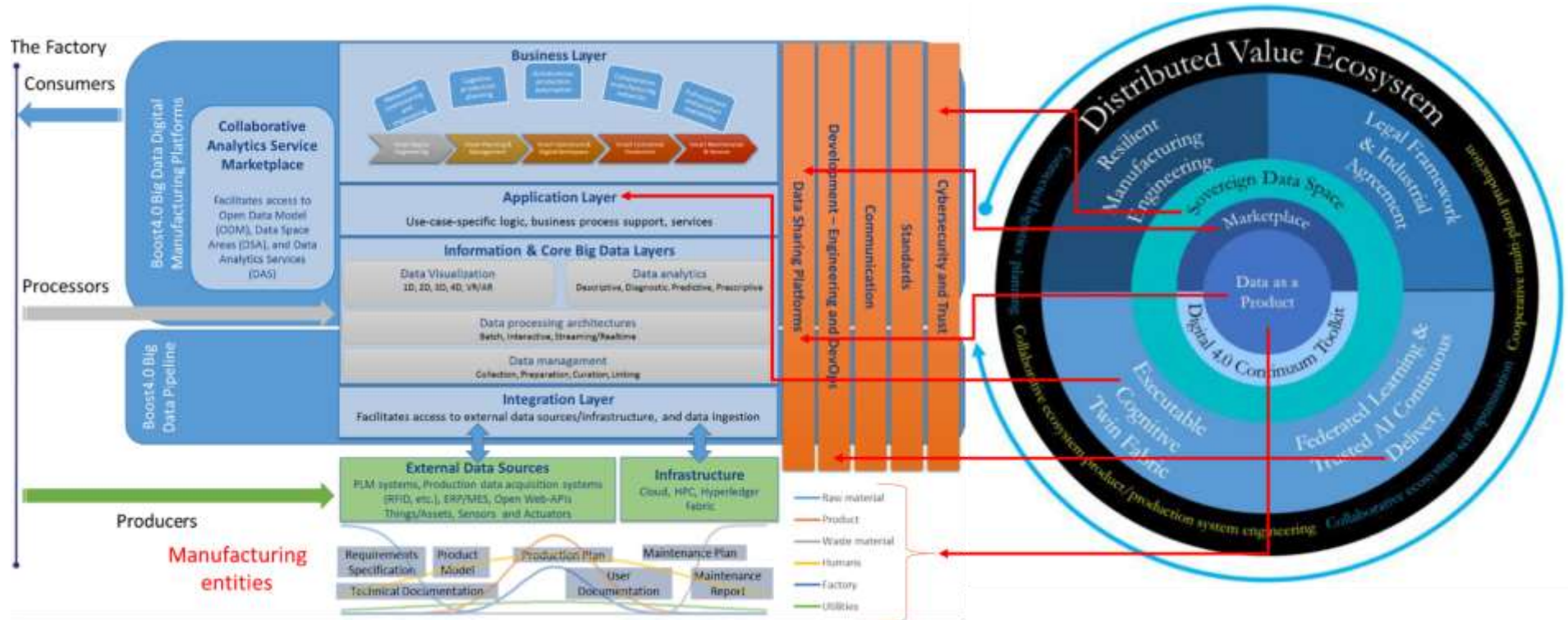
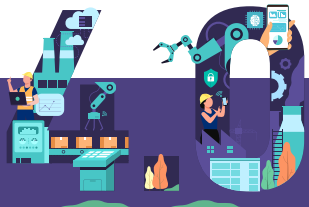


Figure 9: BOOST 4.0 mapping to RE4DY main needs





## 2.5 Bottom-up synergies

Concurrently with the examination of various blueprints and key results from the past, a bottom-up strategy has been employed in synergy with Work Package 3 (WP3), which primarily concentrates on technological advancements. Consequently, an assessment of the existing software assets applicable to its diverse tasks has been initiated as a foundation for proposing essential functionalities and, subsequently, the essential building blocks for the design of the RE4DY RA. The following screenshot provides an illustration of how each asset has been described, showcasing one of its aspects (as outlined in D3.1). At WP2 side special attention has been placed on the features section, which provides insights into the capabilities of the assets and their potential as a source for architectural functionalities.

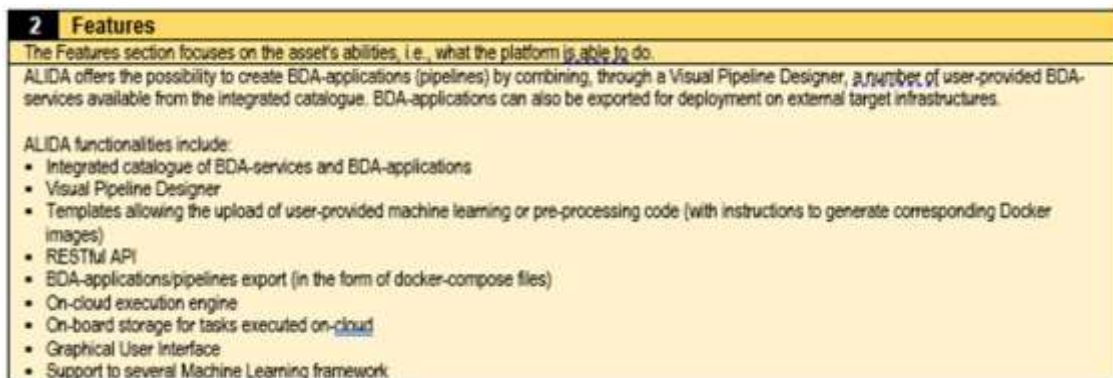


Figure 10: Example of ALIDA asset features collection

## 2.6 RE4DY RA v1 design

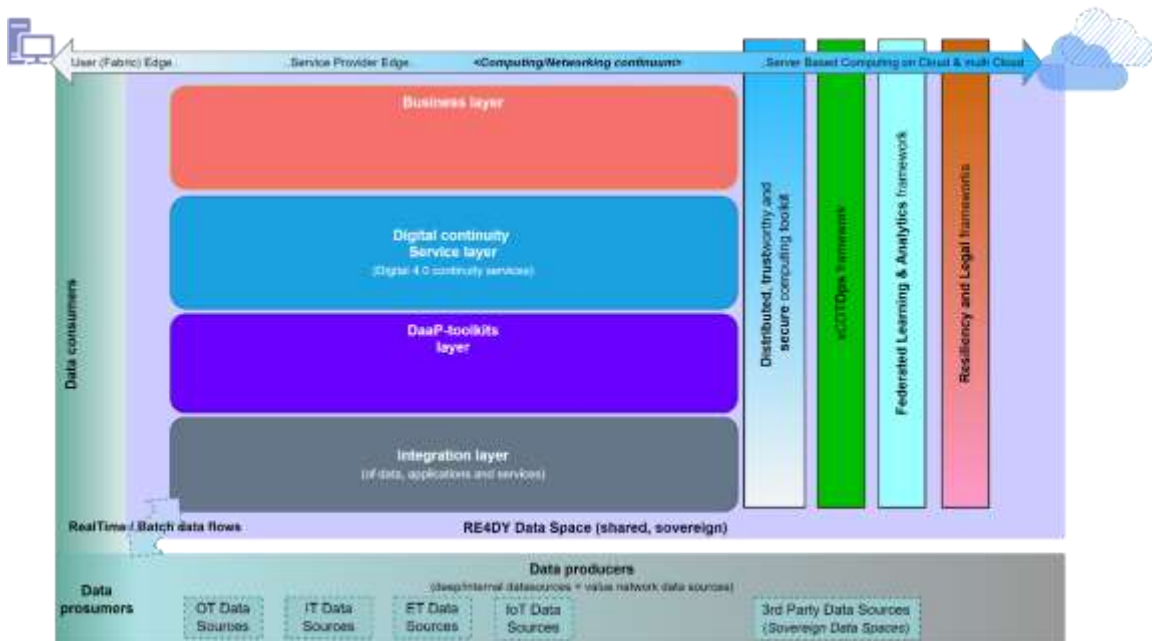


Figure 11: RE4DY Reference Architecture (RA)

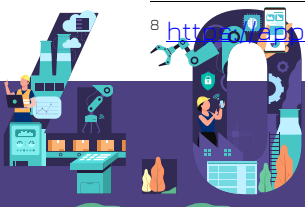


This phase encompasses the architectural design process, during which all the diagrams were created using the draw.io<sup>8</sup> tool that is an open-source technology stack renowned for constructing diagramming applications. The architectural design, being the focal point of this deliverable, is described in the higher-level section §3.

## 2.7 Designed to uptake and exploitation

The RE4DY RA has been aligned with a well-established model from the DFA. As introduced in §3.1, the DFA (DFA, 2021) aims at building and nurturing a community where factories, manufacturing and digital industries can see and can be seen, where they get to know and are known. It recognizes that nowadays it is still difficult for manufacturing industry to find common frameworks that support them in identifying the value added by innovations and digital transformation in their specific contexts and value chains. For this reason, it is crucial for individual pilots to embrace common models and implement pilot 4.0 practices to yield collective advantages in the medium and long run. These include improved synchronization between digital infrastructures and digital manufacturing platforms, enhanced cost-effectiveness in system and factory integration, and significant economies of scale in the implementation of manufacturing 4.0 processes powered by big data and AI technologies. Given this perspective, it is evident that maintaining alignment between RE4DY RA and the DFA model, as illustrated by *Figure 12*, is crucial. This mapping is in line with ensuring the effective adoption and utilization of the Reference Architecture, while facilitating its widespread implementation and exploitation. In fact, members of this Alliance gain access to cutting-edge knowledge, emerging trends, and readily deployable products in the digital manufacturing domain. This exposure opens doors to a growing marketplace centered around Zero X Manufacturing, providing brand recognition and access to new business prospects.

<sup>8</sup> <https://draw.io>



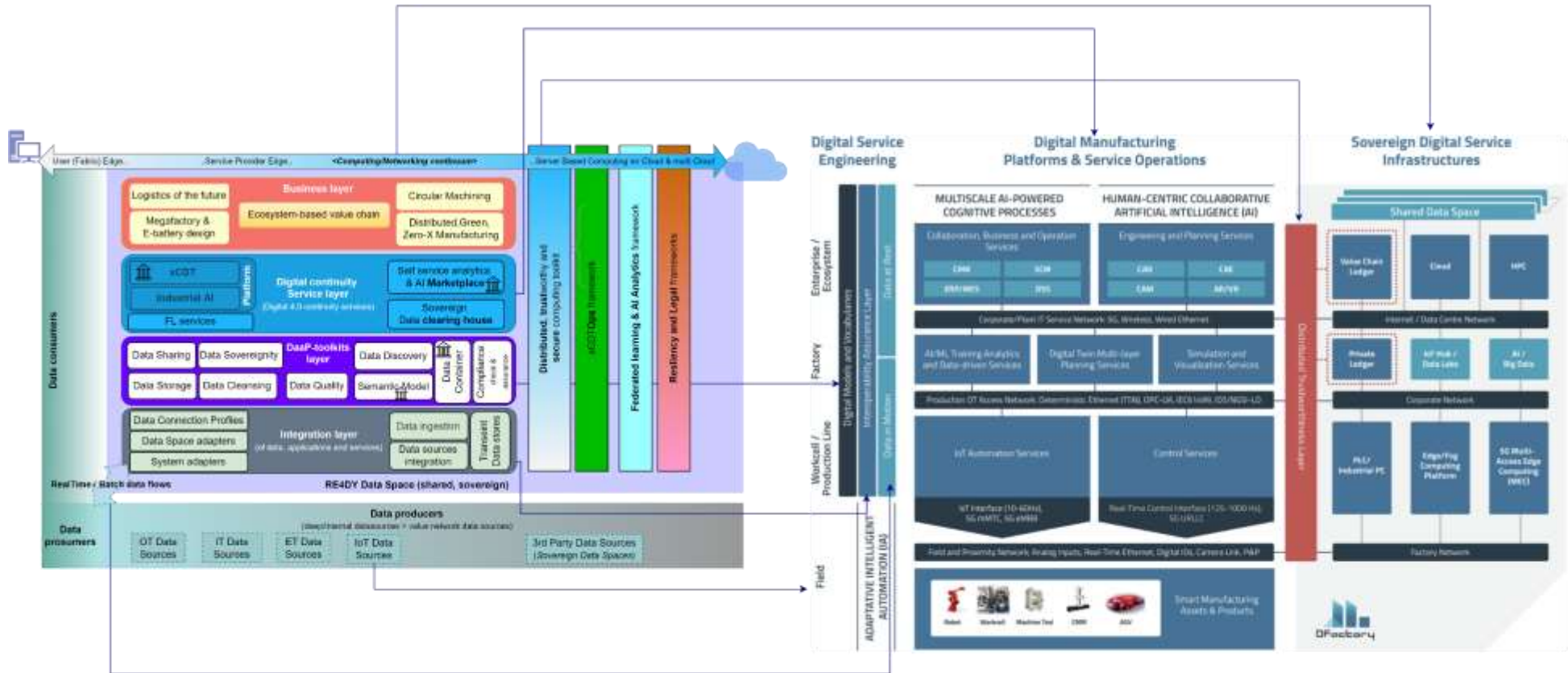


Figure 12: RE4DY RA mapping to DFA-Digital Service Integration Reference Architecture



# 3 RE4DY Reference Architecture

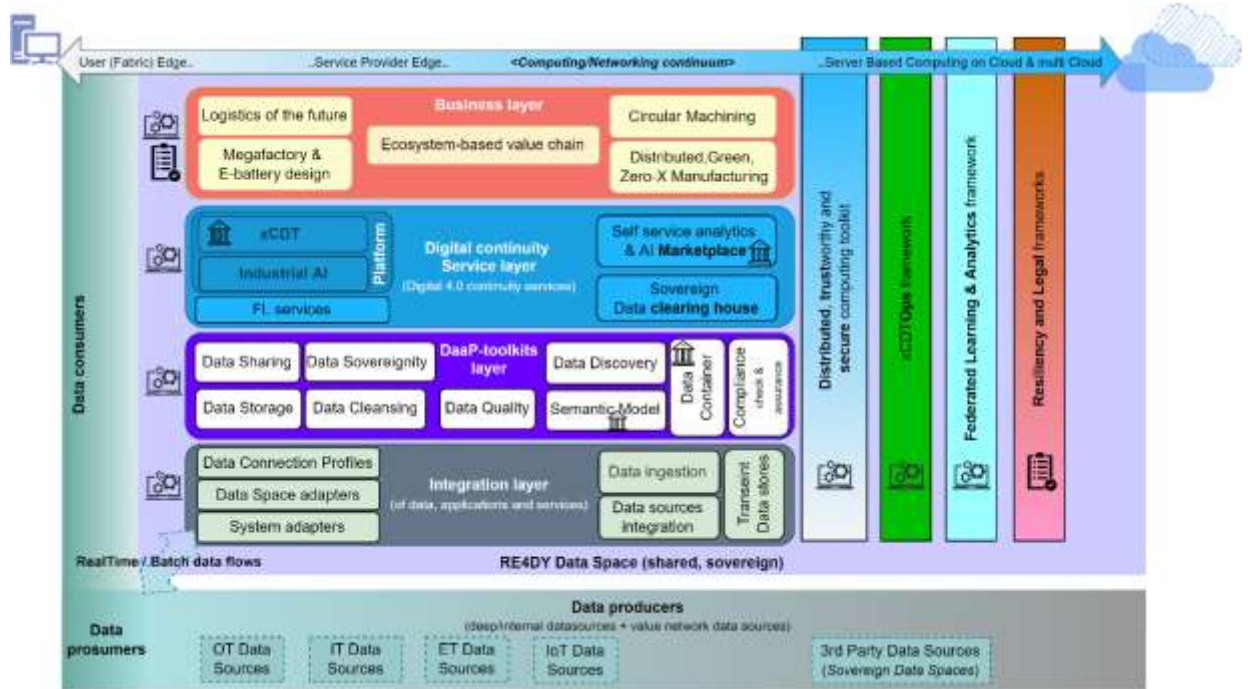
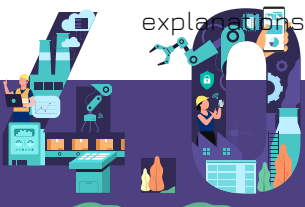


Figure 13: RE4DY RA building blocks

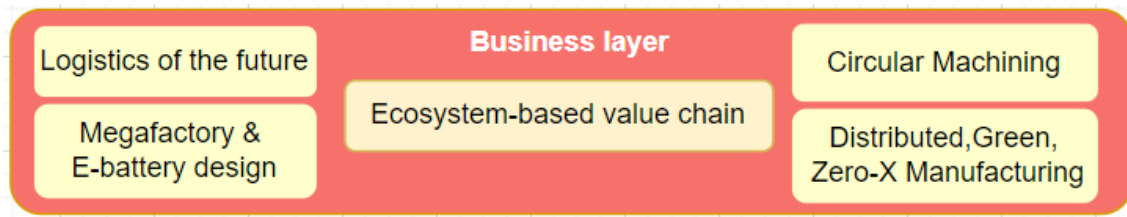
	SW components (tools, toolkits, services)
	Guidelines (set of rules, laws, best practices, etc.)
	RE4DY pillars

The architecture consists of four layers and as many vertical dimensions, with an additional double-headed arrow emphasizing the importance of Digital Continuum in computing and networking. This ensures a seamless utilization of the digital thread, irrespective of the location of data and applications.

All these dimensions work together to establish a conceptual model for creating a shared, sovereign data space, represented by the purple area encompassing these dimensions. The primary participants engaged in this model are the data consumers and producers, commonly referred to as "prosumers" when they assume both roles. Every layer consists of distinct building blocks that represent the provided functionality, as depicted in the dedicated sub-sections. Certain blocks are marked with a "temple" icon, signifying their importance as core functionalities within the project scope. The layers and vertical dimensions are distinguished by specific icons that indicate whether they primarily consist of software components, as is often the case, intrinsically, or if they provide guidelines such as rules, laws, and best practices. The vertical dimensions, similar to the layers, are described in the following dedicated sub-sections, providing detailed explanations for each.



### 3.1 Business layer



*Figure 14: Business layer building blocks*

The RE4DY RA Business Layer inherits several high-level business functionalities from BOOST4.0 RA and applies them to the four specific business cases that have the highest priority in the EU for the development of a common manufacturing data space. These four business cases are:

1. Logistics of the future
2. Megafactory & E-battery design
3. Circular Machining
4. Distributed, Green, Zero-X Manufacturing

The functionalities provided by the Business Layer include:

1. Integrity assurance of the functions in the value stream: This ensures that the functions within the value stream are reliable and maintain data integrity throughout the manufacturing process.
2. Mapping out business models and overall processes: The Business Layer defines and maps out the business models and the resulting overall processes for each of the four business cases.
3. Providing a link between different business processes: It establishes connections and links between different business processes, facilitating information flow and collaboration between various aspects of the manufacturing operations.
4. Receiving events for advancing business processes: The Business Layer receives events and triggers related to the advancement of business processes, allowing for real-time updates and progress tracking.

The Business Layer combines the functionalities of the layers below it to realize business processes and establish links between different business processes. These processes are designed to be applicable across multiple sectors, offering high growth and replication potential.

It's important to note that the Business Layer focuses on the higher-level business functionalities and does not concern itself with concrete systems like ERP (Enterprise Resource Planning) software. If necessary, the functions of specific systems are integrated through the dedicated Integration Layer. In this context, the RE4DY RA Business Layer aligns with the RAMI4.0 Business Layer, following the reference architecture of RAMI4.0 (Reference Architecture Model Industry 4.0).





### 3.1.1 Ecosystem-based value chain

The ecosystem-based value chain represents the flow of data and information within the manufacturing ecosystem to create value and enable informed decision-making. It encompasses the collection, processing, analysis, and utilization of data across the various stakeholders and processes involved in the manufacturing industry.

The ecosystem-based value chain leverages digital technologies and connectivity to enable seamless data exchange and collaboration between different entities within the ecosystem, through Manufacturing Data Networks (MDN). It involves the integration and interplay of various data sources, such as sensors, machines, systems, and human input, to generate actionable insights and drive improvements in manufacturing operations.

In the context of the RE4DY project, the ecosystem-based value chain refers to the interconnected network of stakeholders and organizations involved in the four specific business cases identified as the highest EU priorities. These business cases, such as logistics of the future, megafactory & e-battery design, circular machining, and distributed, green, zero-X manufacturing, form the core focus of the value chain.

The ecosystem-based value chain aims to unlock the value of data by enabling seamless data flow, collaboration, and insights generation within the manufacturing ecosystem. By leveraging data-driven insights, organizations can optimize operations, enhance productivity, improve product innovation, and deliver enhanced customer experiences.

### 3.1.2 Logistics for the future

"Logistics for the Future" focuses on reimagining and optimizing logistics operations in the manufacturing industry by leveraging cutting-edge technologies, innovative strategies, and sustainable practices. It aims to address the evolving needs and challenges of logistics in an increasingly connected and dynamic business environment.

Particularly in the RE4DY project the focus is on internal logistics, addressing the need to optimize the movement, handling, and storage of materials, components, and finished products within manufacturing facilities or production sites. It aims to streamline internal logistics processes to enhance operational efficiency, reduce costs, and improve overall productivity.

Key objectives of "Logistics for the Future" are:

1. Implement autonomous/automatic planning for logistics processes to reduce the time and effort required for cost-efficient scenario deployment. This involves developing a system configuration that allows logistics planners to validate optimal scenarios on a monthly basis, enabling them to easily deploy changes to the shop floor with a click of a button. The objective is to relieve logistics planners from time-consuming tasks, allowing them to focus on other significant projects within the planning department.
2. Reduce lead time in design iterations and implementation by providing stakeholders with efficient calculation tools. The objective is to speed up the process of designing optimal and cost-efficient solutions for logistics processes, particularly in asset efficiency calculations. By leveraging sophisticated simulation tools, adjustments and scenario changes can be made more efficiently, leading to significant time savings.



3. Increase overall efficiency of processes and assets by optimizing line feeding equipment and processes through the use of machine learning and digital twin systems in simulation environments. The objective is to find optimum solutions that improve efficiency and productivity, resulting in cost reduction and enhanced performance.
4. Improve flexibility by shortening changeover times and streamlining the communication and deployment of adaptations to optimal scenarios in logistics processes. The objective is to enhance process flexibility, particularly in handling unforeseen or out-of-process situations that may impact production and the internal supply chain. Adapting optimally to these vulnerabilities is crucial for establishing resilient logistics processes and reducing costs.
5. Reduce operating costs by optimizing internal stakeholder efforts and improving the efficiency of the internal supply chain on the shop floor. The objective is to minimize costs associated with each service provided by streamlining processes and ensuring smooth operations within the internal supply chain. The more efficient the processes are, the fewer costs are incurred.

By embracing "Logistics for the Future", the manufacturing industry can achieve enhanced operational efficiency, cost savings, sustainability, and improved customer satisfaction. The integration of advanced technologies and innovative strategies in logistics operations helps organizations stay competitive and responsive to changing market demands.

### 3.1.3 Megafactory & E-battery design

The "Megafactory and E-Battery Design" business case within the RE4DY project focuses on the development and optimization of manufacturing processes for electric vehicle (EV) batteries in large-scale production facilities, often referred to as megafactories. This business case recognizes the increasing demand for EVs and the critical role of high-performance and cost-effective batteries in the transition to sustainable transportation.

The objectives of the "Megafactory and E-Battery Design" business case are:

1. Establish a resilient manufacturing engineering reference framework for digital smart products and production value ecosystems in connected factories 4.0, while developing a general model for optimizing individual and value chain industrial process efficiency, sustainability, and reliability.
2. Increase the autonomy and interoperability of big data pipelines, digital threads, and digital twins by integrating open Digital 4.0 continuum toolkits, enabling value networks to treat "Data as a Product" and enhancing the reusability and portability of data and pipelines.
3. Accelerate the implementation of integrated intelligence and active knowledge within holistic cognitive and collaborative connected factories 4.0 Zero X smart manufacturing, aiming to improve efficiency, scale, and trust-building in distributed Industrial Internet value networks' setup and data sharing.
4. Democratize industrial data spaces and cognitive digital twins, maximizing their commercial impact, optimizing upskilling and reskilling needs, and driving adoption in both the manufacturing and digital sectors. This objective involves identifying access to valuable data sets, supporting the development of resilient digital twin processes, and empowering stakeholders to become trusted value network prosumers.



The "Megafactory and E-Battery Design" business case in the RE4DY project aims to accelerate the development and adoption of advanced battery technologies for electric vehicles. By optimizing manufacturing processes, enhancing efficiency, and fostering collaboration, this business case contributes to the widespread adoption of electric vehicles and the transition to sustainable mobility.

### 3.1.4 Circular Machining

The "Circular Machining" business case within the RE4DY project focuses on the implementation of circular economy principles in the machining processes within the manufacturing industry. It aims to reduce waste, optimize resource utilization, and promote the reuse and recycling of materials in machining operations.

The objectives of the "Circular Machining" business case include:

1. Establish an open and transparent data ecosystem: Develop a framework within the Machine Tool sector that enables the creation of an open and transparent data ecosystem. This objective aims to address the challenges of proprietary and scattered data silos by facilitating data connectivity and sharing among service providers and users.
2. Address data security and ownership concerns: Define clear guidelines and protocols to address data security and ownership issues, fostering a conducive environment for data sharing between companies. This objective aims to build trust and encourage data exchange, which is essential for realizing data-based services on industrial assets.
3. Foster the adoption of pay-per-use business models: Promote the adoption of pay-per-use business models in the Machine Tool sector to overcome the risk of under-utilization of machines. This objective aims to create a critical mass for such business models, encouraging stakeholders to maximize the utilization of machine tools and optimize their performance.
4. Enable cross-company, multi-cloud data sharing: Adopt the RE4DY Framework to facilitate cross-company data sharing in multi-cloud environments. This objective seeks to ensure the availability, collation, and sharing of data among all involved actors in the machine tool domain, including OEM machine tool builders, tier 1 tool producers, and end-users.
5. Enhance reliability and quality of machining processes: Utilize the integrated data sources made available through the RE4DY Framework to improve the reliability and quality of machining processes. This objective aims to enable better decision-making, leading to enhanced product quality and customer satisfaction.
6. Increase transparency and autonomy of machining and supply chain processes: Leverage the shared data and services within the ecosystem to enhance the transparency and autonomy of machining and supply chain processes. This objective aims to reduce manual efforts and streamline operations, resulting in increased efficiency and productivity.
7. Reduce energy consumption and environmental footprint: Utilize the data-driven insights and integrated services to optimize machining processes and reduce energy consumption. This objective seeks to contribute to sustainability efforts by minimizing the environmental footprint associated with machine tool operations and the production of products.
8. Improve diagnostic and predictive capabilities: Harness the power of the shared data to enhance diagnostic and predictive capabilities in the machine tool





sector. This objective aims to enable advanced data-driven services such as auto-diagnostics, predictive maintenance, and predictive planning of manufacturing and supply chain processes.

By embracing circular machining principles, the RE4DY project aims to transform machining processes into more sustainable, resource-efficient, and environmentally friendly operations. The business case promotes the circular economy paradigm, enabling the manufacturing industry to optimize resource utilization, reduce waste generation, and contribute to a more sustainable and circular future.

### 3.1.5 Distributed, Green, Zero-X Manufacturing

The "Distributed, Green, Zero-X Manufacturing" business case within the RE4DY project focuses on the development and implementation of manufacturing systems that are distributed, environmentally sustainable, and have zero carbon emissions (Zero-X). It aims to transform traditional manufacturing practices by leveraging decentralized production, renewable energy sources, and green technologies.

The Distributed, Green, Zero-X Manufacturing Business case focuses on the development of a robust platform that enables near real-time predictive quality analysis for distributed multi-plant manufacturing processes. The business case is built upon four main pillars that aim to revolutionize the industry and address existing challenges.

1. **Develop a platform for near real-time predictive quality:** The first objective of the business case is to create a platform that empowers manufacturers to proactively detect and address potential defects or quality issues in their manufacturing processes. Leveraging advanced AI/ML techniques, the platform will analyse real-time data from various sources and provide insights to operators, highlighting areas that require further investigation. By detecting quality issues in advance, manufacturers can take prompt actions to mitigate risks and ensure high-quality production.
2. **Implement AI/ML-based defect detection tools:** To support quality inspection operations, the business case will deploy AI/ML-based defect detection tools. These tools will assist operators in their inspection tasks by leveraging machine learning algorithms to identify potential defects or anomalies in manufactured parts or components. By automating certain aspects of the inspection process and providing real-time insights, these tools will enhance the efficiency and accuracy of quality control procedures.
3. **Create predictive quality algorithms for continuous improvement:** The business case aims to develop predictive quality algorithms that are tailored to specific manufacturing processes. These algorithms will leverage historical and real-time data to predict and prevent potential quality issues. By continuously analysing manufacturing data, the algorithms will identify patterns, trends, and potential risks, enabling manufacturers to implement continuous improvement processes. This proactive approach to quality management will result in optimized manufacturing operations and improved overall product quality.
4. **Improve training processes for inspection operators:** To reduce dependency on specialized and experienced inspection operators, the business case aims to improve training processes. By leveraging AI/ML tools and advanced simulations, the pilot aims to provide operators with comprehensive training materials and interactive learning experiences. This will enhance their expertise in quality



- inspection, reduce training time, and empower a broader workforce to contribute to quality control processes.
5. Establish a Manufacturing Digital Thread: The business case recognizes the need for seamless data sharing and consumption in a multi-Digital Twin environment across geographically distributed plants. Therefore, the pilot aims to establish a Manufacturing Digital Thread that connects different stakeholders involved in the manufacturing processes. This digital thread will enable efficient collaboration, information exchange, and visibility across the manufacturing ecosystem. By creating an integrated data environment, manufacturers can make more informed decisions, enhance process transparency, and optimize the entire supply chain.
  6. Ensure scalability and extension to value chain players: The Manufacturing Digital Thread developed in the business case will be designed with scalability in mind. It will be capable of accommodating additional value chain players, such as suppliers and service providers, to foster a more integrated and connected ecosystem. This extension will enable seamless collaboration and data sharing among all stakeholders involved in manufacturing processes, resulting in enhanced efficiency, productivity, and innovation throughout the value chain.

By adopting the principles of "Distributed, Green, Zero-X Manufacturing," the RE4DY project aims to promote sustainable manufacturing practices, reduce carbon emissions, and drive the transition towards a greener and more environmentally conscious manufacturing industry. This business case aligns with the broader global goals of mitigating climate change, conserving resources, and achieving a sustainable future.

### 3.2 Digital continuity service layer

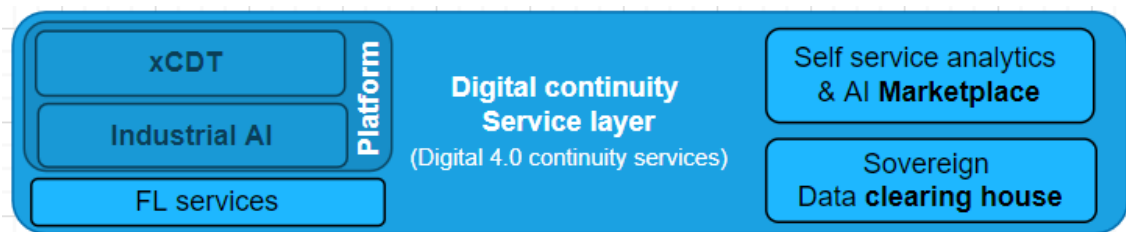


Figure 15: Digital continuity service layer building blocks

The RE4DY “Digital continuity service layer”, as an extension of the Boost 4.0 reference architecture, encompasses eXecutable Cognitive Digital Twins (xCDT), Industrial AI, Federated learning services, Self-service analytics, an AI and Data marketplace, and the Sovereign data clearing house. These components collectively enable seamless integration, advanced analytics, secure collaboration, and efficient data management within the manufacturing sector, empowering organizations to unlock the full potential of digital transformation.

- eXecutable Cognitive Digital Twins (xCDT): The eXecutable Cognitive Digital Twins (xCDT) are virtual representations of physical assets, processes, or systems within the manufacturing environment. These digital twins combine real-time data with advanced analytics and AI capabilities to provide a holistic understanding of the physical counterparts. xCDTs allow manufacturers to monitor, simulate, and



optimize operations, facilitating predictive maintenance, process optimization, and scenario planning.

- **Industrial AI:** Industrial AI involves the application of artificial intelligence (AI) techniques specifically tailored for manufacturing processes. It utilizes AI algorithms and machine learning models to analyse data generated by machines, sensors, and other industrial assets. Industrial AI enables manufacturers to gain insights into production efficiency, quality control, predictive maintenance, and anomaly detection. By leveraging AI, manufacturers can optimize processes, reduce downtime, improve product quality, and enhance overall operational efficiency.
- **Federated Learning Services:** Federated learning services enable collaborative model training across multiple manufacturing sites or partners while preserving data privacy. Instead of sharing raw data, federated learning allows the training of machine learning models on decentralized data sources. This approach enables manufacturers to harness the collective knowledge and insights from distributed datasets without compromising data security or violating privacy regulations.
- **Self-Service Analytics and AI Marketplace:** The Digital continuity service layer also includes self-service analytics and an AI and Data marketplace. Self-service analytics empowers users within the manufacturing sector to independently access, explore, and analyse data without heavy reliance on IT or data science teams. It provides user-friendly tools and interfaces that allow users to create custom dashboards, perform ad-hoc queries, and derive actionable insights from manufacturing data.

The AI and data marketplace is a curated platform that facilitates the discovery, procurement, and integration of pre-built AI models, algorithms, and applications tailored to the manufacturing industry. It serves as a centralized hub where manufacturers can explore and leverage a wide range of AI solutions offered by third-party providers. The marketplace fosters innovation and accelerates the adoption of AI technologies within the manufacturing sector.

- **Sovereign Data Clearing House:** The sovereign data clearing house acts as a trusted intermediary within the Digital continuity service layer. It ensures data privacy, security, and compliance by managing the exchange of data between different stakeholders in the manufacturing ecosystem. The clearing house establishes protocols and mechanisms for data sharing, consent management, and auditing, enabling secure collaboration and data-driven decision-making while maintaining data sovereignty.

### 3.2.1 Integrated trusted Industrial AI and eXecutable CDT Platform



Figure 16: Cognitive digital twin and industrial AI platform



Digital Threads are the map for the digital journey based on specific business workflows. They digitally connect all tasks and processes of the entire lifecycles of product and production to solve the challenges. The Digital Threads provide digitalized processes and revolutionize the way products are developed, produced, and optimized. On the one hand, Digital Threads refer to horizontal integration. The horizontal integration includes the entire value chain from design, to production, service and re-cycling. It connects everything over its lifecycle from product innovation to production through product in use. On the other hand, Digital Threads include beside the horizontal interoperability the perspective of vertical integration. Field devices, such as controllers on the shop floor, generate a huge amount of data. Vertical integration leverages the same data analytics capabilities from information technology on the top floor to operations technology on the shop floor to get the most value from that data.

The Digital Twin creates new insights, thanks to the combination of physics-based simulations with data analytics in a fully virtual environment. Performance data from the real world allow to feed those back to the virtual model to continuously optimize product and production including the necessary logistic processes. Therefore, all relevant physical assets need to be connected and interoperable. Once the product is built in real the data collection of the product behavior, usage and performance is possible.

The goal of the cognitive digital twin is in the closed-loop connection between the real and the digital world. Through this connection actionable insight gained from the physical world for informed decisions throughout the lifecycle of products and production operations.

The Digital Twin for products enables inter-disciplinary collaboration and parallel development to create an accurate virtual representation of a new product in almost no time.



Figure 17: Product manufacturer perspective

The Digital Twin for production or line machine builder represents the virtual model of the shop floor, which is modelled to optimally manufacture the new product.



Figure 18: Line machine builder perspective

In a further step the cognitive digital twin contains and represents the actual knowledge of a solution architecture. Moreover, the cognitive digital twin is intended to be a hybrid,



adaptive system and is characterized by the fact that it evolves with the collected data and AI. A cognitive digital twin is representing the real situation capable to consider and involve real time conditions of all connected devices and systems. It's a continuously calibration and update of a machine learning model while its running.

Ensuring the inclusion of real time conditions, physical assets must be integrated into the CDT. Exemplarily the Siemens PLM system Teamcenter (Teamcenter, 2022) enriches and concentrates data which represents the assets and product through the whole lifecycle. PLM Teamcenter helps industrial customers to generate representative datasets in a flexible and efficient way. Those datasets are building the foundation for a Digital Product Passport or the instances of the Asset Administration Shell.

### 3.2.2 Federated Learning Services

The possibility given by Federated Learning services to train machine learning models by leveraging a wide range of datasets by simultaneously offering privacy guarantees, allows for the realisation of eXecutable Cognitive Digital Twins (xCDT). In other words, tools that through AI and data collected from various (even geographically distributed) data sources can learn, monitor and predict in real-time the status of many kinds of assets. Such tools, complemented by additional services and offered to the end users usually through Industrial AI platforms often as part of wider pipelines, find applications in many industrial business scenarios and can bring unprecedented benefits to the organizations.

A first example, applicable to many manufacturing contexts, is the ability to predict the Remaining Useful Life (RUL) of a machine or piece of equipment. Knowing how long it will take for a machine to breakdown allows shopfloor operators to perform maintenance activities before incurring in a disruption of the productive process (predictive maintenance), thus avoiding expensive unplanned emergency interventions and delays in the delivery of the final product. Furthermore, Machine Learning models can offer valuable insights to the quality of a product. Namely, in the manufacturing community unexpected machinery failures can negatively impact the final product's quality. To mitigate this issue, AI algorithms are trained to support and provide the ability to early detect and subsequently identify quality defects arising from faults in machinery equipment. The operators can be notified on time in order to activate the proper mechanisms to avoid the defects. In the context of Machining, similar purposefully trained AI algorithms are also able to determine the wear level of a tool, and therefore suggest a timely replacement. By replacing the tool before it exceeds a certain wear level allows for the refurbishing of the tool itself but also increases the quality of the being-worked-on piece by reducing the number of defects. The same considerations apply not only to the tool but also to the machine tool. Moreover, because of the increased equipment effectiveness, the amount of rework or scrap is lowered and the overall carbon footprint as well as energy consumption reduced. Finally, the opportunity given to collect, while still preserving privacy, insights from the production processes of various customers also allows for an increase in the product quality due to an optimised use of the tools and can enable the reduction of the inventory through predictive ordering of the tools.

The enhancement of data privacy and security is a key aspect of the federated services, as, most times, sensitive data remain within the limits of the device or service that produces them. As a result, the risk of a data breach is minimised. However, the challenge of this type of services is the requirement of establishing an efficient communication protocol between the different members of the federated architecture and the development of



efficient mechanisms for aggregating the results sourced from the multiple participants in a synchronized and secure manner.

In the context of Zero-X Manufacturing, federated learning services can be employed to design, develop and train AI models for inspecting images of produced pieces in search for defects. By highlighting the portion of piece containing or that could contain defects, the algorithm assists the operator in inspecting the piece and triggering, if required, a rework or scrapping process. As a result, (i) the time required to identify a defect is lowered and (ii) the overall quality of the product is increased (higher customer satisfaction) with particularly important benefits in terms of safety for the final users. Moreover, given the possibility to use these algorithms to support junior inspectors in practising while reducing the level of reliance on senior inspectors, the resiliency of the entire inspection process is increased. In manufacturing sectors where the authorities permit, it is even possible to envisage fully autonomous defects identification.

Beyond specific domains of application, federated learning services also enable a new business model where companies can offer their data to other companies behind payment and without physically sharing the datasets.

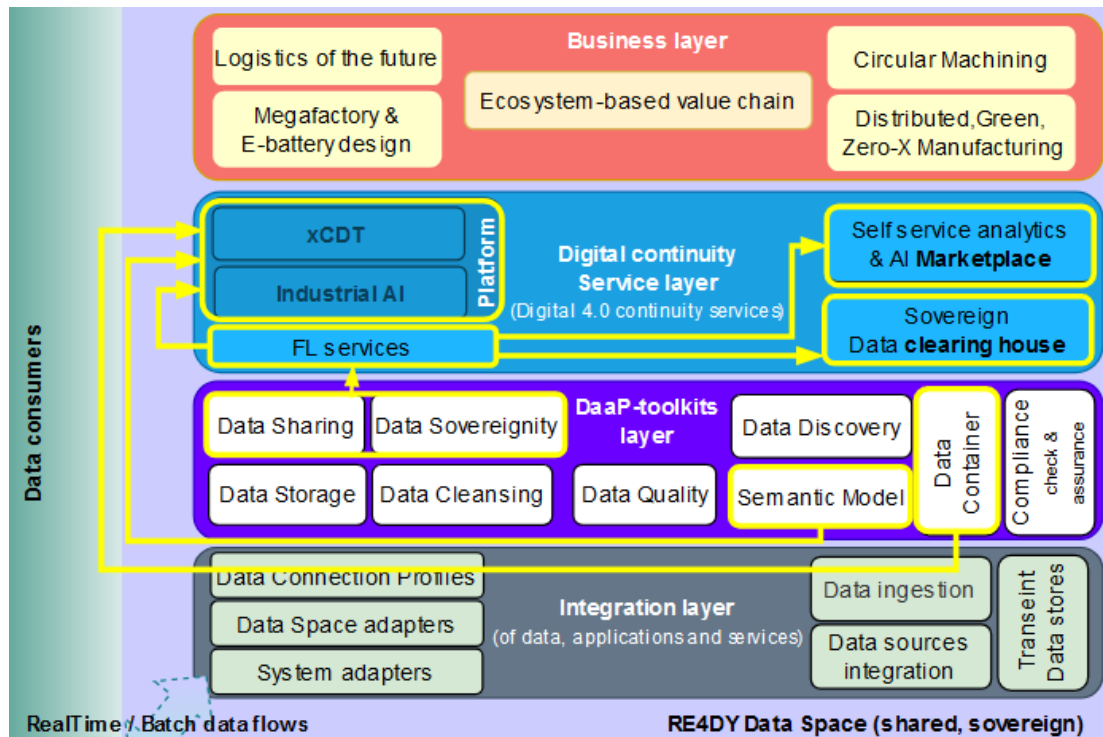


Figure 19: RA interactions among modules and layers involved in/by “FL services”

For AI/ML applications and models to be leveraged by Industrial AI platforms, it is necessary that Federated Machine Learning platforms expose them through interoperable and well-defined APIs (application programming interfaces). The same interfaces can then also be used by Self-service and AI Marketplaces where AI/ML applications, along with the available datasets, are catalogued, presented and made available to the end users. Eventually, the availability of AI/ML assets to Industrial AI platforms and Marketplaces allow end users to devise relevant solutions for their business scenarios (defined in the Business Layer above).



To enable this innovative approach to Data Sharing, Federated Learning tools and platforms take advantage of a well-defined underlying DaaP-toolkits layer which guarantees the interoperability of datasets and algorithms by Data Containers and Semantic Models, and high data quality via specialized cleansing and quality tools capable of making datasets compliant with pre-defined quality standards. Throughout the entire data pipeline, data owners always maintain control over their datasets and the usage to which they are exposed through Data Sovereignty mechanisms alongside the Sovereign Data Clearing House services.

### 3.2.3 Self-Service analytics and AI marketplace

The RE4DY self-service analytics and AI marketplace is a component that enables the discovery, procurement, and integration of data assets from various sources. It serves as a curated environment where users can explore and access diverse datasets relevant to their specific needs. The marketplace facilitates the discovery of data assets and ML/AI models, supports data procurement processes, and provides the necessary infrastructure for seamless integration of data into users' analytics workflows. It acts as a central hub for data-driven innovation and collaboration within the manufacturing ecosystem.

It relies on concepts such as Data Containers (DC) and Data Connection Profiles (DCP) for efficient data management. It is also an enabler for federated learning and on-demand AI to develop advanced analytics and AI capabilities. The marketplace also supports data-driven business models, providing opportunities for organizations to apply different monetization schemas on their data assets. Additionally, data sharing, along with initiatives like IDSA and GAIA-X data spaces, ensures secure and trusted data exchange within the marketplace. Thus, within the AI Marketplace:

- Data Containers (DC) provide a standardized format and structure for data assets available in the marketplace. These containers encapsulate data generated from machines, sensors, and other sources, making it easier to manage and exchange these data. Specific use cases needs are implemented by Data Connection Profiles (DCP), which enable simple, durable, and context-sensitive integration between complex systems without any wasteful custom development. Thus, the marketplace enables users to discover, procure, and utilize data assets with ease, ensuring seamless integration and accessibility of data.
- It can facilitate federated learning and on-demand AI by enabling collaborative model training across distributed datasets and provide users with access to pre-built AI models and algorithms. By integrating federated learning and on-demand AI, the marketplace facilitates advanced analytics and AI capabilities for users, driving insights and innovation.
- Data sharing is a fundamental aspect of the marketplace's functionality. The marketplace provides a platform for users to share and collaborate on data assets, enabling cross-functional analysis and fostering innovation. Data sharing within the marketplace can be facilitated through controlled access mechanisms, data sharing agreements, and secure data exchange protocols.
- It enables the exchange of data among stakeholders, allowing the development of data-driven business models and monetization schemas. Data marketplace functionalities provide opportunities for organizations to leverage data assets for





various purposes, including commercialization, partnerships, and value creation. The marketplace facilitates the exploration of new business models and monetization strategies that are driven by data and analytics.

- The International Data Spaces Association (IDSA) and GAIA-X initiatives – which are focused on creating secure and trusted data exchange frameworks – aim to establish data ecosystems that prioritize data sovereignty, privacy, and security. In the context of the AI Marketplace, the integration of IDSA and GAIA-X data space concepts and services ensures adherence to data governance principles, privacy regulations, and secure data exchange practices. This integration enables the marketplace to operate within a trusted framework, enhancing the confidence and reliability of data sharing and collaboration among participants.

### 3.2.4 Sovereign data clearing house

The Sovereign Data Clearing House component in the Digital Continuity Service Layer refers to a central entity or platform that facilitates secure and controlled data sharing among participants while ensuring data sovereignty and compliance with relevant regulations.

In addition, the Sovereign Data Clearing House can typically be integrated with an AI/Data marketplace within the Digital Continuity Service Layer, serving as a trusted data exchange mechanism within the marketplace. The clearing house ensures the privacy, security, governance, and compliance aspects of data sharing, enhancing the trustworthiness of data transactions and fostering collaboration among participants within the marketplace, where data providers can share their data with authorized consumers.

The concept of a Sovereign Data Clearing House is also closely related to the specifications and principles promoted by the International Data Spaces Association (IDSA) and GAIA-X, as both have shared objectives of promoting secure and trusted data exchange, data sovereignty, and data privacy within the context of data spaces. Linked to this, IDSA and GAIA-X frameworks, encompass five key aspects that are fundamental to the management and exchange of data within the ecosystem. These aspects are:

- **Data as an Economic Good:** Data as a valuable economic asset. IDSA and GAIA-X promote the concept of treating data as a tradable and monetizable resource, enabling businesses to generate value from their data assets. By considering data as an economic good, they encourage the development of data-driven business models and the establishment of fair and transparent mechanisms for data exchange and monetization. This can be achieved by putting in place clearing and billing mechanisms that may involve the use of standardized protocols and contracts for data exchange, as well as mechanisms for tracking data usage and facilitating payment or compensation based on agreed-upon terms.
- **Data Ownership:** recognition that data is typically owned by individuals or organizations that generate or collect it. Data owners have the right to control and govern how their data is shared, accessed, and used by others. The data space frameworks ensure that data owners retain control over their data assets, allowing them to make informed decisions about data sharing and collaboration.
- **Data Sovereignty:** referring to the concept that data should be subject to the laws and regulations of the jurisdiction in which it is generated or stored. IDSA and GAIA-X place a strong emphasis on data sovereignty, ensuring that data remains under





the jurisdiction and control of the data owner. This aspect ensures compliance with legal and regulatory requirements, protecting the rights and interests of data owners and enabling trust in data exchange.

- **Data Quality:** Data quality is crucial for reliable and meaningful data exchange, recognizing the significance of data quality and promotes mechanisms to ensure data accuracy, integrity, and completeness. By focusing on data quality, these data space frameworks enable participants to have confidence in the data they exchange, leading to more accurate and valuable insights and decision-making.
- **Data Provenance:** referring to the origin and history of data, including its sources, transformations, and any changes made to it over time. IDSA and GAIA-X emphasize the importance of data provenance, ensuring transparency and traceability in data exchange. By maintaining a clear record of data provenance, the data space enables participants to understand the lineage and trustworthiness of the data they access and use.

On a more technical level, both IDSA and GAIA-X promote the use of self-descriptions, which are standardized metadata representations of data assets that enable secure and trustworthy data exchange. Self-descriptions provide detailed information about the data, including its structure, format, quality, and usage policies<sup>9</sup>.

The Sovereign Data Clearing House can leverage the self-description standards to ensure that data assets shared through the marketplace include comprehensive metadata. This allows participants and consumers to discover, evaluate, and utilize data resources while ensuring transparency and compliance with data governance requirements and have a clear understanding of the data's characteristics and usage conditions, facilitating secure and informed data exchange.

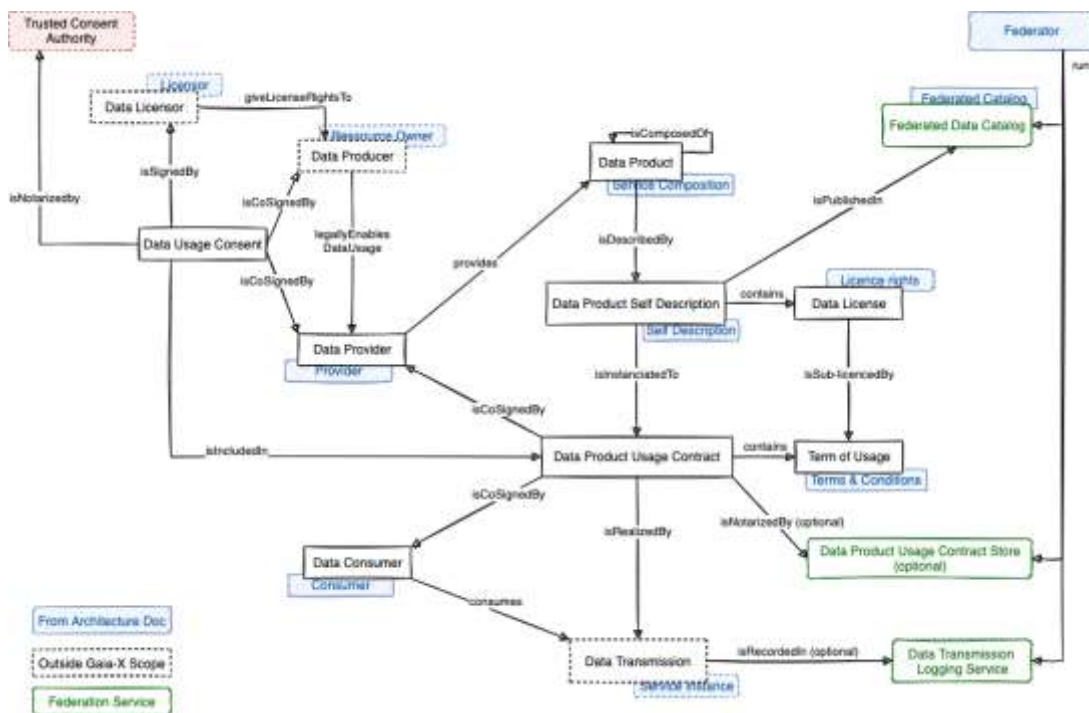


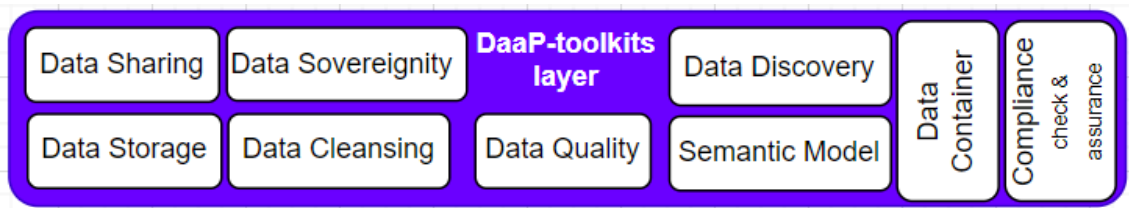
Figure 20: GAIA-X Data Product Self-Description Conceptual Model<sup>10</sup>

<sup>9</sup> This also closely links with RE4DY Data as a Product (DaaP) concept.

<sup>10</sup> Source: <https://docs.gaia-x.eu/technical-committee/data-exchange/22.10/dewg>



### 3.3 DaaP-toolkits layer



*Figure 21: DaaP-toolkits layer building blocks*

This layer is responsible for the realization of an application agnostic data-centric middleware platform. Through its modules DaaP-toolkits layer guarantees the interoperability of datasets through the use of Data Containers and Semantic Models, the high quality of datasets through the Data Cleansing, Quality and Compliance modules and reusability through Data Discovery, Data Sharing and Data Sovereignty modules. Via the tools provided, this layer not only brings all the data together but also transforms and processes them to meet the needs of AI and ML automatically and at sustainable levels, thereby increasing reusability and avoiding the repetition of the same functionality in each individual digital twin application.

The modules defined in the DaaP-toolkits layer are the following:

- Data Sharing
- Data Sovereignty
- Data Discovery
- Data Storage
- Data Cleansing
- Data Quality
- Semantic Model
- Data Container
- Compliance Check & Assurance

Detailed descriptions of these components are provided in the following sub-sections.

#### 3.3.1 Data sharing

Data sharing is a concept that has become increasingly important in the digital economy. As companies rely more and more on data as a strategic resource, they are finding that they need to collaborate with other organizations to achieve their goals. However, this collaboration can be complicated by the need to protect sensitive data from unauthorized access or misuse. In this section, we will explore the concept of data sharing in more detail, including its benefits and challenges.

One of the key benefits of data sharing is that it allows organizations to leverage their collective knowledge and expertise. By pooling resources and sharing information, companies can gain insights that would be difficult or impossible to obtain on their own. For example, a group of healthcare providers might share patient data to identify patterns and trends that could help them develop new treatments or improve existing ones. Another benefit of data sharing is that it can lead to new business models and revenue streams. By creating data marketplaces or other collaborative platforms, companies can monetize



their data assets in new ways. This can be especially valuable for organizations that have large amounts of valuable but underutilized data.

However, there are also significant challenges associated with data sharing. One of the biggest concerns is privacy and security. Companies need to ensure that sensitive information is protected from unauthorized access or misuse. This requires robust security measures such as encryption, access controls, and monitoring systems. Another challenge is ensuring that all parties involved in the collaboration can benefit fairly from the shared information. This requires careful negotiation and agreement on issues such as data ownership, access rights, and revenue sharing. In some cases, it may be necessary to involve third-party intermediaries such as data brokers or escrow services to ensure that everyone's interests are protected. Finally, there is the challenge of ensuring that the shared data is accurate and reliable. This requires careful attention to data quality issues such as completeness, consistency, and timeliness.

### Patterns for data sharing

There are several general architectural patterns for data sharing. These patterns provide a framework for designing systems that enable data sharing between organizations while ensuring security, privacy, and reliability:

- One common pattern is the use of *APIs* (Application Programming Interfaces) to enable data exchange between different systems. Different applications can communicate with each other through APIs in a controlled and secure manner.
- Another pattern is the use of *data warehouses or data lakes* to store and manage large amounts of shared data. Having access to these centralized repositories enables multiple organizations to share information while maintaining control of their own data.
- A third pattern is the use of *federated identity management systems* to enable secure access to shared resources. These systems allow users from different organizations to access shared resources using their own credentials, without requiring them to create new accounts or passwords.
- Finally, there is the pattern of using blockchain technology for secure and transparent sharing of information. Blockchain provides a decentralized ledger that can be used to record transactions and other types of information in a tamper-proof manner. This makes it ideal for applications such as supply chain management or digital identity verification where trust and transparency are critical.

### RE4DY approach for data sharing

With respect to the principles of the International Data Spaces Association (IDSA), the RE4DY project aims to support data sharing practices that promote self-determination and control. With IDS-certified technology, data will be shared in a secure manner while maintaining data sovereignty for the creator and establishing trust among participants. This approach aligns with the model of a sovereign data space based on IDS principles, and its core building blocks. (see next section, 3.3.2)

Data sharing in RE4DY will leverage connectors for sovereign and secure data sharing in various parts of the project's data fabric. Different usage control policies will be implemented, allowing a Data Provider using an IDS Connector to customize the access and use of their data. These policies will vary, from allowing unrestricted data usage to



imposing specific restrictions regarding usage intervals, usage frequency, and required deletion times, among others.

### 3.3.2 Data sovereignty

Data sovereignty is the concept that data are subject to the laws and governance structures of the nation where they are collected. Data sovereignty is closely linked with concepts like data security, cloud computing, network sovereignty etc.

RE4DY project will deliver a sovereign data space following the principles of International Data Spaces Association (IDSA) that aiming to be a global standard for data exchange transactions that will be based on IDS-certified technology.



IDSA enables self-determination and control over data usage to remain in the hands of those who collect, store and provide it, rather than passing to large data exchange platforms and others, as is often the case today.

IDSA offers:

- Assurance of data sovereignty for the creator of the data and trust among participants.
- Equal opportunities through a federated design.
- Data privacy and security that's the most trusted in the world.

IDSA provides its Reference Architecture Model (IDSA RAM) that is presented in chapter 2. It enables companies to develop and utilize vendor-independent data ecosystems and marketplaces, open to all, at low cost and with low entry barriers.

This fits greatly with data sovereignty concept and needs. IDSA RAM contains the conceptual level including technology-agnostic specifications and it is based on five layers (Business, Functional, Process, Information, System) and the three perspectives (Security, Certification, Governance).

IDSA RAM supports various roles such as Data Owner, Data Provider, Data Consumer, Service Provider, App Provider, Vocabulary Provider and Broker Provider. Each one of these roles is strictly connected with core components in IDSA RAM.

Figure 22: Key Idea of Data Sovereignty



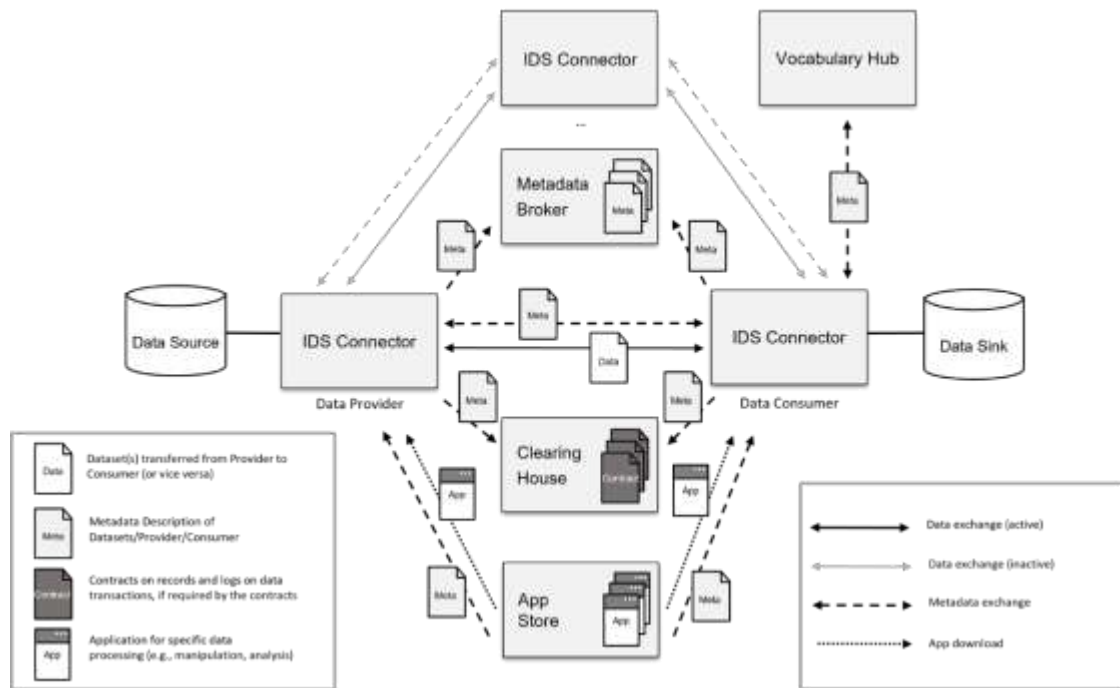


Figure 23: Roles, Interactions and Core Components of IDSA RAM

As depicted in the previous picture a sovereign data space based on IDS principles is based on various core building blocks, IDS Connectors (IDSA, IDS connector), Clearing House (IDSA, Clearing House), Broker (IDSA, Metadata Broker), App Store (IDSA, App Store), Identity Provider (IDSA, Identity Provider) and Vocabularies (IDSA, Vocabulary hub).

RE4DY approach for sovereign data sharing

RE4DY will achieve data sovereignty services by building on IDSA principles and by using components such as connectors for sovereign and secure data sharing in various parts of project’s data fabric. Currently various connectors are under evaluation and planned to be used such as Dataspace Connector (IDSA, Dataspace Connector), Eclipse Dataspace Connector (EDC (Eclipse)) etc.

Sovereign Data Sharing will be enabled as a Data Provider using an IDS Connector is able to setup various Usage Control policies:

#	Title	Description
1	Allow the Usage of the Data	provides data usage without any restrictions
2	Connector-restricted Data Usage	allows data usage for a specific connector
3	Interval-restricted Data Usage	provides data usage within a specified time interval
4	Duration-restricted Data Usage	allows data usage for a specified time period
5	Restricted Number of Usages	allows data usage for n times
6	Use Data and Delete it After	allows data usage within a specified time interval with the restriction to delete it at a specified time stamp
7	Local Logging	allows data usage if logged to the Clearing House



8	Remote Notifications	allows data usage with notification message
---	----------------------	---

### 3.3.3 Data discovery

Data discovery is a fundamental step in the data lifecycle and a key component within the RE4DY architecture. It involves techniques and practices for identifying, locating, and understanding their structure and content, and determining their suitability for specific use cases. It may involve cataloguing data assets, metadata management, semantic annotation, and the use of search mechanisms to facilitate efficient data discovery.

By enabling data discovery, organizations can effectively leverage data assets for analytics, decision-making, and innovation. It supports the identification of valuable insights, patterns, and correlations that drive operational efficiencies, optimize processes, and create new business opportunities.

Effective data discovery is essential in the context of data sharing and IDSA/GAIA-X data spaces. By adopting the concepts of Data as a Product (DaaP), self-descriptions, DCAT, vocabularies, semantics, FAIR data principles, and integrating with AI4EU on-demand APIs, RE4DY enables data providers to package and market their data assets, enhance metadata representation, and leverage advanced techniques for efficient and personalized data discovery. These efforts contribute to the overall goal of creating secure and trusted data exchange ecosystems in the manufacturing sector.

The RE4DY Data as a Product (DaaP) and IDSA/GAIA-X self-descriptions share the common goal of enhancing data interoperability, discoverability, and value realization. By utilizing standardized self-descriptions, data assets can be described in a consistent and structured manner, facilitating their packaging, presentation, and exchange as products. Self-descriptions provide a means to convey relevant information about the data (by encapsulating additional metadata and information, such as quality attributes, provenance, usage policies, and pricing models,) to make it easier to discover, consume, and monetize, and thus enabling potential consumers to make informed decisions about data acquisition and utilization.

These concepts, along with the use of DCAT, enable data providers to describe their data assets effectively and make them discoverable within data marketplaces and other platforms. Moreover, the use of domain vocabularies and metadata standards provides a common language for describing data assets, enhancing the consistency and clarity of metadata representation. Semantics also play a crucial role in data discovery by enabling machines to understand the meaning and relationships between different data assets, facilitating more precise and relevant search results.

In addition, RE4DY adheres to FAIR principles, therefore making data findable, accessible, interoperable, and reusable, which significantly improves the discoverability of data, enabling effective data discovery and promoting data sharing and collaboration.

Finally, use of AI4EU On-Demand APIs within marketplaces in the context of data discovery can serve as a valuable resource as it provides intelligent search capabilities, recommendation systems, and data analysis functionalities that facilitates the discovery, procurement, and integration of pre-built AI models, algorithms, applications, and services that can be accessed on-demand.



### 3.3.4 Data storage

The Data Storage Component plays the role of the repository for all the existing datasets that have made their way in after being pre-processed. It can be seen as a pseudo data-lake, containing all the pre-processed datasets. This comes with all the dataset general requirements, such as scalability, portability, data efficiency, etc.

After a dataset has been collected and pre-processed, it then gets stored in the Data Storage Component. For example, when data is requested to be applied in a machine learning flow, the training dataset will be stored.

The requirements that the Data Storage block must comply with are the following:

- Handling of large volumes of data at high speed with a scale-out architecture.
- Unstructured, semi-structured, or structured data storage.
- Enabling easy updates to schemas and data fields.
- Developer-friendly.
- Take full advantage of cloud to deliver zero downtime.

The Data Storage component consist of various data storage solutions. Pilots can select the specific solution/s that best suit their particular requirements.

Consequently, a NoSQL database could be used as an auxiliary for temporary (or permanent) data storage options, carrying modifications and custom parameters, to comply with the project's efficiency standards. Some examples for NoSQL databases are MongoDB<sup>11</sup> and Cassandra<sup>12</sup>. Both sharing many similarities but also having their own unique elements and differentiating sections. For example, both are open-source NoSQL databases that support horizontal partitioning and are horizontally scalable. They also support both support replication. The biggest differences are:

- MongoDB is a document store database that works with collections containing multiple documents, whereas Cassandra is a column-oriented database.
- MongoDB has a master-slave architecture, while Cassandra has a peer-to-peer architecture where all are master nodes in communication with each other.
- MongoDB uses binary JSON or BSON format, an extremely expressive data format, to store data while Cassandra uses a columnar style and tables.

In addition, a different type of database (for example, PostgreSQL<sup>13</sup>) could be included to store other type of data for which NoSQL might not be the best fit, such as camera images for detecting faulty pieces.

### 3.3.5 Data cleansing

Data cleansing is a procedure which involves the identification and resolution of numerous issues regarding the integrity and correctness of a dataset. A non-exhaustive list of such issues includes duplicate or redundant records, missing values, inaccurate or inconsistent records, errors, and outliers. In cases where regulatory compliance is necessary, data cleansing may also involve the anonymization of data through the removal of sensitive information. The large volume of data disseminated in RE4DY along

<sup>11</sup> <https://www.mongodb.com/>

<sup>12</sup> <https://cassandra.apache.org/>

<sup>13</sup> <https://www.postgresql.org/>





with the establishment of an AI Marketplace for datasets and other AI assets make the development and integration of a data cleansing process a highly valuable addition to the ecosystem.

While data cleansing may be accomplished manually, this is not always possible or efficient. The large number of datasets circulating the RE4DY data space justifies tackling this issue in an automatic manner in order to maximize time efficiency and complete the process with the least amount of manual human intervention. Within the context of an IDS compliant data space, data cleansing is a process which can be easily integrated into the data sharing procedure, either before or after it, by utilizing the functionality offered by Data Apps.

IDS Data Apps are pieces of software which may be discovered through an IDS App Store and installed in a connector. Their main purpose is to process data in diverse ways and output the resulting data in a predetermined format, making them ideal candidates for data transformation and cleansing. Additionally, it is possible to chain a number of Data Apps so that the output of one Data App is piped into the input of the next Data App. This may be used to achieve data cleansing in a modular way and allow the development of dedicated, single-purpose Data Apps composing the data cleansing chain. For instance, the chain could be composed of apps performing duplicate elimination, outlier detection with statistical methods and finally transformation to a format which complies with the needs of the data consumer.

It is crucial to ensure that the data being shared and processed by the stakeholders is error-free, reliable and of superior quality. Acquired datasets may be imperfect and applying data cleansing techniques before their use in order to rectify such issues is of utmost importance. Therefore, a robust data cleansing process constitutes a strict requirement. The process is to be applied before or after the actual data exchange, in order to guarantee the high quality of datasets published in the AI marketplace, the necessary preparation of data for training, the facilitation of data processing and the prevention of errors due to faulty data which contribute to inconsistent results.

### 3.3.6 Data quality

Data quality encompasses a number of highly desired characteristics regarding a dataset, including completeness, accuracy, validity, consistency and reliability. It may additionally form an umbrella term for various techniques aiming to reduce the impact of lower quality data such as data transformation, cleansing and FAIRness.

Rendering managed data operational is of utmost importance since poor quality data introduces friction during data sharing, processing, and analysis, which has the potential to be detrimental in a data-driven ecosystem such as RE4DY. For instance, inaccurate or missing data used as training input in machine learning algorithms constitutes a significant risk of reduced model accuracy. In turn, the integrity of the results and therefore their interpretation is weakened, which leads to potentially false analysis and flawed decision-making. The same points apply to datasets published in the AI asset marketplace.

In the domain of data sharing and especially when considering Data as a Product (DaaP) principles, it is crucial that managed data is refined until suitable for use and reuse to maximize its usefulness, its capacity to easily undergo processing and analysis, and therefore its value. Data quality is also related to the Data Connection Profiles (DCP), an



elementary component of data access and sharing in RE4DY, which offers a trusted computing continuum where data can be stored and accessed by both data providers and consumers. It is of critical importance that data circulating the Data Containers (DC) meets the criteria of high data quality. Finally, from the point of view of data governance, it is essential to ensure the accuracy of the recorded information by eliminating inconsistencies in order to guarantee validity and reliability.

The implementation of a sovereign data space and digital thread fabrics by RE4DY warrants the establishment of a comprehensive data quality control and enforcement procedure. With value creation being the focal point of such a platform, it is vital to ensure data is up to predefined standards to unveil its potential value. Therefore, adherence to quality requirements will benefit all aspects of the highly data-oriented ecosystem implemented by RE4DY.

### 3.3.7 Semantic model

Semantic models are considered as conceptual data models in which semantic information is included. This means that the model describes the meaning of its instances. In particular, semantic data models include the capability to express and exchange information which enables parties to interpret meaning (semantics) from the instances, without the need to know the meta-model. Semantic models offer:

- a. Easy to understand data relationships plus that the relationships are on the surface so no need to derive more data.
- b. Effortless development of applications and services.
- c. Better data visualization and reporting.

As semantic data modelling refers to processes to build conceptual data models that include semantic information that adds a basic meaning to the data and the relationships that lie between them, concepts like taxonomies, ontologies and knowledge graphs are adopted.

Semantic models and ontologies are a core part of RE4DY framework and its DaaP. Even its core building block for the creation of a sovereign data space, the International Data Spaces Information Model (IDSA, International Data Spaces Information Model, 2022) , is expressed as semantic model.

Moreover, semantics and ontologies will be used to model data that will be exchanged over the various project's components and of course during the pilot cases implementation. Ontologies and vocabularies for modelling concepts related to industry 4.0/5.0 and smart manufacturing will be studied and used. Existing widely used data models and ontologies for the aforementioned domains have been gathered in the following figure:





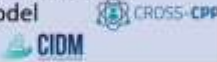


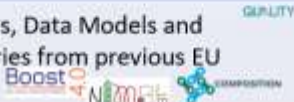
Ontologies/ Data Models/ Vocabularies	Short Description
OntoCommons & IOF 	Ready-to-use Ontology Commons EcoSystem (OCES) for data documentation, including a set of ontologies and tools
OPC UA Information Model and Data Models 	OPC is the interoperability standard for the secure and reliable exchange of data in the industrial automation space and in other industries. It is platform independent and ensures the seamless flow of information among devices from multiple vendors
Cross-CPP -Standardized Cross Industrial Data Model 	Provides a single point of access to data streams from multiple smart products in easily accessible non-proprietary data formats
OGC SensorThings 	It provides a common data model and services that allow IoT devices and applications to CREATE, READ, UPDATE, and DELETE IoT data and metadata. The data model is designed based on the ISO/OGC Observation and Measurement (O&M).
Asset Administration Shell Models and Sub-models 	Asset Administration Shell (AAS) is the digital representation of an asset. The AAS consists of a number of sub-models in which all the information and functionalities of a given asset – including its features, characteristics, properties, statuses, parameters, measurement data and capabilities – can be described.
Ontologies, Data Models and Vocabularies from previous EU Projects 	Various data models and ontologies has been developed in EC-funded projects and they are related to manufacturing and supply-chain domains.

Figure 24: List of Ontologies and Data Models related to Industry 4.0 domain

The semantic models and ontologies that will be developed and used during the project will be the ‘targets’ for data integration and transformation services. Data from disparate data sources will be transformed to data that will be modelled and linked based on these ontological concepts. The RE4DY semantic models will be developed on the latest state-of-the-art methods/tools developed or recommended by the EU project OntoCommons<sup>14</sup> and the Industrial Ontologies Foundry<sup>15</sup> More particularly, the OntoCommons EcoSystem (OCES) hierarchical architecture will be followed for the development of the RE4DY semantic models, reported in the OntoCommons Roadmap<sup>16</sup> and illustrated in the figure below:

<sup>14</sup> <https://ontocommons.eu/>

<sup>15</sup> <https://www.industrialontologies.org/>

<sup>16</sup> <https://zenodo.org/record/7544509#.ZHcGN3ZBzMs>



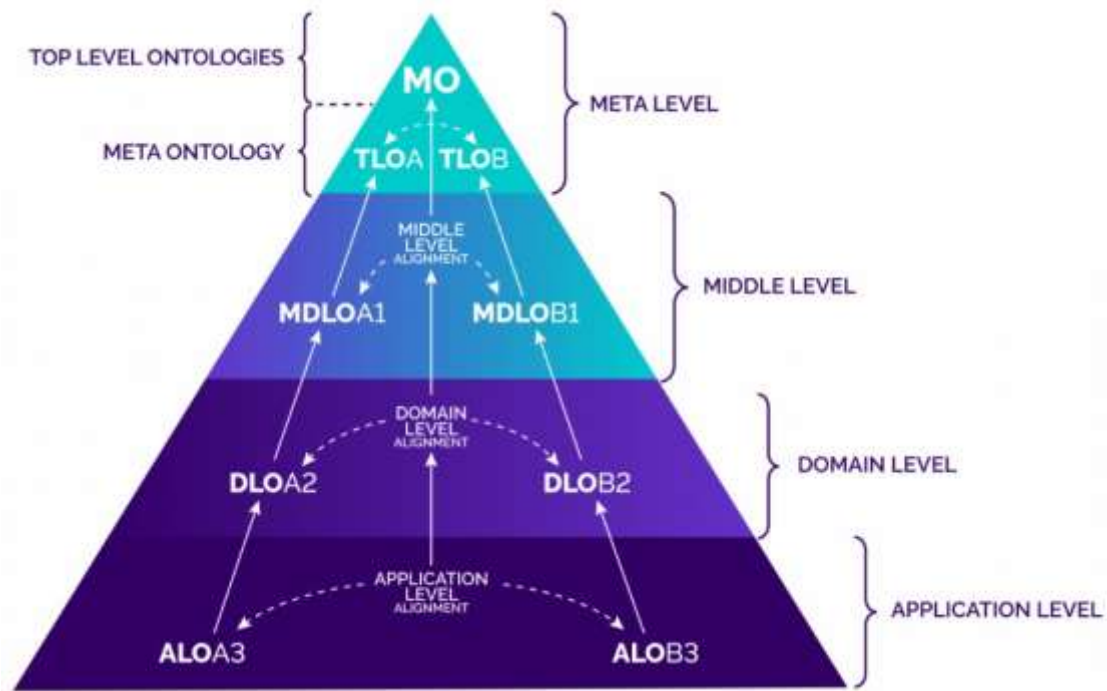


Figure 25: OntoCommons Hierarchical Architecture for building Semantic Models

Furthermore, the most recent IOF ontologies<sup>17</sup> developed by IOF will be used.

Beyond the research regarding the current state-of-the-art of modelling of manufacturing assets and services, RE4DY will employ the Asset Information Modelling Framework (IMF) that enables transitions to information models of facility assets from current documentation practice. In order to achieve its purpose, the IMF framework is comprised of methods and resources designed to support incremental and scalable implementation. IMF provides incremental implementation where current ways of working are dominated by legacy systems, tools, and work processes, by creating value at each step of a gradual modelling exercise. Rather than aiming at developing one single model of an asset from the start, IMF identifies different types of models, where context models are the most basic ones.

<sup>17</sup> <https://industrialontologies.org/the-iof-ontology-version-1-has-been-launched-in-february-2023/>



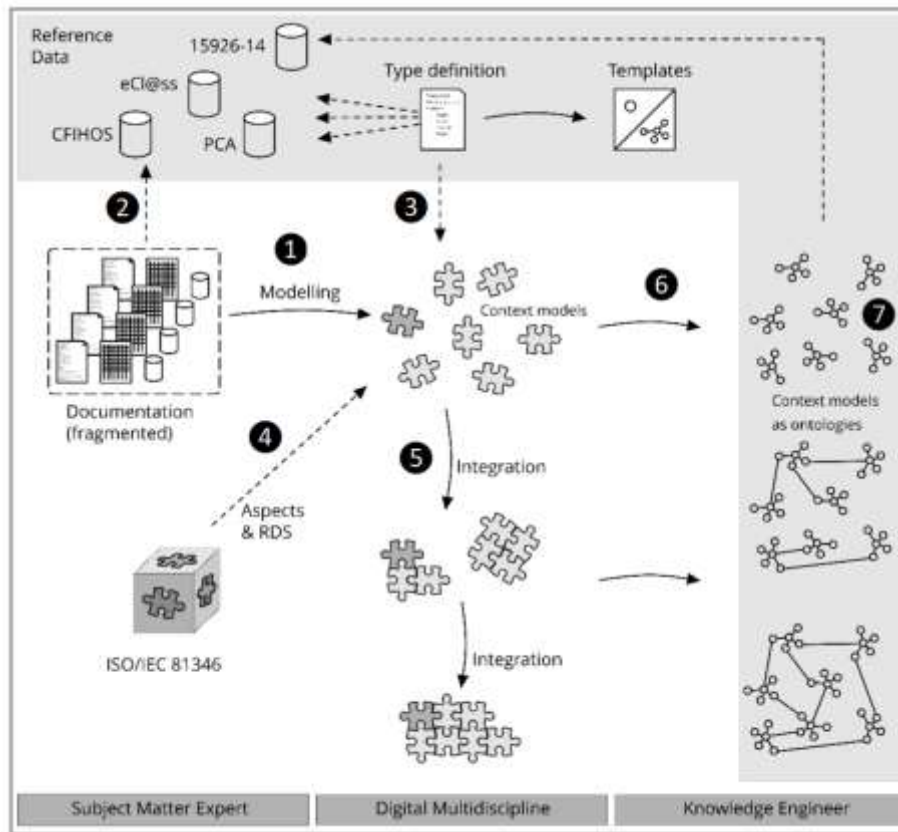


Figure 26: Process to extract Context Models based on IMF

### 3.3.8 Data container

Data virtualisation is one of the key data integration techniques used in the RE4DY digital thread management. It provides access to information via a virtualized service layer, providing a unified view of the data from the different sources and defining a unified way to access the data from heterogeneous sources regardless of their location.

The RE4DY Data Containers (DC) will provide access to the data following the Data as a Product approach while, at the same time, creating an abstraction layer that will hide all the underlying technical complexity from the users. DC will allow the applications to access the data in a trusted and reliable way regardless of the location (Edge or Cloud) and particularities of the data sources. The RE4DY DC will provide the data according to the consumer’s requirements in terms of format and data quality, with the proper considerations about security, privacy, and performance.

A DC has the following functionalities:

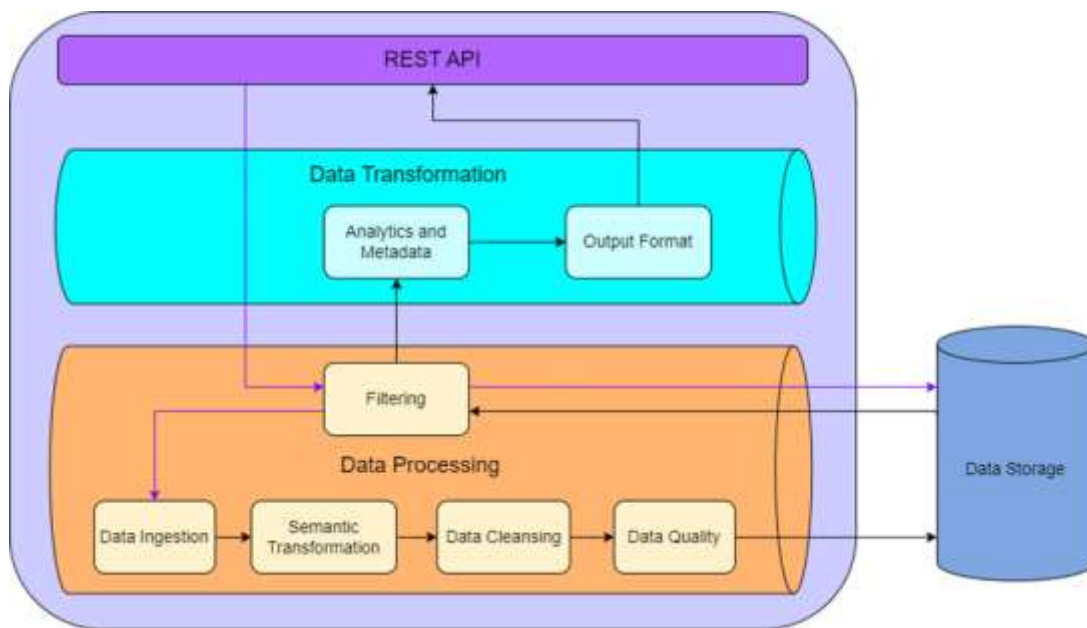
- Enable access to the data following the Data as a Product principles. This includes both historical and (near) real-time data.
- Provision of the necessary metadata to facilitate the use of the data by the upper layers and applications, as well as the implementation of the Data as a Product features.
- Analytics and metadata extraction, which could vary from basic ones (means, sums, variances or standard deviations) to more complex tasks, such as the combination of different datasets.



- Data transformation to a format suitable for the data recipient (for example, Parquet, JSON or CSV).
- Composition of processing techniques in pipelines to adapt the data to the requirements of the consumer.

A DC is composed of the following modules as depicted in [Figure 27](#)

- REST API: it provides access to the data its associated metadata.
- Data Transformation module: it converts the data to the requested format and prepares it to be forwarded to the client.
- Data Processing module: it communicates with the lower components, such as the data storage, to retrieve data based on the request made by the client. This module may perform additional tasks based on the request, such as apply filtering rules to select only a subset of the available data.



*Figure 27: Data Container submodules - example of pipelines for historical data which include calls to other RE4DY components*

Although the RE4DY Data Container will establish a set of filtering rules to allow a further selection of the relevant data by the user and ensure the data is transformed into the requested format, the incoming data will also need to be prepared for the DC itself. This can be achieved through the integration of the DC with other blocks of the RE4DY architecture, which will be responsible for different tasks associated with preparing the data for consumption, such as pre-processing, semantic translation, data cleansing, etc. DC will communicate with the other components of the DaaP Toolkits layer (and Integration Layer) and combine the different steps into a pipeline [Figure 27](#) in order to serve the data according to the application requirements and constraints in terms of quality and format while hiding the complexity of the underlying infrastructure, which will not be accessed directly from external components. More specifically, the RE4DY Data Container objectives should aim to:

- Further processing the datasets, which leads to analytics and metadata extraction.
- Transforming the data to a format more suited for the data recipient.





- Apply filtering rules to the data in order to select the most relevant information for the consumer.
- Integrate with other blocks of the RE4DY architecture in order to provide the data following the DaaP approach regardless of the nature and location (Edge or Cloud) of the data.
- Allow data sharing following the IDSA and Gaia-X specifications. This can be achieved through the integration of the DC with IDSA and Gaia-X components such as an IDS connector.

### 3.3.9 Compliance check and assurance

Within the sphere of a digital thread DaaP solution where data is collected, analysed and provided as a service, quality and compliance are of paramount importance. Compliance check and assurance will be applied to data exchanged through the Data Connection Profiles and Data Containers with the intent of rendering it compliant and compatible with their usage patterns and ameliorating data quality across the data value ecosystem implemented by RE4DY.

The integration of a FAIRness assessment process constitutes one of the primary approaches which will assist in unfolding the real value of the data. The Guidelines on FAIR Data Management in Horizon Europe released by the European Commission encourages the data handled by European-funded project beneficiaries to be compliant with the FAIR principles and procures directions for that purpose.

FAIR data refers to data that adhere to the four principles of being Findable, Accessible, Interoperable, and Reusable based on principles established in Wilkinson et al.<sup>18</sup>, attempting to address pain points in the reuse of research data. The FAIR principles aim to facilitate the use of computational systems to manage data, as the volume, complexity and rate of data creation continue to grow. The emphasis is on machine-actionability, meaning that data should be easily found, accessed, and reused with minimal human intervention.

The FAIR principles released by GO FAIR<sup>19</sup> are enumerated below:

#### Findable

- F1. (Meta)data are assigned a globally unique and persistent identifier.
- F2. Data are described with rich metadata (defined by R1 below).
- F3. Metadata clearly and explicitly include the identifier of the data they describe.
- F4. (Meta)data are registered or indexed in a searchable resource.

#### Accessible

- A1. (Meta)data are retrievable by their identifier using a standardised communications protocol.
  - A1.1 The protocol is open, free, and universally implementable.
  - A1.2 The protocol allows for an authentication and authorisation procedure, where necessary
- A2. Metadata are accessible, even when the data are no longer available.

<sup>18</sup> <https://doi.org/10.1038%2FSDATA.2016.18>

<sup>19</sup> <https://www.go-fair.org/fair-principles/>





### Interoperable

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (Meta)data use vocabularies that follow FAIR principles.
- I3. (Meta)data include qualified references to other (meta)data.

### Reusable

- R1. (Meta)data are richly described with a plurality of accurate and relevant attributes.
- R1.1. (Meta)data are released with a clear and accessible data usage license.
- R1.2. (Meta)data are associated with detailed provenance.
- R1.3. (Meta)data meet domain-relevant community standards

In essence, the FAIRness principles define a number of optimal qualities data should have in each step of interaction with it. Locating the data is the first step. Therefore, to enable discoverability to both human and computer users, data and metadata should be uniquely identified and located. Furthermore, it is imperative to have machine-readable metadata to enable automated discovery of datasets and services, as this is one of the cornerstones of the FAIRification procedure. Once the desired data has been identified, the user must be able to learn how to access it, including the use of authentication and authorization protocols, if necessary. Next, it is typically necessary for the data to be integrated with other relevant data. In that case, it must be interoperable with other applications or workflows for storage and analysis. Finally, optimizing data reusability constitutes the primary objective of FAIR. In order to accomplish this, it is essential that metadata and data are comprehensively described to enable their replication or combination in diverse settings.

Adhering to the FAIRness principles is vital in data-driven environments and especially in a DaaP context where the value of the platform is intrinsic to the data it provides. Therefore, it is crucial for data to remain findable, accessible, interoperable, and reusable in an ecosystem composed of diverse, interconnected elements, where it is processed and reused by various stakeholders, both humans and machines. In the context of a sovereign data space, data FAIRness holds substantial importance as it promotes data interoperability through common models, facilitates the data exchange process and results in superior data quality.

It is worth mentioning that the sheer volume of circulating data sets within the ecosystem warrants the development of an automated FAIRness assessment procedure. This procedure is to be transparently integrated into the implementation of the platform in order to further optimize all stages of data use and reuse. Ensuring that automatic data FAIRness assessment is put into practice within the domain of a data-oriented infrastructure with diverse data sources and formats will enhance the ease of discovery, access, and harmonization of data throughout the ecosystem.



### 3.4 Integration layer

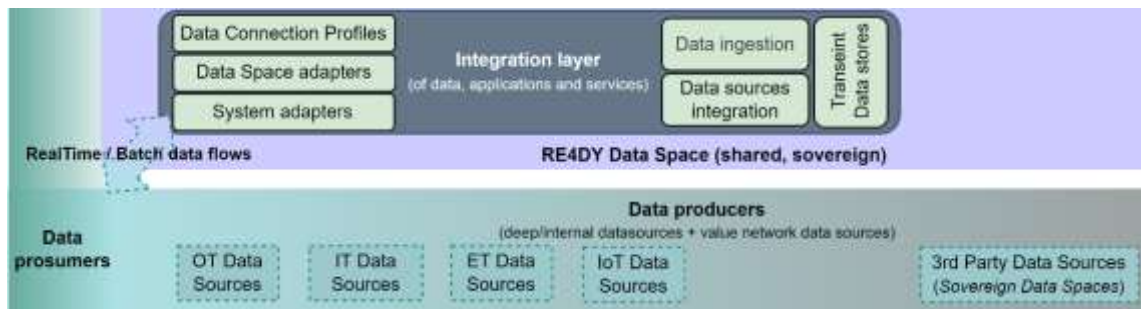


Figure 28: Integration layer building blocks

This layer is dedicated to the integration of different type of data, which can flow in both directions from and to the upper layers of the architecture as well directly from the shopfloor, even externally to the enterprise, in a transparent way, both batch and real-time. It represents conceptually the outwards interface for the DaaP, namely the contact point of the enterprise Data Space(s).

Data producers and consumers:

Data generated by data producers include data at different levels of elaboration, from operational to engineering technologies; for example, operational data over a period of time, (internally to the factory as well as externally, depending on the cases) that can be coupled with those from design phases, or from PLM models with which to allow the building of even more precise digital twins and thus new data containers. These data containers, which implement -through the above layers- the DaaP RE4DY main concept, are precious both internally for the factory (Digital Thread) and externally for the market, when properly secured in order to assure data sovereignty and guarantee trusted data spaces.

Examples of data sources from the data producers are:

- (OT) Operational technology Data Sources: shopfloor product and/or process data.
- (IT) Information Technology Data Sources: CRM Data, ERP data, Decision support system data and other legacy data.
- (ET) Engineering Technology Data Sources: product design data, process design data.
- (IoT) Internet of Things Data Sources: data from product in the field, sensors data (e.g., machinery, environment) and product line data.
- 3rd Party data sources: data form external providers, data from customers and data from external data space end points.

The integration layer is formed by six building blocks which are listed and described as follows:

1. Data Ingestion and ETL Services
2. Data sources integration
3. Transient data stores
4. Data Connection Profile
5. System Adapters
6. Universal connector

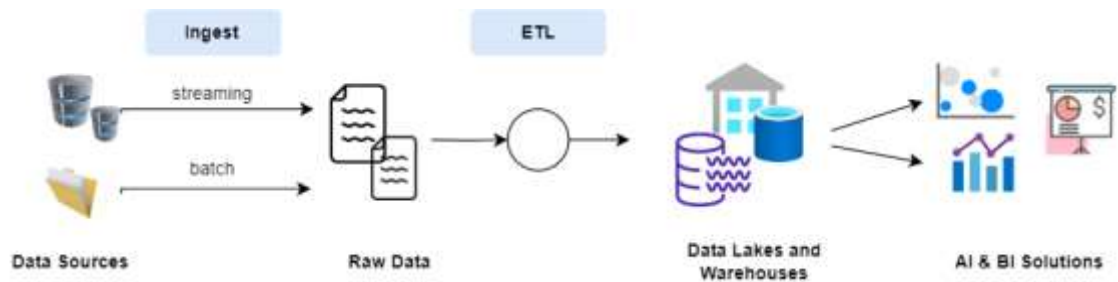


### 3.4.1 Data Ingestion and ETL Services

Data Ingestion is the process of importing and loading data into a system. Generally, it is considered as one of the first steps that should be followed for using any data for further analysis and application of AI/ML services on them. In particular, it is the process of introducing data into a database or other storage repository. Often data ingestion involves ETL (extract, transform, load) tools to move information from a source system to another.

Data Ingestion can be: (a) Real-time ingestion that involves streaming data into a data warehouse in real-time. Examples of these types are mainly cloud-based systems that can ingest the data quickly, store it in the cloud and share it to various users. (b) Batch ingestion related to collection of large amounts of raw data from various data sources (that should be processed all together at once) to process it later.

Data ingestion processes increase reliability and accuracy regarding data and boost flexibility & simplicity and speed regarding data analysis procedures. To provide the previous mentioned advantages, data ingestion services have to tackle with challenges like data quality, latency, security, data capturing and collection services and maintenance.



*Figure 29: Data Lifecycle from Raw Data to Transformed Data*

RE4DY will consider all challenges in order to deliver data ingestion services with special focus to be given in advantages like flexibility and simplicity by combining ingestion with ETL services so to restructure enterprise data to predefined formats and makes it easier to use by project tools that are capable of processing a range of standardized and widely used data formats.

ETL, which stands for Extract, Transform and Load, lies between data ingestion and integration processes as it combines data from multiple data sources into a single, consistent data store that is loaded into a data warehouse or data lake etc .

RE4DY is going to combine open-source ETL tools such as Jolt with data sovereignty and integration services so to provide data lifting and transformation as-a-service available to various tools in its data fabric targeting to Data-as-a-Product delivery.

### 3.4.2 Data sources integration

Data integration refers to the process of bringing data from disparate sources together to provide users with a unified view. This process aims to reduce IT costs, free-up resources and improve data quality. A set of services is required for integration system such as pipelines that can automatically move, consolidate, and transform (big) data from multiple data sources while maintaining lineage. Various challenges arise regarding scalability, performance, data quality and real-time integration of continuously streaming data.



The task of disparate data sources integration is strictly connected with data ingestion and ETL services that have been analysed in the previous section. RE4DY will combine this type of services by using integration engines like Apache NiFi<sup>20</sup> by following and extending the work done in previous research projects such as EFPF Data Spine<sup>21</sup>.

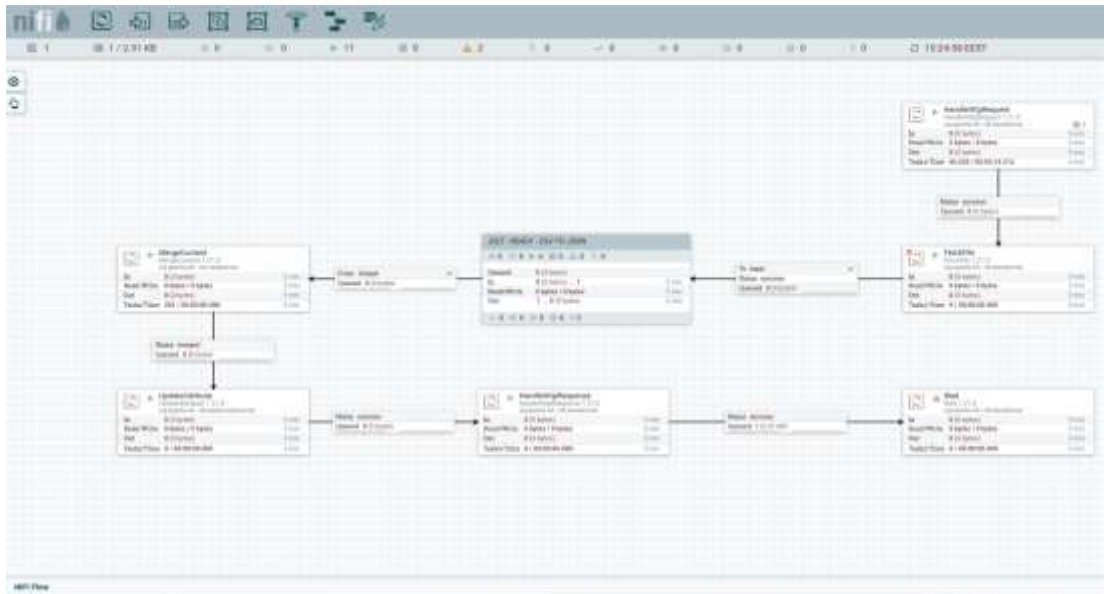


Figure 30: Apache NiFi Flow example regarding Data Transformation

As mentioned before the data integration services in RE4DY will be coupled with ETL services to deliver data transformation across the project data fabric. NiFi will be coupled with tools like Jolt for data transformation. Already various processors have been tested in a RE4DY deployed instance of Apache NiFi and Jolt was used as well. An integration flow that has been created in the project’s T3.3 for handling an http request for transforming a csv from a data source to a JSON format to be consumed by an analytic tool is presented in the previous figure.

In addition to the purely technical challenges of data integration, proposed to be solved thanks to the aforementioned tools, a severe issue for the data interoperability is represented by the interpretation of the data themselves. This issue is crucial for most of the system integrators in manufacturing sector, whenever they are required to connect an equipment asset to a software platform. In most of the manufacturing companies, indeed, data available are seldom documented in an interoperable way, meaning that data flowing to software are sent as pure data, without metadata payload to contextualise their production.

In order to mitigate this issue, the usage of AAS packages is forecasted in the RE4DY activities, considering the transmission of a one-off package comprehensive of all the information of the physical asset producing data. This package is supposed to contain all the relevant information about the asset itself, as well as the information about the streamed data, such as data type, URI and text descriptors for each variable.

<sup>20</sup> <https://nifi.apache.org/>

<sup>21</sup> <https://www.mdpi.com/1424-8220/21/12/4010>



As demonstrated in the DIMOFAC project, the integration of AAS as an information container allows a dramatic reduction of configuration times. In RE4DY, this approach is being extended to different applications, as well as the data model paradigm is being extended in order to be able to include different types of assets, such as software tools.

### 3.4.3 Transient data stores

The *Transient Data Store* is a passive component of the architecture where transient representations of data objects are stored temporarily during the execution of the algorithms or data pre-processing. The main characteristic of transient data stores is their short lifespan. The data stored in these systems is typically temporary and disposable, meaning it is not meant to persist for a long time and can be discarded after its immediate purpose is fulfilled. Transient data stores are often contrasted with persistent data stores, such as databases or file systems, which are intended for long-term data storage.

Transient data stores offer flexibility, speed, and scalability for various computing scenarios. However, it is important to note that since the data stored in these systems is temporary, it can be lost if not properly handled or replicated. In some cases, if a system outage or reboot occurs, the data stored in this type of storage is likely to be lost. Therefore, they are best suited for data that can be regenerated, reconstructed, or is not critical for long-term persistence.

Due to their different characteristics, transient data stores and regular data storage serve distinct purposes in the RE4DY architecture. Transient data stores excel at fast access, temporary data storage, and real-time processing, while regular data storage systems provide durability, scalability, and long-term data retention. By combining both types of storage, RE4DY architecture can achieve high-performance data processing, data durability, fault tolerance, scalability, and cost optimization, catering to the diverse needs of data-driven applications and workflows.

Transient data stores have many possible uses, some of which are:

- **In-Memory Caches:** These can significantly improve performance by reducing the need to fetch data from slower persistent storage by storing data in in-memory caches, which are high-speed storage systems that hold frequently accessed data closer to the application or service for quick retrieval.
- **Message Brokers:** Transient data stores in combination with message brokers can allow for the temporary storage and exchange of messages between various components or systems in a distributed architecture. They provide reliability and scalability while facilitating real-time data processing.
- **Caching Layers:** Transient data stores can be employed as caching layers in front of slower data sources, such as databases or APIs. Caches store frequently accessed data, reducing the load on the underlying systems and improving response times.

### 3.4.4 Data Connection Profile

The Data Connection Profiles (DCP) are a fundamental element to provide a trusted computed continuum where the data can be shared in a harmonised and simple way. The DCP aim to provide the means for achieving an integration among complex systems in a standard, simple, durable, and context-sensitive way. DCP will provide a standardized description of how the data is going to be accessed and used by other RE4DY components



in the use cases to ease the integration and avoid the need for custom development for each possible actor in a use case. The DCP will be defined on a use case basis, and they will provide information to facilitate the use of the data (either from the sources or from a DC), such as the description of the purpose of the data, its granularity and data quality considerations, while also addressing IP and privacy needs. To this respect, the Data Containers implementation will be based on a set of common DCP that will describe how to communicate with the sources to retrieve the data. Similarly, the components of the upper layers will make use of the DCP to access the corresponding DC. Moreover, it will be possible to define a DC to provide a combination of data from different sources, which could be other DC.

DCP will extend the Open Group's IoT standards O-MI<sup>22</sup> and O-DF<sup>23</sup>, which were initially defined in the context of H2020 Project bloTope<sup>24</sup> to enable interoperability among devices and applications within an ecosystem. O-MI enables the definition of an interface (as open API) and O-DF provides complementary data model to describe the information using a generic structure. O-DF is described using XML schema (although it is also possible to use JSON) and was designed to be generic enough to describe any object or information needed for data exchange. It is structured hierarchically as an object tree with any number of levels and properties.

The information provided by a DCP should refer to:

- The data sources, considering the type of data they provide (historical or real-time).
- The exposed APIs to access the data, as well as the output data format and data model.
- Properties to facilitate the use of the data, for example, purpose of the data, data quality, granularity (for accessing the data) or parameters related with QoS (latency, availability, bandwidth).
- How the data from the data sources needs to be processed before making them available through the DC.
- Licensing, privacy, integration with Data Governance or definition of data usage policies.

In addition, to facilitate the integration of the RE4DY assets in a dataspace, the DCP will be aligned with IDSA data asset model and DIN SPEC 27070 trusted information gateway model adopted by Gaia-X.

### 3.4.5 Data Space Adapters

The Data Space Adapter building block is essential for enabling trust and interoperability in data sharing and exchange within data spaces, which are designed to provide data sovereignty. Data spaces, and with that, data sovereignty will be the level playing field on a global scale. This represents a significant advantage and revolutionizes the data economy of the future – with the goal of benefiting society, businesses, and individuals.

Technical Interoperability is a major requirement in data spaces. It should be realized by data connectors, based on specifications and standards rather than relying on singular

<sup>22</sup> <http://www.opengroup.org/iot/omi/index.htm>

<sup>23</sup> <http://www.opengroup.org/iot/odf/index.htm>

<sup>24</sup> <https://biotope-project.eu/>



implementations or reference implementations. To do that, multiple levels of interoperability must be addressed: first, the general interaction between the connectors for the description of data assets and the related endpoints must be addressed including the definition of policies for access control and usage control, followed by the negotiation of those policies and contracts. The initiation and management of the data exchange process needs a clear specification, which can be mapped then to tangible protocols, like https, MQTT, web sockets or others. This is the handover to use case specific, domain-specific or ecosystem-specific definitions and standards. General interactions require a robust standard that can be implemented by the different connectors, while the subsequent data exchange makes use of domain or use case-specific standards. The same applies to semantic interoperability, which can be achieved on the foundation of the Data Catalog Vocabulary (DCAT)<sup>11</sup>. The further definition of the data exchanged is handled by semantic models, taxonomies, schemas or other similar mechanisms, the so-called “vocabularies”.

To achieve robustness and reliability in a data space, the interoperability of connectors requires verification. Based on standards and specifications, compliance to those can and must be continuously evaluated to maintain this foundation in addition to the continuous management and verification of security aspects related to the data connectors.

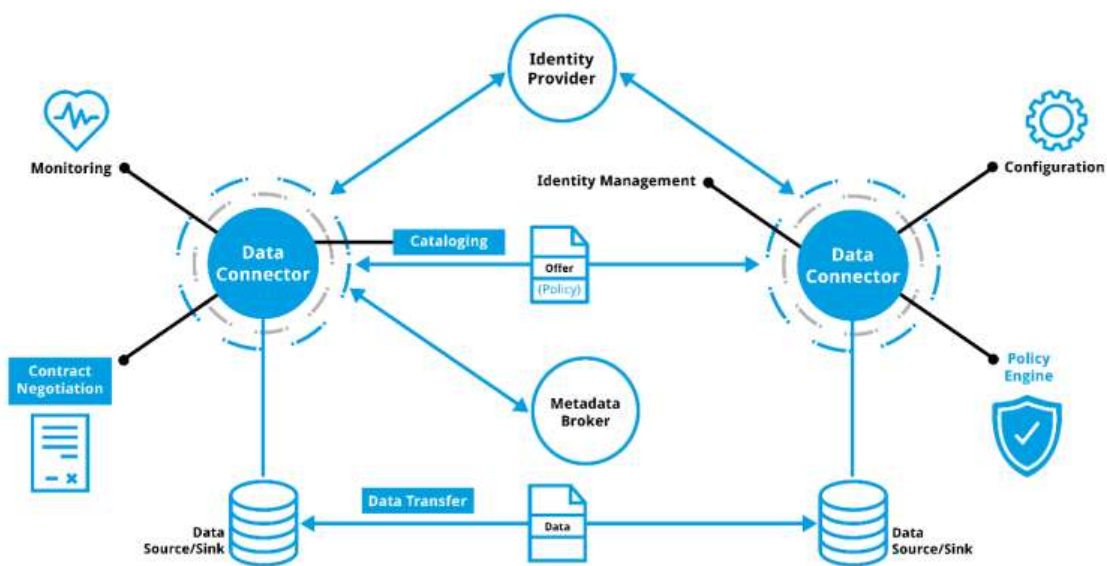


Figure 31 Data Exchange Services realized by a data connector

Today there is already a set of usable standards to achieve the goals described above, but additional standards are required. The interaction of the connectors on the general level, as described in Figure 31, requires a protocol agnostic standard as foundation for interoperable data spaces. For this reason, some of the organizations involved in the development of the different Reference Models and Architectures are working on a specific Dataspace Protocol, a set of specifications designed to facilitate interoperable data sharing between entities governed by usage control and based on web technologies. These specifications define the schemas and protocols required for entities to publish data, negotiate usage agreements, and access data as part of a federation of technical systems termed a data space. The Dataspace Protocol, therefore, represents the foundation for technical interoperability in data spaces.





### 3.4.6 System adapters

The concept of System Adapters within the "Integration layer" in RE4DY reference architecture refers to the components or interfaces that enable seamless integration and interoperability between different systems. These adapters play a crucial role in facilitating data sharing, data transfer, and data transformation while ensuring confidentiality and privacy.

System Adapters serve the purpose of establishing connections and bridging between various systems, particularly legacy systems. They enable data exchange and collaboration, allowing organizations to securely share and access data across different platforms and ecosystems.

System Adapters also play a crucial role in data transformation and integration by facilitating the mapping, conversion, and harmonization of data between different systems and formats. They handle data normalization, validation, and enrichment processes to ensure data consistency and quality during data transfer or integration. System Adapters may implement data transformation rules and workflows to support data harmonization.

In addition, they can also enforce encryption mechanisms, access controls, and data protection measures to ensure that sensitive information remains secure during data transfer. System Adapters may implement authentication and authorization protocols to verify the identities of data recipients and ensure that only authorized parties can access the data.

In that regard, leveraging on Data Containers (introduced in §3.3.8), the System Adapters can encapsulate all these underlying complexities and facilitate the interoperable distribution and integration of data by using standardized format and APIs that can be easily shared and consumed by other systems or stakeholders.

## 3.5 Computing/Networking continuum



Figure 32: Computing/Networking continuum dimension

Cloud computing has dominated the arena of IT world during the past decades. Despite having faced some challenges, it has remained the reference in most modern IT deployments. However, the advent of the new generation of internet, characterized by ever-increasing demands of bandwidth, latency and computing power has brought to light some flagrant shortcomings of this paradigm.

In the last years, the development of edge computing has placed data processing and storage capabilities closer to the edge devices. This way, data can be processed closer to the sources, which results in increased data privacy and security, as well as better performance and less network usage and latency.



Recent advances have seen the development of more intelligence devices, capable of applying on-device processing combined with the production of federated AI architectures across devices. This is combined with intelligent and programmable networks, development of cognitive cloud systems and advances in orchestration across different cloud environments which has resulted in the device to cloud continuum.

The combination of edge and cloud resources and services, known as the digital continuum (or computing continuum), is changing the technological landscape. It is the combination of resources and services at the centre of the network (cloud), at its border (edge) and in transit (fog). Data is generated and pre-processed at the edge, partially processed by intermediate nodes and, if necessary, transferred to the cloud.

The structure of edge-cloud continuum computing (*Figure 33*) can be separated in three main levels:

- The front-end level encompasses various endpoint devices, including sensors, actuators, mobile devices (such as smartphones, tablets, and wearables), personal computers, and other data-generating elements like cameras and Bluetooth devices. In most deployments, these components are not equipped to handle extensive computing requirements, if any at all. Consequently, their primary role within this layer is to act as data forwarders, transferring data from the source to subsequent stages of the system for processing and analysis.
- The near-end level plays a crucial role in the context of edge computing. It serves as a significant aspect of the paradigm by executing specific tasks aimed at reducing traffic towards the cloud and attaining the benefits mentioned earlier. Typical tasks assigned to this level include filtering, pre-processing, aggregation, content caching, device management, and privacy protection. On the other hand, the back-end level represents the traditional element centered around cloud computing (CC). This is where most of the bulk processing, storage, and analysis of data takes place.
- The back-end level pertains to the traditional aspect of cloud computing, focusing on centralized processing and storage operations. This layer, which can be located in a single remote site or distributed across multiple locations, serves as the primary host for the main workload associated with cloud services. In certain cases, this level may also act as a centralized controller within the edge computing environment, overseeing, and coordinating various tasks and resources.



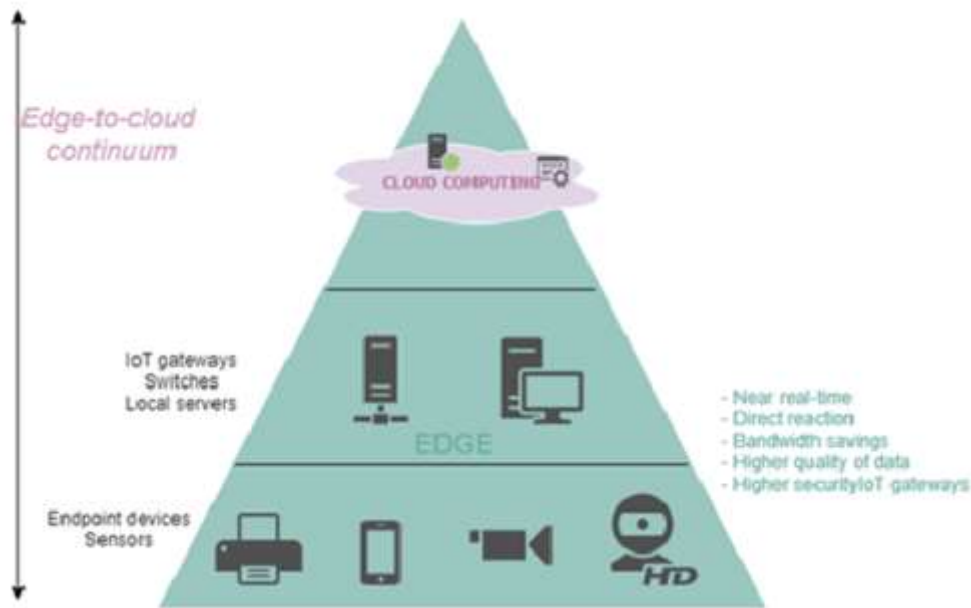


Figure 33: Edge-to-cloud computing continuum approach and its advantages (C. Avasalcai, 2020)

The networking infrastructure is also key to achieve the digital 4.0 continuum since it provides the capacity to efficiently interconnect data sources that are hosted on the cloud, edge or endpoint devices and systems as well as large networks of intelligent digital twins. Proper planning and commissioning of the networking infrastructure (based on wired and wireless communication technologies) is then necessary to ensure performance in terms of capacity and latency among others and achieve the digital 4.0 continuum.

Current approaches for the planning and commissioning of factories mostly consider the computing and networking infrastructure as external and independent to production systems. However, the computing and networking infrastructure must be considered as an additional asset of the factory, and as such, the planning and commissioning of the computing and networking infrastructure must be done jointly with the rest of the factory's assets with the aim of achieving production targets. The Computing/Networking continuum layer of the Re4dy RA will then include the necessary tools and solutions for the joint planning and commissioning of factory and digital twin fabrics with the underlying computing and networking infrastructure which is critical to achieve the digital 4.0 continuum.

One key aspect for the design of these tools and solutions is the integration of the networking infrastructure in the digital twin fabric of the factory. In this context, the Computing/Networking continuum layer of the Re4dy RA will also contain the Asset Administration Shells (AASs) of the networking infrastructure that are being developed in the project (in particular, for 5G networks). The AASs are used to create fully digital versions or digital twins of a factory's asset (Bader (SAP), 2022)[ref] and it facilitates the integration of the communication network into a factory management and control system (5G-ACIA, 2021) [ref]. The definition of the 5G AAS is being based on the indications and specifications given by 5G-ACIA and Platform Industrie 4.0. The availability of digital twins of the computing and networking infrastructure will allow the development of innovative digital continuum planning toolkits that integrate and jointly plan the interactions and



dependencies between the physical, computing and networking, and intelligence and data domains, which is key to provide the capacity to efficiently and resiliently support a digital 4.0 continuum.

### 3.6 Distributed, trustworthy and secure computing toolkit

The distributed trustworthy and secure computing toolkit will support the implementation of data provenance and traceability within the RE4DY infrastructure. In the context of RE4DY, data provenance and traceability involve mechanisms for tracking data flow and maintaining a secure and auditable record of transactions. These mechanisms will enable the capability to generate provenance records for each piece of data that has been provided, consumed and processed, and provide the means to create a chronological documentation of events regarding this data. A typical use case would be to trace the origin of a product or its materials as well as their chain of custody within the supply chain.

Distributed ledger technologies constitute a highly suitable building block for developing a data provenance and traceability framework since a significant number of their features are highly desirable. In particular, they offer immutable and append-only storage by maintaining a complete and auditable record of all transactions taking place on the network. Transactions, when processed, are able to modify the state of the network, which is observable by all network participants. Furthermore, the decentralized nature of blockchain enables high data availability due to eliminating single points of failure. However, blockchain is not only utilized as secure and transparent storage. Certain blockchain platforms support smart contracts which allow the secure execution of custom business logic which may update the network state. In this manner, secure computing is achieved without the need for any intervention from third parties, a quality which satisfies the transparency and auditability requirements of a data provenance and traceability framework.

Although the role of blockchain in the distributed trustworthy and secure computing toolkit is prominent, the aspect of identity management is also of notable significance with regard to the emerging security and privacy requirements within a distributed system composed of undetermined participants, both individuals and organizations. Through identity management and the establishment of roles and rights within the system, authentication and authorization may be implemented. Therefore, an access control plan may be formulated to protect sensitive information from unauthorized access, ensure the integrity and security of the data in question along with providing essential non-repudiation capabilities. Additionally, based on the confidentiality level of the stored data, the inclusion of encryption mechanisms will ensure that only authorized parties may decrypt and access it. The aforementioned components are to be integrated with the data provenance and traceability framework based on distributed ledger technology.

Concerning identification of stakeholders, particularly natural persons, the usage of self-sovereign identity (SSI) is of notable mention. The European Union has capitalized on the establishment of eIDAS<sup>25</sup>, a regulatory framework for a unified digital identity for individuals among all EU member states, in order to implement the European Self-Sovereign Identity

<sup>25</sup> <https://digital-strategy.ec.europa.eu/en/policies/eidas-regulation>



Framework (ESSIF)<sup>26</sup>. ESSIF is compatible with eIDAS and offers users the advantages of a digital identity and the ability to control how the data linked with their identity is shared to web services and applications during the authentication process. To achieve this, ESSIF utilizes decentralized identifiers (DIDs)<sup>27</sup> along with the European Blockchain Services Infrastructure (EBSI)<sup>28</sup>.

To conclude, data provenance and traceability represent essential features of the distributed trustworthy and secure computing toolkit which will aid stakeholders in verifying and validating the lineage of data. Leveraging the features and capabilities of a technology stack consisting of distributed ledger technology reinforced with additional security and privacy standards which enhance integrity and transparency and foster trust, will ensure all requirements for establishing data provenance and traceability within RE4DY are fulfilled.

### 3.7 xCDTOps framework

The cognitive digital twin operations framework contributes guidelines and methods in order to provide a framework with which digital twins can be generated and deployed. With the purpose to tackle the engineering challenges, actionable insights from analytics at the edge to the cloud are necessary. Bringing Operational technology (OT) and IT closer together in industrial production, a seamless flow of data from the field level to the cloud can be ensured.

Enabling the AI to detect and evaluate business-relevant anomalies, the machine-learning algorithms are trained on the basis of process data and then concentrated to determine which anomalies have an impact on the economic efficiency of the plant. The plant operator themselves then defines the further focus of the AI using the app dashboard, where anomalies can be selected, evaluated and commented. This evaluation phase is accompanied by several feedback loops, so that the plant operator ends up with well-trained, focused AI that is able to evaluate anomalies, based on the process data, for their business relevance.

<sup>26</sup> <https://ssimeetup.org/understanding-european-self-sovereign-identity-framework-essif-daniel-du-seuil-carlos-pastor-webinar-32/>

<sup>27</sup> <https://www.w3.org/TR/did-core/>

<sup>28</sup> <https://ec.europa.eu/digital-building-blocks/wikis/display/ebsi>



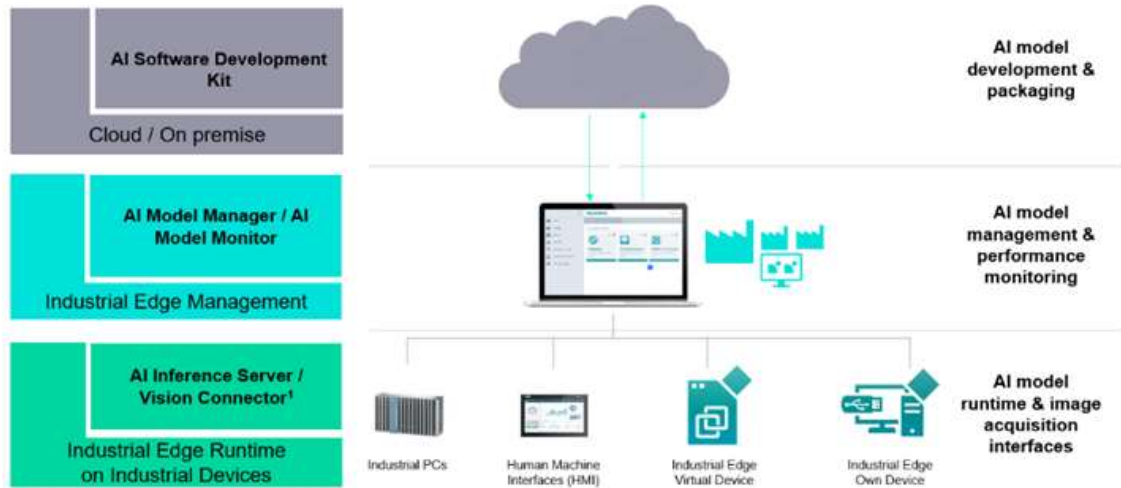


Figure 34: Seamless automation, deployment and monitoring from cloud to shopfloor

When implementing and setting up the AI model in an operational framework, there are some challenges to overcome to avoid the failure of an advanced project. The difficulty of extracting relevant and labeled data from shop floor must be considered. Managing and tracking multiple assets, datasets and models is also a challenge. Similarly, incomplete versioning of models, data, and software complicates transparency and traceability. Continuous evaluation of model performance in different configurations presents an additional challenge.

The AI lifecycle provides an eight-step framework for data scientists and automation engineers to set up and deploy digital twins. The eight-steps include plan, model, validate, deploy, sense, infer, monitor and retrain. The Figure 35 shows the framework and the involved steps. Below the eight steps are described more in detail.

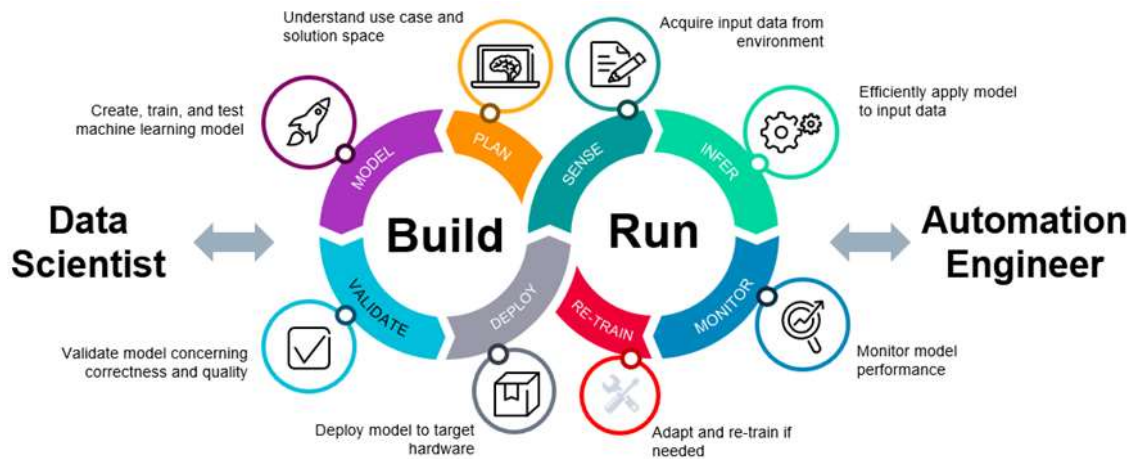


Figure 35: Industrial AI lifecycle: building and operating AI models on scale

Plan

- Understand use case and business aspects (e.g., added value of artificial intelligence).
- Elucidate key requirements and collect initial data.
- Define KPIs and metrics.



- Define software architecture and used technology stacks.
- Identify risks and potential showstoppers.

#### Model

- Select suitable method (e.g., neural network).
- Acquire and (optionally) label input data.
- Create and implement machine learning model.
- Train model using given data.
- Perform local testing.
- Optimize model according to predefined metrics.

#### Validate

- Run automated tests, optionally in virtual environment.
- Determine key metrics (e.g., accuracy, precision).
- Identify regressions w.r.t. given metrics and (optionally) inference performance.
- Re-train dependent models.
- “Training data coverage”.

#### Deploy

- Generate optimized deployable artifact (e.g., ONNX file, container).
- Update version information and initiate approval.
- Upload package to binary repository (e.g., artifactory).
- Download to device(s).
- Install and start as part of lifecycle management.

#### Sense

- Acquire data (e.g., by sub-subscription to sensor / video stream).
- Perform necessary pre-processing steps (e.g., FFT).
- Merge data from multiple sources (sensor fusion).
- Filter and select relevant inputs for inference.

#### Infer

- Analyse consolidated input data using deployed model.
- Generate output data based on inference results (e.g., classification).
- Requires appropriate frameworks and potentially hardware accelerators.

#### Monitor

- Collect key performance indicators (e.g., confidence levels) according to predefined metrics.
- Trigger events based on custom rules (e.g., thresholds).
- Send statistics to backend including selected input data (e.g., images with poor classification results).





The Figure 36 describes tasks along the complete lifecycle of AI solutions. Those tasks are part of previous described steps from Figure 35. A detailed description of the tasks is given below.

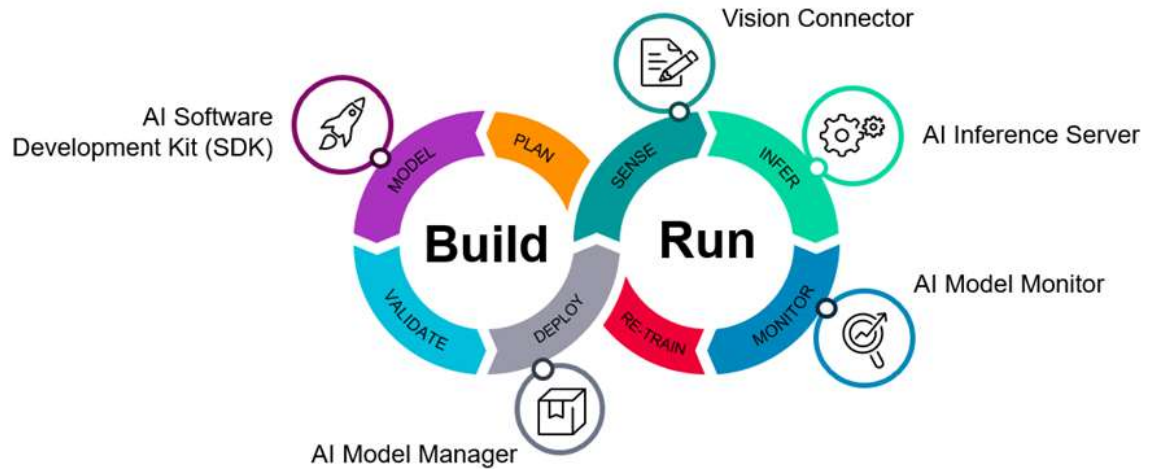


Figure 36: Lifecycle of AI solutions



### AI Software Development Kit (SDK)

Extend your existing ML design and training environment with our Python Software Development Kit

- Package your model into standard inference pipeline running on AI Inference Server using our Python SDK.
- Test and validate your package locally prior deployment.
- Create projects based on standard use case templates (image classification and time series).
- Keep using your existing Python coding and training environment (Visual Studio Code, JupyterLab, Azure ML Studio).
- Integrate packaging into your CI/CD or MLOps environment (Azure ML, AWS Sagemaker, GitLab CI + DVC).

### AI Model Manager

Keep all AI models up and running

- AI Model Manager lets you keep track of all AI model deployments on Edge.
- Configure automated model download and deployment from cloud of AI providers (Google VertexAI, AWS SageMaker, Azure).
- Deploy existing models to multiple Edge devices and keep track of versioning.
- Customize model deployment configuration to each Industrial Edge Device you deploy to.

### Vision Connector

Integrate industrial Cameras to Industrial Edge

- Vision Connector allows Industrial Edge apps to connect to multiple industrial cameras via standard protocol 'GENeric Interface for CAMeras' (GenICam), making image or video data available via IE Databus and ZMQ.
- It allows low throughput data sharing via Industrial Edge Databus and a higher throughput alternative to stream data to other apps running on Industrial Edge.

### AI Inference Server

Run AI Models directly on Industrial Edge:

- AI Inference Server is a ready-to-install Edge App that can instantiate AI models via configuration packages.
- AI Inference Server orchestrates the different AI models and inference pipelines.
- AI Inference Server provides also a graphical representation to monitor the AI solution.
- Basic Python Runtime that can be extended with customer specific needs or runtimes (e.g., TF Serving).
- Data ingestion can be mapped from IE data bus leveraging the rich set of IE southbound connectors.

### AI Model Monitor

Make sure AI models are delivering value continuously:



- AI Model Monitor is an integrated module for AI Model Manager.
- Keep track of inference performance, detecting HW and SW issues.
- Evaluate data quality metrics and potential data drifts.
- Check if model outputs are meeting expectations, detect model performance degradation in advance.
- Receive automated alerts whenever your model or data are not behaving as expected.

## 3.8 Federated Learning and Analytics framework

The Federated Learning and Analytics framework aims at offering a set of cross-cutting guidelines and tools with which to support the development of federated machine learning based applications and to ease the implementation of AI pipelines.

Federated Machine Learning (FML) refers to the ability to train a machine learning model in a geographically distributed environment by leveraging local datasets and without having them moved outside the data owner's domain of control. Rather, it is the algorithms that are deployed on the data owner's premises.

Federated Machine Learning services allow for such distributed training through the deployment and execution of two types of modules: *Aggregator* and *Participant*. At the beginning of each round, a copy of the global Machine Learning model is sent to each of the *Participants* and there trained on the local datasets. At the end of the round, all the *Participants* send their locally updated copies of the model to the *Aggregator* where they are aggregated to update the global model. The resulting global model is then sent back to the *Participants* for the next round of local training. The process continues until a certain convergence criterion is reached.

The *Aggregator* typically runs on cloud - although in some scenarios it can be found on local premises - and coordinates the *Participants*, the exchange of the model updates and the overall Federated Training process throughout its multiple rounds. The *Participants* must be able to interact with the local data source and with the *Aggregator*. Multiple *Participants* never interact directly with each other but always with the *Aggregator*. For the realization of FML systems, it is possible to rely on ready-to-use Federated Machine Learning frameworks offering the base (customisable) logic for the coordination of *Aggregator* and *Participants* as well as of the training procedure. Examples of such frameworks are Flower and TensorFlow Federated.

There are various strategies to follow for the aggregation of model updates of the *Aggregator*, such as federated averaging, secure aggregation, and differential privacy, which can enhance the privacy, efficiency, and performance of the learning process. More specifically:

- Federated averaging runs on multiple participants and periodically averages the developed ML model across them.
- Secure aggregation is built out of distributed trust across multiple *Aggregators* that keep individual participant updates private as long as one server is honest, and in this way can defend against malicious clients while remaining efficient.



- Differential privacy is the methodology of adding noise to the participant updates to further protect the privacy of the data.

Although Centralized Machine Learning usually performs better than Federated Machine Learning, the benefits of the latter are manifold. Firstly, by not having to move datasets outside a data owner's domain of control, privacy is respected and, as a result, third-party companies are more willing to share their datasets, being reassured that no information about specific data samples is leaked. Secondly, not moving datasets also makes it more sustainable to tap into a larger pool of datasets each of which may also feature a larger number of data samples. In addition, the cost for transmitting data is vastly reduced, especially in cases where the volume is high, and the data is transferred across different regions. Thirdly, processing speed is increased due to the parallel computational nature of the concept, the execution of local model updates and the distribution of the computational load across all aggregators.

The Reference Architecture presented in this document does not preclude the extension of FML to its super case of Federated Analytics (FA). In this case, machine learning is achieved through a flexible approach where the Data Scientist can run different algorithms on different local nodes, compute some custom features and then transfer them to a cloud node for further processing.

Regardless of it being FML or FA, a service should offer a set of features aiming at making it easier and more effective for Data Scientists to adopt federated learning and solve AI tasks. First, the services should offer, wherever possible, support to multiple Machine Learning frameworks in order to allow the reuse of existing code, take advantage of the latest cutting-edge ML libraries or simply allow the user to work with the framework of their choice. In addition, the Data Scientist should also be able to fully customise and tune the involved modules, for instance the Aggregation logic or the centralized processing unit. The modules that run locally should be able to interact with a wide variety of data sources - even legacy - and process them. Furthermore, given the distributed nature of Federated Learning, it is also very important to be as independent as possible from the characteristics of the final execution environments, which can be very heterogeneous. To this end, the use of containerization technology is highly recommended for the deployment of the software components on a wide range of underlying hardware and software systems. Containerization also fits well with orchestration strategies able to speed up the deployment and update of AI assets.

To make Federated Learning services accessible to Data Scientists or other users with no or little programming experience, it is also recommended that the platforms offer visual pipeline editors allowing for the creation of end-to-end pipelines without writing code. To this regard, nowadays more and more platforms offer integrated tools that enable the creation of pipelines by combining a number of reusable blocks through graphical user interfaces. Moreover, when needed, it should also be possible to create these blocks from scratch and catalogue them for later reuse.

Added value can be given by an integrated orchestrator allowing for a seamless deployment of the developed AI pipelines on a variety of target environments. The advantages of such a mechanism are manifold. First, it allows users with little experience to easily manage the pipeline, from development to operation, through a graphical user interface. Moreover, it also allows for the monitoring and control of the assets running on both cloud and edge without having to directly access the specific nodes. Last but not



least, the visualization features. These enable users to visualise information about the training of the model - thus getting insights into the quality of the resulting model itself - but also to view the results of a prediction. Visualization features can include customisable dashboards, charts but also more traditional reports that can be exported and shared.

Finally, it is also essential to investigate the relation between Federated Learning and Data Spaces. Let us describe an example showing how FML can be employed on top of IDS.

The typical IDS use case involves the physical exchange of datasets between two or more parties interacting through IDS connectors. In this context, it is critical that the data owner always maintains control over the shared datasets and how they are used. This is achieved through IDS Usage Control Policies, IDS contract negotiation, IDS connectors and Policy Execution Points (PXP). Usage Control Policies can only be applied on *exchanged* datasets. However, in Federated Learning, because of privacy requirements and sustainability reasons, datasets are not to be transmitted. Still, the IDS principles find application. First, it is necessary to identify the data offerings, meaning the data over which the owner wants to maintain sovereignty. There are two kinds of offering. For Federated Machine Learning platforms, the offerings are the algorithms. From the data owners' point of view, the offerings are the datasets they own and make available for training AI models. Optionally, these data offerings can be registered on an IDS Metadata Broker along with contracts to make them discoverable.

Now, let us assume for example that - through an orchestrator - a Data Scientist wants to train a machine learning model on a distributed set of local datasets. The orchestrator can take in input a set of *references* to algorithms and datasets and can access the local nodes through a Docker Swarm or Kubernetes cluster. Here is where IDS comes into play. At the beginning, the orchestrator *will only own references to the assets but not the assets themselves*, it will therefore activate to get access to them through IDS. To start with, it will negotiate a contract with each of the data owners and with the FML platform. In agreeing on a contract, it will also accept the related usage control policies. Once the contracts are stipulated, it will get access to the physical algorithms from the FML platform by downloading them via APIs, whereas for the datasets it will not download them but will retrieve (exchange) only a set of corresponding representing metadata. These metadata can be passwords, access tokens, labels identifying their position inside the cluster and so on. Once obtained these metadata and therefore the due authorizations, it can proceed by deploying - outside IDS and through Docker Swarm, Kubernetes or a similar technology - the algorithms where the datasets are - on the local nodes. Since, according to IDS, usage control policies can only be enforced on *exchanged* data, the idea is that by having access to the *exchanged metadata*, the orchestrator (through its IDS connector) can enforce the usage control policies on them and mirror the enforcement also to the corresponding physical assets thanks to the Policy Execution Points; modules able to execute custom code for the enforcement of usage policies. For instance, if a control policy defines that an access token (exchanged metadata) can be used only for one week, then at the end of the week that token will be revoked, and it will not be possible to access the corresponding physical dataset anymore. In this way, even without having exchanged the physical datasets, the orchestrator can still ensure data sovereignty through the related metadata. It is worth to mention that the approach according to which IDS is used only for data discovery, contract negotiation and usage control policies enforcement is



called, in IDS literature [Dataspace Connector Manual<sup>29</sup>]: *out-of-band* and it used in situations where the IDS connectors pose limitations to the way data are to be transmitted or processed.

### 3.9 Resiliency & Legal Frameworks: resiliency

The RE4DY active resilience framework aims to identify the different factors that manufacturing companies require in order to build and enhance the resilience of their operations. By resilience we mean the ability of the organization to revert back or move to new and better states of operations when they are exposed to disruptive events. Hence, the aim of such a framework is to provide manufacturing organizations with the ability to incorporate ‘resilience thinking’ such that they are able to effectively manage disruptions and develop corresponding mitigation strategies so that they can strengthen their operations as well as that of their value chains.

The RE4DY active resilience framework will demonstrate in a realistic manner, a highly standardized and unified manufacturing framework for product and value networks. Data-driven agility is required so that manufacturing industries achieve long-term resilience and meet future EFFRA Industry 4.0 resilience challenges. The framework includes data-driven capabilities and demand and supply-driven strategies in the context of the WEF manufacturing global response initiative and aims to provide guidelines to manufacturing companies on how to be resilient. Some of the different aspects considered in the building of the RE4DY active resilience framework (and as already introduced in D2.1) can be depicted by *Figure 37*.

---

<sup>29</sup> <https://international-data-spaces-association.github.io/DataspaceConnector/CommunicationGuide/v6/Streaming>



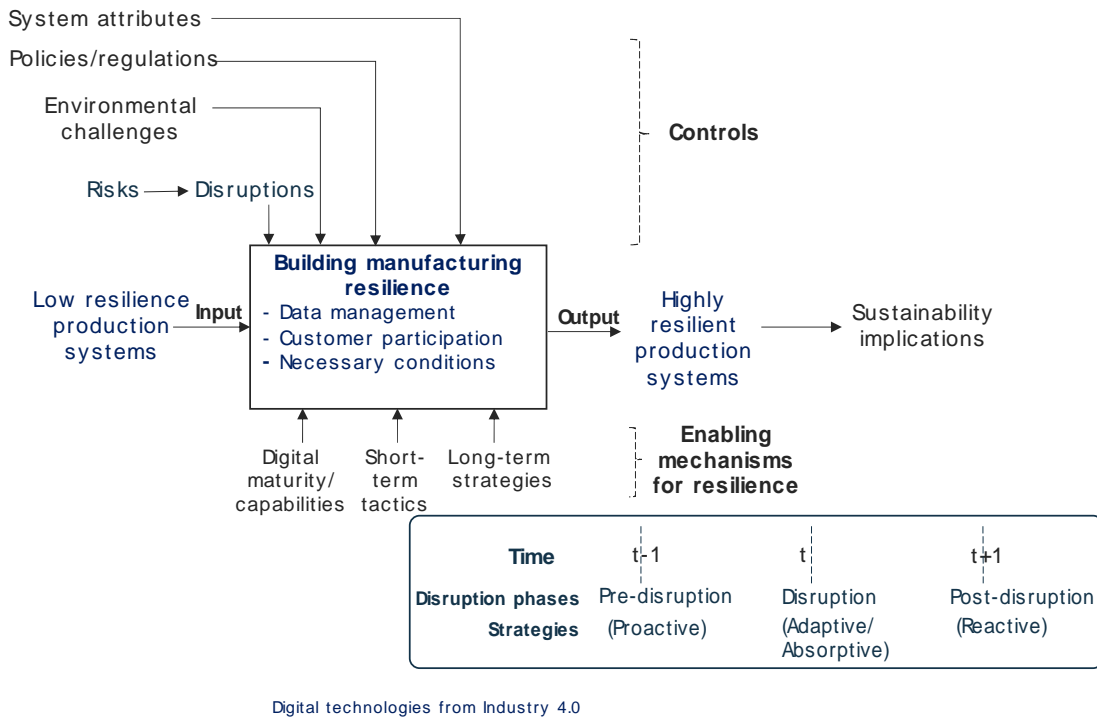


Figure 37: The RE4DY active resilience framework.

The four pilots chose one of their business processes in the RE4DY project to elucidate the building of resilience. D2.1 explains which business processes were chosen by the pilots to incorporate Resilience Engineering (RE) factors. Since then, the framework has evolved into understanding how the different elements can be applied to the pilots so that it can be generalized and validated in the different manufacturing contexts.

For instance, pilot 2 would like to perform ecosystem optimization through the application of simulation models when product changes occur. For this, they want to understand the tolerance levels to manage the quality of products when such changes or requirements from customers occur. In addition, if such changes occur, then how would production get affected? What minimum deviations in geometry would give rise to such disruptions? What other factors would need to be considered that would pose as risks?

Accordingly, the different aspects that need to be considered for building manufacturing resilience will be applied to the specific business processes chosen in the next stages. It will also be explored, how the business processes can be 'resilient by design' so as to incorporate holistic resilience engineering thinking.

Some guidelines identified are:

1. Perform a thorough risk management process. This includes several steps such as (i) risk identification (ii) source of the risk (iii) risk assessment and evaluation (severity and likelihood of risks) (iv) risk prioritization (v) identification and implementation of appropriate risk mitigation strategies (proactive, adaptive/absorptive, reactive, and transformative)
2. Identify the corresponding disruptions (external/internal, severity and likelihood)
3. Identify the boundaries of the risks and disruptions ('where' they occur, for instance the factory level, the worker, the supply chain, etc.)





4. Identify the mitigation strategies [pre-disruptive, disruptive (adaptive/absorptive), and post-disruptive (transformative, reactive)]
5. Identify which Industry 4.0 technologies can help or hinder the development of active resilience strategies and which phase they are/could be used.
6. How do the resilience strategies impact sustainability (environmental and social aspects)?

## 3.10 Resiliency & Legal Frameworks: legal

The Digital 4.0 continuum reference framework is designed to enable the realization of the digital thread and cognitive digital twin fabric scenarios outlined by pilots in D2.1. From a legal perspective, this realization is contingent upon two factors. The first factor is the ensured protection of traditional intellectual property rights (IPRs) (including copyright, patent, and trade secrets) and novel data control rights (e.g., the proposed IoT data access right under the Data Act). The second factor is compliance with applicable data governance legislation, potentially including the Data Governance Act (DGA) and the proposed Data Act (DA).

### 3.10.1 Protection of IPR and Data Control Rights

The first step in safeguarding IPRs is identifying which rights may subsist in data and, separately, in the computational models that generate that data.

Individual datums (here understood as any digital representation of acts, facts, or information) are not protected by copyright under international and European Union law (Art. 10(2) TRIPS Rec. 9 Copyright Directive; Rec. 45 Database Directive; (Giannopoulou, 2018)). Pursuant to the Database Directive, only compilations of data, whose selection or arrangement renders them intellectual creations of the author, are subject to copyright. In a similar vein, data itself is not patentable, though computer-implemented data structures and formats may qualify as patentable inventions (EPO Guidelines for Examination, Part G-II, 3.6.3). Further, there is no legally recognized “ownership” of data within Union law, in the sense that Union law has not recognized an exclusive property right that subsists in data (Geiregat, 2022).

Nevertheless, data readily falls under the scope of novel data control rights, which are rights *in personam*. More concretely, RE4DY digital twin scenarios may produce data that falls within the DA’s access right, which will entitle users of IoT products to obtain from the product’s manufacturer the data that the product generates. It is vital to query the precise extent of this proposed right, as it foresees exceptions for data derived from complex, proprietary algorithms, such as those which fuse multiple different types of sensors into a single input (DA Proposal, Art. 3(1)). A deeper understanding of the access right and its exception will be vital for furnishing Digital 4.0 partners with legal certainty regarding their rights and obligations when it comes to data control.

Moving beyond data and looking instead at the processes that transform data, it is relevant to first query whether, when, and how computer simulations (including digital twins) qualify for patentability. Since each pilot employs both machine learning and digital twins, it is likely that patents will constitute a key modality of IP protection for project partners active in the Digital 4.0 continuum reference framework.

The European Patent Office has already confirmed via case law (EPO EBA, G 0001/19) and guidelines (EPO Guidelines for Examination, Part G-II, 3.3.2) that computer-implemented



simulations and machine learning models are patentable inventions in principle. In practice, applications for patents on computer simulations must still demonstrate technical features that contribute technical effects towards the solutions of a technical problem in order to qualify for protection.

While the basic premise of patenting digital twins is not in doubt, the precise legal requirements behind patentability sometimes warrant clarification. For example, the technical character of an invention may be evinced by the presence of ‘functional data’ that in itself has a technical function, such as controlling a device’s operation (EPO Guidelines for Examination, Part G-II, 3.6.3). Patent applicants may find ‘functional data’ to be an important element in their digital twins patent applications. Yet, the concept of functional data has not been extensively developed and may be difficult to utilize in practice. This and other similar ambiguities surrounding the patentability of digital twins warrant clarification in the Digital continuum 4.0 legal framework.

Beyond clarifying the emergence and nature of IPRs and data control rights, the RE4DY legal framework will also have to accommodate the existence of a multilateral network of rightsholders. Several pilots entail ongoing data sharing between multiple entities, each of which may enrich, transform, and employ the data and the underlying simulations generating that data in different ways. Where the law is unclear on how collective contributions to a single simulation or dataset by different entities may result in either shared control of old IPRs or the creation of new IPRs, the legal framework should envision industrial agreements that facilitate easy and reliable rights management and tracking.

To properly track and manage IPR transactions, the legal framework will also need to countenance licensing issues. Existing data licenses are modeled after open software licenses, which are predicated on the fact that software is protected by copyright. Since data is not similarly protected by copyright, there exists the prospect that current data licenses are an inappropriate legal mechanism. As stated by Giannopoulou: “The application of a license inherently presumes the existence of property rights” (Giannopoulou, 2018). To avoid legal uncertainty, tailored data licensing agreements must be developed, which take into account the actual IPRs and data control rights that subsist within RE4DY pilot IP.

To reliably track IPR across complex data transactions and transformations, semantic interoperability must extend to the legal dimension. A machine-readable legal ontology of IPR would be the first step toward such interoperability and is therefore under development. Its eventual deployment will complement similar ontologies in the reference framework (e.g. for resilience).

### 3.10.2 Compliance with existing data regulations

RE4DY pilots wishing to create manufacturing data spaces will have to comply with novel legislation on data spaces. This includes both the DGA and the DA.

In such a scenario, special technical and organizational measures will be necessary for the functioning of data space operators (specifically, data intermediation service providers). The DGA obliges data intermediaries to observe strict neutrality requirements (including the partition of data intermediation activities into a separate legal entity), as well as to offer FRAND-like access to the data space’s services (DGA Ch. 3).



The DA obliges data space participants to ensure interoperability via publication of relevant dataset content, licenses, data quality, data collection methodology, data format, code lists, vocabularies, taxonomies, and APIs and the terms thereof (DA Proposal, Art. 28). It provides for the possibility that responsibilities for these aspects may be internally assigned between the various data space participants and/or may be discharged by adhering to an EC-adopted harmonized standard (DA Proposal, Art. 28). Hence, industrial agreements within the Digital continuum 4.0 legal framework should allow entities wishing to participate in a manufacturing data space to internally allocate data space responsibilities and choose between appropriate harmonized standards.

To assist the establishment of RE4DY manufacturing data spaces, the legal framework may adapt the contractual framework for data networks established in the SITRA Rulebook for a fair data economy (SITRA, 2022).



## 4 Conclusions

This report described the approach and first results achieved in defining the Digital 4.0 Continuum Reference Architecture, which aims to achieve different and challenging objectives with a holistic perspective. In fact, balancing the need for seamless data sharing with the requirement to protect sensitive information poses a significant challenge. Data sovereignty involves maintaining control over data and ensuring it is subject to the jurisdiction and regulations of the country where it resides. Adhering to EU data sovereignty principles and laws mentioned by this report (e.g., DGA and DA as mentioned in §3.10.2) while embracing data spaces and the "Data as a Product" (DaaP) business model requires careful consideration of legal frameworks, contractual agreements, and technical solutions. Collaboration between stakeholders, including businesses, policymakers, and regulatory bodies, is essential to establish a common understanding and a framework for data sovereignty in the context of data spaces and the DaaP business model.

This report outlined a first framework for all tech providers and trials to align to a common and integrated Reference Architecture. This first solution together with the highlighted challenges will provide a foundation for further evolutions in the tasks carried in WP2 and the technology advancements in WP3.

Additionally, the testing activities going to be carried out by the trial work packages, such as WP4 and WP5, will serve as a first evaluation of the architectural and implementation decisions, which will be further refined throughout the project as well as tested and exploited directly in DFA and related initiatives.

Therefore, the architecture will continue to evolve, characterize, and solidify as the pilot experiments progress, refine, and clarify the requirements. Next step within WP2 will be indeed the definition the chapter nr. 3 of the Trial Handbook with which to focus on pilots' data characterisation and system requirements.

Finally, the advancements in WP3 will play a crucial role in providing valuable feedback to this architectural framework. The approach adopted for this project embraces an iterative process that adheres to the principles of agile best practices. It implies the initial iteration of the framework, documented in this report, which serves as a foundation for the continuous enhancement and refinement of the ultimate structure expected to be achieved by M24. By following this iterative approach, the framework will evolve and integrate valuable feedback and insights acquired throughout the development lifecycle, ensuring an optimized end outcome.



## 5 References

### 5.1 EU Legislation and International Treaties

Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases (Database Directive) [1996] OJ L 77, 27.3.1996

Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (Copyright Directive) [2019] OJ L 130, 17.5.2019.

Proposal for a Regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data (Data Act) COM/2022/68 final.

Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act) [2022] OJ L 152, 3.6.2022.

TRIPS: Agreement on Trade-Related Aspects of Intellectual Property Rights, Apr. 15, 1994, Marrakesh Agreement Establishing the World Trade Organization, Annex 1C, 1869 U.N.T.S. 299, 33 I.L.M. 1197 (1994).

### 5.2 Guidelines and Case Law

EPO EBA, Case G 0001/19 (Pedestrian simulation) of 10.3.2021.

EPO Guidelines for Examination (2023). Available at: <https://new.epo.org/en/legal/guidelines-epc>

### 5.3 Bibliography

5G-ACIA. (2021, Feb). *Using Digital Twins to Integrate 5G into Production Networks*. Retrieved from <https://5g-acia.org/whitepapers/using-digital-twins-to-integrate-5g-into-production-networks/>

Bader (SAP), B. (. (2022, May). *Details of the Asset Administration Shell. Part 1 - The exchange of information between partners in the value chain of Industrie 4.0*. Retrieved from [https://www.researchgate.net/publication/361547453\\_Details\\_of\\_the\\_Asset\\_Administration\\_Shell\\_Part\\_1\\_-\\_The\\_exchange\\_of\\_information\\_between\\_partners\\_in\\_the\\_value\\_chain\\_of\\_Industrie\\_40\\_Version\\_30RC02](https://www.researchgate.net/publication/361547453_Details_of_the_Asset_Administration_Shell_Part_1_-_The_exchange_of_information_between_partners_in_the_value_chain_of_Industrie_40_Version_30RC02)

BDVA. (2017). Retrieved from [http://www.bdva.eu/sites/default/files/BDVA\\_SRIA\\_v4\\_Ed1.1.pdf](http://www.bdva.eu/sites/default/files/BDVA_SRIA_v4_Ed1.1.pdf)

C. Avasalcai, I. M. (2020). *Edge and Fog: A Survey, Use Cases, and Future Challenges in Fog Computing*. Wiley.



- DFA. (2021, 12). Retrieved from [https://digitalfactoryalliance.eu/wp-content/uploads/2021/12/Digital-Factory-eBook-reviewed\\_v1.5\\_graficas.pdf](https://digitalfactoryalliance.eu/wp-content/uploads/2021/12/Digital-Factory-eBook-reviewed_v1.5_graficas.pdf)
- DSBA. (2023, 4 21). Retrieved from <https://data-spaces-business-alliance.eu/download/33968/>
- Eclipse. (n.d.). *EDC*. Retrieved from <https://projects.eclipse.org/projects/technology.edc>
- Geiregat, S. (2022). *The Data Act: Start of a New Era for Data Ownership?* Retrieved from <http://dx.doi.org/10.2139/ssrn.4214704>
- Giannopoulou, A. (2018). Understanding Open Data Regulation: An Analysis of the Licensing Landscape. In *Open Data Exposed* (pp. 101-125). The Hague: Asser Press.
- IDSA. (2022, 09 16). *International Data Spaces Information Model*. Retrieved from <https://international-data-spaces-association.github.io/InformationModel/docs/index.html>
- IDSA. (n.d.). *App Store*. Retrieved from [https://docs.internationaldataspaces.org/ids-ram-4/layers-of-the-reference-architecture-model/3-layers-of-the-reference-architecture-model/3\\_5\\_0\\_system\\_layer/3\\_5\\_3\\_app\\_store\\_and\\_data\\_apps#app-store-and-ids-apps](https://docs.internationaldataspaces.org/ids-ram-4/layers-of-the-reference-architecture-model/3-layers-of-the-reference-architecture-model/3_5_0_system_layer/3_5_3_app_store_and_data_apps#app-store-and-ids-apps)
- IDSA. (n.d.). *Clearing House*. Retrieved from [https://docs.internationaldataspaces.org/ids-ram-4/layers-of-the-reference-architecture-model/3-layers-of-the-reference-architecture-model/3\\_5\\_0\\_system\\_layer/3\\_5\\_5\\_clearing\\_house#clearing-house](https://docs.internationaldataspaces.org/ids-ram-4/layers-of-the-reference-architecture-model/3-layers-of-the-reference-architecture-model/3_5_0_system_layer/3_5_5_clearing_house#clearing-house)
- IDSA. (n.d.). *Dataspace Connector*. Retrieved from <https://international-data-spaces-association.github.io/DataspaceConnector/>
- IDSA. (n.d.). *Identity Provider*. Retrieved from [https://docs.internationaldataspaces.org/ids-ram-4/layers-of-the-reference-architecture-model/3-layers-of-the-reference-architecture-model/3\\_5\\_0\\_system\\_layer/3\\_5\\_1\\_identity\\_provider#identity-provider](https://docs.internationaldataspaces.org/ids-ram-4/layers-of-the-reference-architecture-model/3-layers-of-the-reference-architecture-model/3_5_0_system_layer/3_5_1_identity_provider#identity-provider)
- IDSA. (n.d.). *IDS connector*. Retrieved from [https://docs.internationaldataspaces.org/ids-ram-4/layers-of-the-reference-architecture-model/3-layers-of-the-reference-architecture-model/3\\_5\\_0\\_system\\_layer/3\\_5\\_2\\_ids\\_connector#ids-connector](https://docs.internationaldataspaces.org/ids-ram-4/layers-of-the-reference-architecture-model/3-layers-of-the-reference-architecture-model/3_5_0_system_layer/3_5_2_ids_connector#ids-connector)
- IDSA. (n.d.). *IDS-RAM 4.0*. Retrieved from <https://docs.internationaldataspaces.org/ids-ram-4/>
- IDSA. (n.d.). *Metadata Broker*. Retrieved from [https://docs.internationaldataspaces.org/ids-ram-4/layers-of-the-reference-architecture-model/3-layers-of-the-reference-architecture-model/3\\_5\\_0\\_system\\_layer/3\\_5\\_4\\_metadata\\_broker#metadata-broker](https://docs.internationaldataspaces.org/ids-ram-4/layers-of-the-reference-architecture-model/3-layers-of-the-reference-architecture-model/3_5_0_system_layer/3_5_4_metadata_broker#metadata-broker)
- IDSA. (n.d.). *Vocabulary hub*. Retrieved from [https://docs.internationaldataspaces.org/ids-ram-4/layers-of-the-reference-architecture-model/3-layers-of-the-reference-architecture-model/3\\_5\\_0\\_system\\_layer/3\\_5\\_6\\_vocabulary\\_hub#vocabulary-hub](https://docs.internationaldataspaces.org/ids-ram-4/layers-of-the-reference-architecture-model/3-layers-of-the-reference-architecture-model/3_5_0_system_layer/3_5_6_vocabulary_hub#vocabulary-hub)
- IIC. (2022, 11 7). Retrieved from <https://www.iiconsortium.org/wp-content/uploads/sites/2/2022/11/IIRA-v1.10.pdf>



- NIST. (2019, October). *NIST Big Data Interoperability Framework: Volume 6, Reference Architecture*. Retrieved from <https://doi.org/10.6028/NIST.SP.1500-6r2>
- SITRA. (2022). *Rulebook for a fair data economy*. Retrieved from <https://www.sitra.fi/en/publications/rulebook-for-a-fair-data-economy/>
- ZVEI. (2019). *RAMI 4.0*. Retrieved from [https://www.zvei.org/fileadmin/user\\_upload/Presse\\_und\\_Medien/Publikationen/2019/Maerz/Vision\\_Antrieb\\_4.0\\_ENGLISCH/ZVEI\\_BR\\_Vision\\_Antrieb\\_4.0\\_ENGLISCH25.03.19\\_\\_003\\_.pdf](https://www.zvei.org/fileadmin/user_upload/Presse_und_Medien/Publikationen/2019/Maerz/Vision_Antrieb_4.0_ENGLISCH/ZVEI_BR_Vision_Antrieb_4.0_ENGLISCH25.03.19__003_.pdf)

