CEF2 RailDataFactory

Deliverable 5.2 – Final Study Results

Due date of deliverable: 31/12/2023

Actual submission date: 13/01/2024

Leader/Responsible of this Deliverable: Patrick Marsch (DB)

Reviewed: **Y**/N

Document status		
Revision	Date	Description
01	23/11/2023	Document structure generated
02	02/01/2024	Chapter 2 substantially updated
03	04/01/2024	Chapter 3 substantially updated
04	10/01/2024	Chapters 1-3 consolidated and chapters 4-7 updated
05	11/01/2024	Clean version established
06	13/01/2024	Version submitted to the project officer

Project funded by the European Health and Digital Executive Agency, HADEA, under Connecting Europe Facilities Digital Grant Agreement 101095272		
Dissemination Level		
PU	Public	X
SEN	Sensitiv – limited under the conditions of the Grant Agreement	

Start date: 01/01/2023 Duration: 12 months

ACKNOWLEDGEMENTS

This project has received funding from the European Health and Digital Executive Agency, HADEA, under Connecting Europe Facilities Digital Grant Agreement 101095272.

REPORT CONTRIBUTORS

Name	Company
Patrick Marsch	DB
Wolfgang Albert	DB
Philipp Neumaier	DB
Alexander Heine	DB
Julian Wissmann	DB
Waseem Ul Aslam Peer	DB
Bart du Chatinier	NS
Philippe David	SNCF

Note of Thanks

We would like to speak out a big thank you toward our Advisory Board members for their detailed input to the project, and the valuable discussion we had throughout the project.

Disclaimer

Licensing

EXECUTIVE SUMMARY

The European rail sector is currently on the verge to the strongest technology leap in its history, with many railway infrastructure managers and railway undertakings striving toward large degrees of automation in rail operation, and mechanisms to increase the capacity and quality of rail operation.

In particular in the pursuit of fully automated driving (so-called Grade of Automation 4, GoA4), where sensors and cameras on trains will be used to automatically detect hazards in rail operation, it is commonly understood that an individual railway company or railway vendor would not be able to collect enough sensor data to sufficiently train the artificial intelligence (AI) eventually deployed in the rail system.

For this reason, it is commonly assumed that a form of Pan-European Railway Data Factory is needed, as a part of the overall ecosystem that allows various railway players and suppliers to collect and process sensor data, perform simulations, develop AI models, certify models, and ultimately deploy the models in the automated railway system.

In close sync with related activities listed in Section 1.2, the **CEF2 RailDataFactory** study has focused in particular on the High-Speed Pan-European Railway Data Factory Backbone Network and data platforms required to realise the vision of the Pan-European Railway Data Factory.

In this deliverable of the study, all key findings of the project are summarised, open points for further investigation are listed, and the project is concluded.

ABBREVIATIONS AND ACRONYMS

Abbreviation	Definition
AI	Artificial Intelligence
ATO	Automatic Train Operation
DSD	Digitale Schiene Deutschland
DZSF	Deutsches Zentrum für Schienenverkehrsforschung (German Centre for Rail Traffic Research)
DWDM	Dense Wavelength Division Multiplex
ETCS	European Train Control System
FRMCS	Future Railway Mobile Communication System
GDPR	General Data Protection Regulation
GoA4	Grade of Automation 4
HPC	High Performance Computing
IAM	Identity and Access Management
IM	Infrastructure Manager
ISP	Internet Service Provider
ML	Machine Learning
PEDF	Pan-European (Railway) Data Factory
PUE	Power Usage Effectiveness
RU	Railway Undertaking
SDN	Software Defined Networking
TSI	Technical Specification for Interoperability
ZTA	Zero Trust Architecture

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

The European railway sector is on the verge to the strongest technology leap in its history, with many railway infrastructure managers (IMs) and railway undertakings (RUs) striving toward large degrees of automation in rail operation, and mechanisms to increase the capacity and quality of rail operation.

In particular, various railway companies – both IMs and RUs – and railway suppliers are currently working toward fully automated rail operation (so-called Grade of Automation 4, GoA4), for instance in the context of the Shift2Rail [1] and Europe's Rail [2] programs, in which sophisticated lidar and radar sensors as well as cameras are used to automatically detect and respond to hazards in rail operation, such as objects on the track or passengers in stations in dangerous proximity of the track. Another important use case is high-precision train localisation by detecting static infrastructure elements and locating them on a digital map, as for instance covered in the Sensors4Rail project [3]. While the rail system has various properties that render fully automated driving principally easier than, e.g., in the automotive sector (for instance, railway motion is only one-dimensional, scenarios are typically much less complex than automotive scenarios, etc.), key challenges on the way to fully automated driving in the rail sector are that hazardous situations have to be detected much earlier due to long braking distances, and it is very challenging to collect and annotate sufficient amounts of sensor data with sufficient occurrences of relevant incidences to perform the required artificial intelligence (AI) training and to be able to prove that the trained AI meets the safety needs.

For this, it is expected that single railway suppliers, IMs and RUs will not be able by themselves to collect and annotate sufficient amounts of sensor data for AI training purposes – but instead, an European data platform and ecosystem is required into which railway stakeholders (suppliers, IMs, RUs, safety authorities, and others) can feed, process and extract sensor data, as well as simulate artificial sensor data, and through which the stakeholders can jointly develop and assess the AI models needed for fully automated driving.

Cross-border data exchange is crucial for railway undertakings, even if nationally different requirements exist. Through an improved use of technology, for example transfer learning or self-supervision learning with existing data, these national requirements can be partially resolved, and a significant acceleration can be achieved. As an example, transfer learning is a machine learning (ML) technique in which knowledge learned from one task is reused to improve performance on a related task. Among other things, cross-border data exchange enables seamless coordination of the development of fully automated driving and interoperability between different national railway networks and ensures efficient and smooth cross-border operations. The EU Directive (EU) 2016/797 [4] on the interoperability of the rail system provides guidelines and rules to promote such data exchange and ensures a standardised and effective approach across Europe.

1.1 AIM AND SCOPE OF THE CEF2 RAILDATAFACTORY STUDY

The CEF2 RailDataFactory study has focused exactly on the stated vision of a Pan-European (Railway) Data Factory (PEDF) for the joint development of fully automated driving. The study, being co-funded through HADEA, has aimed to assess the feasibility of a PEDF from technical, economical, legal, regulatory and operational perspectives, and determine key aspects that are required to make a PEDF a success. For a better understanding of the study's aim and scope, please see Chapter 1.1 in Deliverable D 1 [5].

1.2 DELINEATION FROM AND RELATION TO OTHER WORKS

The Shift2Rail project **TAURO** [6] also looks into the development of fully automated rail operation, for instance focusing on developing

- a common database for AI training;
- a certification concept for the artificial sense when applied to safety related functions;
- track digital maps with the integration of visual landmarks and radar signatures to support enhanced positioning and autonomous operation;
- environment perception technologies (e.g., artificial vision).

The difference of the CEF2 RailDataFactory project is that this has put special emphasis on the **High-Speed Pan-European Railway Data Factory Backbone Network** and **data platform** (located on the infrastructure side, but used for sensor data collected through both onboard and infrastructure side sensors) required for the PEDF, and also investigated **commercial, legal and operational aspects** that have to be addressed to ensure that the vision of the PEDF can be realised.

DB Netz AG and the German Centre for Rail Traffic Research (DZSF) have released OSDaR23, the first publicly available multi-sensor data set for the rail sector [7][8]. The data set is aimed at training AI models for fully automated driving and route monitoring in the railway industry. It includes sensor data from various cameras, infrared cameras, LiDARs, radars, and other sensors, recorded in different environments and operating situations, and annotated with labels for different objects and situations. The data set is utilised in the Data Factory of Digitale Schiene Deutschland (DSD) to train AI software for environment perception, and more annotated multi-sensor data sets will be created in the future.

The Europe's Rail Innovation Pillar **FP2 R2DATO project** [9], overall focusing on the further development of automated rail operations, also has a work package dedicated to the PEDF. Here, however, the main focus is on creating first implementations of individual data centres and toolchains as required for specific other activities and demonstrators in the FP2 R2DATO project, and on developing an **Open Data Set**. A strong alignment between the CEF2 RailDataFactory study and the FP2 R2DATO Data Factory activities is ensured through an alignment on use cases and operational scenarios, though the actual focus of the projects is then different.

EU-wide research programs are being carried out on Flagship Project 2: "Digital & Automated up to Autonomous Train Operations" and in this context the European perspective is discussed. In addition, each country and each railway infrastructure provider have their own programs, where there is usually also an exchange within the Innovation and System Pillar in the R2DATO. The participants in this study also work in these bodies and try to reflect the European picture. Within the sector initiative Digitale Schiene Deutschland (DSD), Deutsche Bahn already started to set up some components of the data centre in Germany [10].

1.3 AIM AND STRUCTURE OF THIS DELIVERABLE

This current document is the final deliverable D5.2 of the CEF 2 RailDataFactory study, which summarises the key findings of the project.

It is structured as follows, to a large extent taking orientation in the work package structure of the project, as shown in Figure 1:

- In **Chapter 2**, the concept and rationale of the PEDF are summarised, with an emphasis on the key design paradigms, expected users and roles and their benefit from the PEDF, and key hypotheses on the PEDF agreed in the project. This more or less summarises the work conducted in WP 1 on the "Data Factory Concept";
- In **Chapter 3**, the possible architecture and key technical elements of the PEDF are summarised, with a reflection on the typical data processing steps in the context of the development of GoA4 and related building blocks that have been identified. The chapter further summarises the considerations and developments in the project on the requirement pan-European backbone network, the security framework incl. identity access management, and data architecture, management and governance. This corresponds to a summary of WP 2 "Data Factory Architecture";
- In **Chapter 4**, investigations of the PEDF from operational, commercial, security and legal perspectives are shortly summarised, as conducted in the project in WP 3 "Commercial and Operational Assessment";
- In **Chapter 5**, a possible deployment strategy and specific rollout steps for the PEDF are laid out, capturing the work in WP 4 "Deployment Strategies and Enablement";
- In **Chapter 6**, open points are listed that were identified during the project and that could not be concluded, and
- In **Chapter 7**, the deliverable and project are concluded.

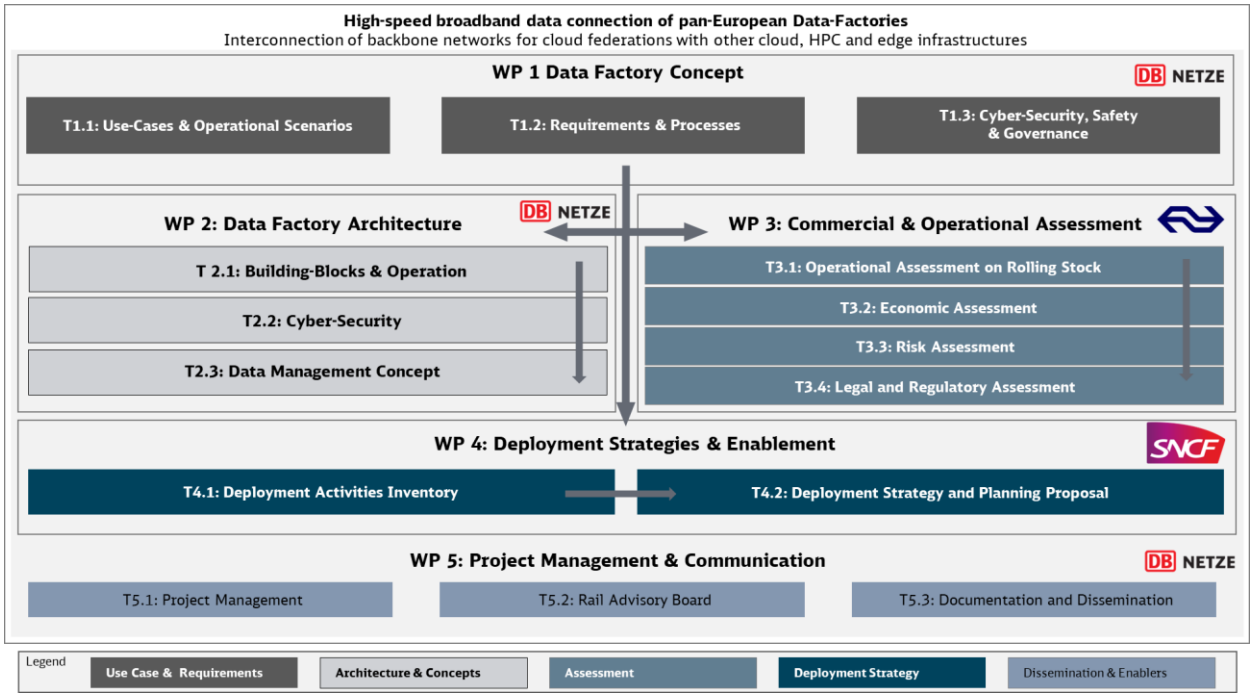

Figure 1. Work package and task structure of the project.

2 CONCEPT AND RATIONALE OF A PAN-EUROPEAN DATA FACTORY

2.1 HIGH-LEVEL CONCEPT AND KEY PARADIGMS

The RailDataFactory study under the CEF2 program has investigated the possible setup of a **Pan-European (Railway) Data Factory (PEDF)** for the railway sector. Such PEDF, deployed on a large scale throughout Europe, will constitute a paradigm shift in how railway data is collected, processed, and utilised across Europe, facilitating a more integrated, efficient, and technologically advanced rail network. In particular, a PEDF is seen as an important enabler for the development of fully automated driving (so-called Grade of Automation 4, GoA4) in the rail sector, as this requires the collection and processing of large amounts for data for artificial intelligence (AI) training, which a single railway or supplier could likely not achieve by itself.

As shown in Figure 2 and detailed in Deliverable D 1 [5], the PEDF constitutes an interconnected set of **individual Data Factories** across Europe, operated by various stakeholders such as railway infrastructure managers (IMs), railway undertakings (RUs), suppliers, etc., and interlinked through a **High-Speed Pan-European Backbone Network**. The **Data Centres** within the individual Data Factories are equipped with computing and storage resources and host an array of tools and services. These enable the collection and processing of railway data related to train operations, infrastructure management, passenger services and maintenance, including data from cameras, lidar or radar sensors on trains or at the trackside that is used for training AI models toward fully automated rail operation. So-called **Touch Points** constitute the entry points of the data to the PEDF, for instance providing high-speed wireless connectivity between trains and ground for the transfer of large amounts of collected sensor data, and allowing for a pre-selection and pre-processing of sensor data before this is transferred to Data Centres.

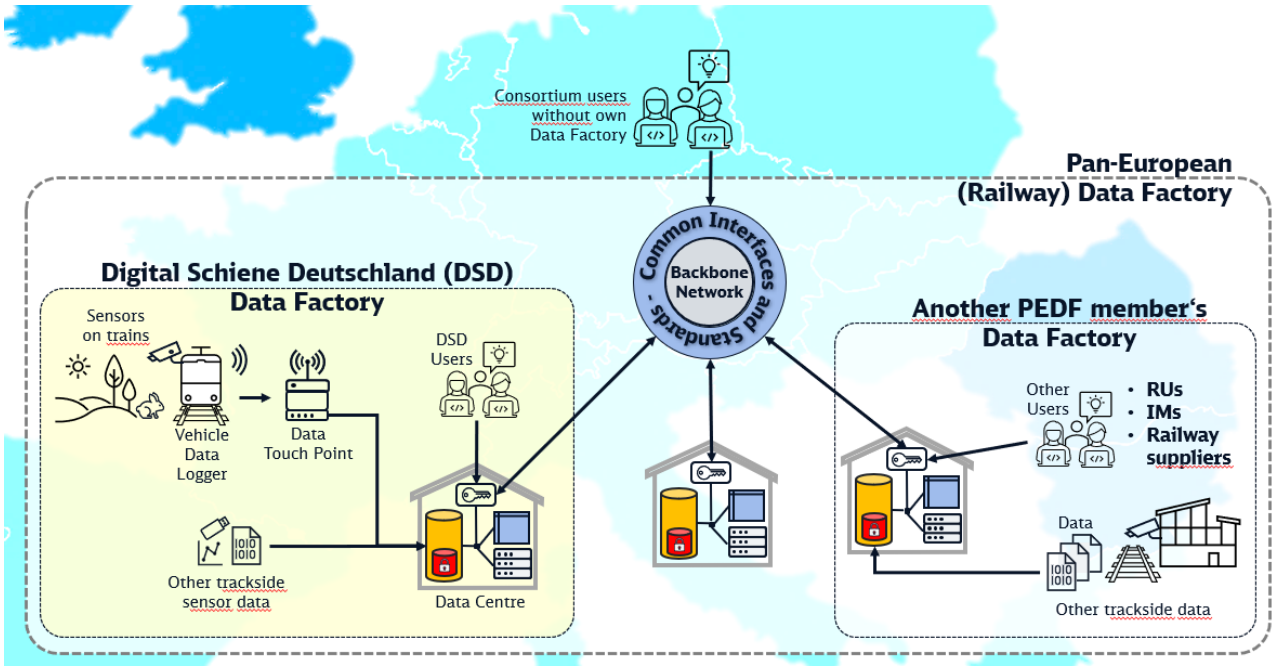

Figure 2. High-level depiction of the envisioned Pan-European Railway Data Factory.

It is important to note that not every stakeholder or European country needs its own Data Centre(s). Through the interconnection of facilities offered by various stakeholders via the High-Speed Pan-

European Backbone Network, data, including high-volume data, can be transported safely and efficiently through the entire network, and a broad set of services is made available to a large ecosystem of partners.

The mentioned Touch Points and connected Data Centres also enable **Edge Computing**, i.e., the notion of processing data close to its source (at the 'edge' of the network) rather than in a centralised data processing facility. This approach reduces latency and bandwidth requirements and enhances real-time data processing capabilities. This is seen as particularly relevant for real-time data analysis in train operations, predictive maintenance, and emergency response scenarios.

The vision of the PEDF builds upon the following key paradigms:

- **Data Integration and Interoperability**: The PEDF is designed to maximally support the joint creation, processing and usage of rail data, basically creating a comprehensive data repository accessible to all stakeholders. More precisely, the PEDF will integrate both real-world sensor data, for instance from lidar and radar sensors or cameras, and artificial sensor data generated from simulations. This combination is vital for training and evaluating AI functions for automated rail operation. This overall ambition obviously requires the standardisation of data formats, protocols and interfaces;
- **Support of AI and Machine Learning**: AI and machine learning algorithms will play an increasingly crucial role in analysing vast amounts of railway data, enabling predictive maintenance, traffic management, and automated train control systems, and will likely form the basis for fully automated rail operation. The PEDF aims to harmonise the development, training, and deployment of AI models, ensuring reliability and safety in their application. More precisely, the PEDF is designed to support the entire lifecycle of AI models – encompassing development, training, certification, and deployment for automated rail operations. It particularly focuses on the concept of Transfer Learning, where AI models trained in one Data Centre can be further evolved in others, accommodating cross-border operational needs;
- **Cybersecurity and Data Protection by Design**: With the increasing reliance of the rail sector on data and digital technologies, ensuring the security and integrity of data and its processing infrastructure is of course paramount. The PEDF hence follows a robust cybersecurity strategy from the beginning, compliant with EU regulations like the General Data Protection Regulation (GDPR), to protect sensitive data and prevent unauthorised access;
- **Standards and Regulations**: The PEDF shall in its core design not only ensure compliance to existing European standards and regulations, such as Technical Standards for Interoperability (TSIs) and GDPR, for railway operations, data protection, and cybersecurity, but also contribute to the development of new standards that cater to the evolving needs of the digital railway environment;
- **Data Sovereignty and Decentralisation**: Despite the bold vision on a joint European infrastructure, it is imperative that the PEDF ensures autonomy for individual stakeholders in managing their data. These have to be able to set up private Data Centres and customise toolchains according to their specific needs. At the same time, the PEDF functions as a federated ecosystem that promotes resource and data sharing while maintaining a low barrier to entry for new participants.

Ensuring data quality and establishing a common understanding and common technologies for data processing are critical to the success of the PEDF. Additionally, harmonising infrastructure and tools is vital for enabling cross-border train operation and ensuring operational efficiency.

The investigated PEDF represents a significant leap forward in digitalising and harmonising the European railway sector. It sets the foundation for a future where data-driven insights lead to safer, more efficient, and sustainable rail operations, aligning with the broader goals of the European Union's transport and digitalisation policies.

2.2 USERS AND THEIR RATIONALE FOR JOINING A PAN-EUROPEAN DATA FACTORY

The concept of a Pan-European (Railway) Data Factory (PEDF), as envisaged in the CEF2 RailDataFactory project, is designed to attract a wide range of users from different sectors of the European railway industry. The rationale for these diverse stakeholders to join the Data Factory is driven by the numerous benefits and advancements it offers.

2.2.1 Potential Users of the Pan-European Data Factory

The envisioned primary users of the PEDF include IMs, RUs, railway suppliers, authorities, as well as research and development entities. These stakeholders are pivotal in the advancements toward fully automated rail operation (GoA4), being central to the vision of the PEDF.

- **Railway Operators (RUs):** These include companies responsible for train operations across Europe. They are primary users of the PEDF, leveraging its capabilities for efficient train scheduling, route optimisation, predictive maintenance, and enhanced passenger services;
- **Infrastructure Managers (IMs):** Being responsible for the maintenance and management of railway infrastructure, IMs could utilise the PEDF for better monitoring and maintenance of tracks, stations, and other infrastructure components. The data-driven insights help in predictive maintenance and infrastructure optimisation. In countries where there is one major rail IM, this may also be a somewhat natural entity to provide a kind of national nucleus for the PEDF, given their role and existing infrastructure. Further, the IM can this way ensure that fully automated driving is introduced in a harmonised way on its rail network;
- **Railway Suppliers:** These are companies that provide technology solutions to the railway sector, ranging from signalling systems to rolling stock manufacturers. They could use the PEDF to develop and test new technologies, to gather insights for product improvements, and to prepare and ultimately obtain authorisation for their products;
- **Regulatory Bodies and Safety Authorities:** These entities use the data for oversight, ensuring compliance with safety standards and regulations. The PEDF provides them with accurate, comprehensive data for better policy-making and regulatory enforcement. Through the notion of the PEDF, authorisation of AI-based solutions in the rail sector may also be better coordinated and harmonised throughout Europe;
- **Research and Development Entities:** Universities, research institutes, and innovation hubs could use the PEDF for academic and applied research. It could serve as a resource for studying railway systems, developing new technologies, and advancing knowledge in the field.

To support the PEDF, the following further entities may be contracted or otherwise provide their services:

- **IT / Tech Providers**: Entities who provide and/or operate parts of the Data Centre infrastructure incl. service toolchains. For these, the PEDF would provide a very efficient platform to offer their services to a large European railway market;
- **Simulation and data processing providers:** These are entities specialised for instance on providing simulations (or simulation infrastructure and toolchains) or processing data (for instance providing annotation services).

2.2.2 Rationale for Joining the Pan-European Data Factory

The rationale for various stakeholders in the European railway sector to join the PEDF is multifaceted, reflecting the diverse benefits and opportunities that this initiative offers. Here are the key reasons potentially driving their participation:

- **Enhanced Data Access and Sharing:** The PEDF provides a centralised platform for managing and accessing a vast repository of railway data. This accessibility is crucial for stakeholders who require comprehensive and integrated data for various operational, safety, and strategic purposes;
- **Innovation and Technological Advancement**: The PEDF fosters an environment conducive to innovation. Access to extensive data resources enables stakeholders to develop, test, and implement advanced technologies and solutions, such as AI-driven predictive maintenance systems or advanced traffic management solutions;
- **Operational Efficiency and Optimisation:** By leveraging the data available through the PEDF, users can significantly improve operational efficiency. This includes optimising train schedules, improving maintenance practices, enhancing infrastructure management, and ensuring better resource allocation. Users can optimise various aspects of railway operations, leading to collaboration and standardisation: Joining the PEDF encourages collaboration among railway operators, infrastructure managers, technology providers, and regulators. This collaborative environment helps in standardising data formats, protocols, and practices, ensuring consistency and interoperability across the European rail network;
- **Safety and Security Enhancement:** The PEDF can provide valuable insights for improving safety and security in railway operations. Analysing data from various sources helps in identifying potential risks, enabling proactive measures for safety and security management;
- **Regulatory Compliance and Quality Control:** With the railway industry being heavily regulated, the PEDF aids stakeholders in complying with various standards and regulations. The access to detailed and accurate data – and a harmonised processing of data - simplifies the process of monitoring, reporting, and ensuring compliance with regulatory requirements;
- **Cost Reduction and Economic Benefits:** By optimising operations and maintenance through data-driven insights, stakeholders can achieve significant cost savings. Furthermore, the PEDF can open up new economic opportunities by enabling the development of new services and solutions;

- **Market Responsiveness and Customer Service:** For service providers, the PEDF offers insights that can be used to better understand and respond to market demands and customer needs, leading to improved service offerings and enhanced customer satisfaction;
- **Research and Academic Insights:** For academic and research institutions, the PEDF is a valuable resource for conducting research on various aspects of railway operations, contributing to the advancement of knowledge and technology in the sector.

In summary, the diverse user base of the PEDF is motivated by the shared goal of harnessing data for operational excellence, safety, innovation, and compliance in the European railway sector. The PEDF serves as a pivotal platform for realising these objectives, fostering a more integrated, efficient, and forward-looking European rail network.

2.3 POTENTIAL ROLES WITHIN THE PAN-EUROPEAN DATA FACTORY

Beyond being a technical infrastructure platform, the PEDF is expected to foster the construction of a broad ecosystem of stakeholders who contribute to, and benefit from, the PEDF in different ways. As detailed in Deliverable D 1 [5] and expanded in Deliverable D 4.1 [11], in particular the following roles as listed in Table 1 could be taken up by the users and stakeholders of the PEDF.

Table 1. Potential roles of users of the Pan-European Railway Data Factory.

Role		Description	Users and other stakeholders who could or should take stated role					
			Railways (IMs or RUs)	Railway suppliers	Regulatory bodies / safety authorities	R&D entities	IT/Tech providers	Simulation and data processing providers
Contributor	**Data provider**	Provides its own high-quality data to the PEDF. This role is expected to have data sovereignty over this data and can release it to other participants for further processing. The role may apply to raw or processed and annotated data, or (trained) AI models.	X	X				
	Service provider	Defines and provides services which consumers of the PEDF can use to access and process data. This for instance includes the provisioning of IT infrastructure including tools and algorithms, or data related services like data annotation. It	X	X		X	X	X

		is possible that a Service Provider connects existing services to a more complex service.						
	Node provider	Supports the PEDF with infrastructure and compute power. A Node Provider also provides information where to run services best.	X	X			X	X
	Instance provider	Defines where and how a service runs, takes care of pipelines and orchestration of processes.	X	X			X	X
	Financial contributor	Provides a financial contribution to the PEDF to be able to use services and tools which are provided. This role could in particular be taken by entities who themselves do not contribute data or services.	X	X	X	X		
	Authorisation provider	Approves that certain services and toolchains provided in the PEDF are suitable to be used for the development of railway products (e.g., trained AI models) that can later obtain authorisation. An authorisation provider may also provide guidelines or process definitions according to which PEDF services are defined.	(X)		X			
Consumer	**Data consumer**	Consumes data provided in the PEDF, for instance processed and annotated sensor data or trained AI models.	X	X	X	X		
	Service consumer	Consumes services provided in the PEDF, such as toolchains, simulation services, AI training services, annotation or other data processing services, etc.	X	X	X	X		
	Infrastructure consumer	Utilises the IT infrastructure provided by Node or Instance Providers, for instance to run its own services.	X	X	X	X	X	X

2.4 EXPECTED OVERALL BENEFITS OF THE PAN-EUROPEAN DATA FACTORY

Beyond the benefits that individual users and stakeholders related to the PEDF may have, it is expected that the PEDF provides overall substantial benefits for the rail sector, which are listed in the following:

1. **Improved Operational Efficiency:** The creation of a unified data platform is expected to significantly enhance the operational efficiency of railway systems across Europe. This efficiency is expected to stem from better data integration, analysis, and application in various aspects of railway operations;
2. **Facilitate enhanced Safety and Reliability:** The PEDF is anticipated to contribute to improved safety and reliability in railway operations. By facilitating comprehensive data analysis, predictive maintenance, and real-time monitoring, the system is expected to help in proactively identifying and mitigating risks;
3. **Facilitation of Cross-Border Interoperability:** A central hypothesis is that the PEDF will greatly facilitate cross-border interoperability within the European rail network. This is based on the idea that standardised data formats and protocols will enable seamless data exchange and coordination across different national railway systems;
4. **Advancement in Railway Technology and Innovation:** The PEDF is expected to act as a catalyst for technological advancement and innovation in the railway sector. The availability of a large pool of data is assumed to drive the development and implementation of advanced technologies, including AI and machine learning, leading to more sophisticated and efficient railway systems;
5. **Data-Driven Decision Making and Policy Development:** Another key hypothesis is that the insights gained from the PEDF will enable more informed decision-making and policy development. This encompasses operational decisions, strategic planning, and regulatory policies, all aimed at improving the overall functionality and service quality of the railway network;
6. **Economic Benefits and Cost Savings:** The initiative is presumed to bring about significant economic benefits and cost savings. By optimising operations, reducing downtime, and improving resource allocation, the PEDF is expected to contribute to the overall financial health of the European rail sector;
7. **Strengthening Data Security and Compliance:** The PEDF is hypothesised to strengthen data security and ensure compliance with regulatory standards such as GDPR. The centralised nature of the data management system is expected to facilitate better control and protection of sensitive information;
8. **Enhancing Customer Experience:** Finally, it is hypothesised that the PEDF will enhance the customer experience in rail travel. This could be through improved service reliability, personalised services, real-time updates, and overall enhanced service quality.

These expectations on the overall benefits of the PEDF collectively form the foundational belief system driving the development and expected outcomes of the PEDF, as also highlighted in Deliverable D 4.2 [12]. The success of the PEDF hinges on the validation and realisation of these claims, which aim to bring a cohesive, efficient, and technologically advanced transformation to the European railway industry.

2.5 KEY HYPOTHESES ON THE PAN-EUROPEAN DATA FACTORY

During the CEF2 RailDataFactory study, it became apparent that there are very diverse views on a potential PEDF, with some controversies on key properties of the PEDF. It was hence seen as beneficial to jointly clarify and nail down key hypotheses on the PEDF that are fundamental to the subsequent detailed design of the PEDF. These hypotheses are shortly listed in the following, with more details provided in Deliverable D 4.2 [12].

Hypothesis 1: There is no "national" Data Factory, but many individual Data Factories

The Data Factories constituting the PEDF are likely not managed nationally, but can belong to different stakeholders, as listed in Section 2.2.1. Hence, it is seen as appropriate to rather refer to individual Data Factories than the notion of "national" Data Factories.

Hypothesis 2: Each individual Data Factory has its own business model

Each individual Data Factory can and should have its own business model and may encompass individual tools to meet the individual requirements of the owners and operators. However, it should be pointed out that the synergy effects within the framework of PEDF can leverage further business potential and, in the long term, increase cost efficiency in meeting individual requirements.

Hypothesis 3: The PEDF is a federation of individual Data Factories

The PEDF is an association of several entities owning and operating individual Data Factories, possibly involving a central governing authority, as elaborated in Deliverable D 4.1 [11].

Hypothesis 4: There is no need for real-time transfer of large data between train and ground

This means that in the case of sensor data or other data with a large data size, no real-time data transmission is required, but instead the data is uploaded or downloaded during maintenance, for instance based on the notion of Touch Points detailed in Deliverable D 1 [5]. Note: The exchange of smaller amounts of data, for instance diagnostics or asset management related information, could and should of course happen in real-time, for which the new 5G-based Future Railway Mobile Communication System (FRMCS) [13] could be used.

Hypothesis 5: The use of the PEDF is not geographically limited to the location of its members

It is assumed that once a software component has been validated by the relevant European safety authorities for use on board a train on European territory, this component may be used on European trains. This applies regardless of whether or not there is an individual Data Factory in the European country in question.

Hypothesis 6: No technical specification will come from the PEDF for the perception systems to the train manufacturers

The PEDF defines requirements with regard to data quality and data formats and can make some recommendations on the technical specifications of the various perception systems, but will not prescribe the latter specifications.

3 POSSIBLE ARCHITECTURE AND KEY TECHNICAL ELEMENTS OF A PAN-EUROPEAN DATA FACTORY

The Pan-European (Railway) Data Factory (PEDF), encompassing a wide spectrum of technological, operational, and strategic components, aims to revolutionise the way railway data is processed, shared, and utilised across Europe. The proposed architecture and its key technical elements hence have to be designed not only to address the current challenges faced by the railway industry, but also to pave the way for future advancements.

At its core, the architecture of the PEDF is a blueprint for an integrated network that binds together disparate data sources from across the continent, creating a unified platform for data exchange and analysis. This structure is crucial for achieving the high levels of interoperability and standardisation necessary for a seamless, cross-border railway operation.

The key technical elements of this architecture are diverse, each playing a specific role in the larger mechanism. These elements include advanced data processing centres, high-speed communication networks, robust Cybersecurity measures, and sophisticated AI and machine learning algorithms. Together, they form the backbone of a system capable of handling the immense volumes of data generated by the European rail network, transforming it into actionable insights and solutions.

In the next sections, we delve into the details of the proposed architecture and its key components, exploring how each element contributes to the overall functionality and efficacy of the PEDF.

3.1 HIGH-LEVEL ARCHITECTURE

3.1.1 Typical Data Life Cycle

For the design of the architecture of the PEDF, it is important to consider the typical life cycle of the data it is expected to generate, store and process. As the training of AI models for fully automated rail operation is one main driving use case of the PEDF, see also Deliverable D 1 [5], let us here hence shortly look into the related data life cycle, as depicted in Figure 3, which takes orientation in the “framework for artificial intelligence system using machine learning” according to ISO 23053 [14].

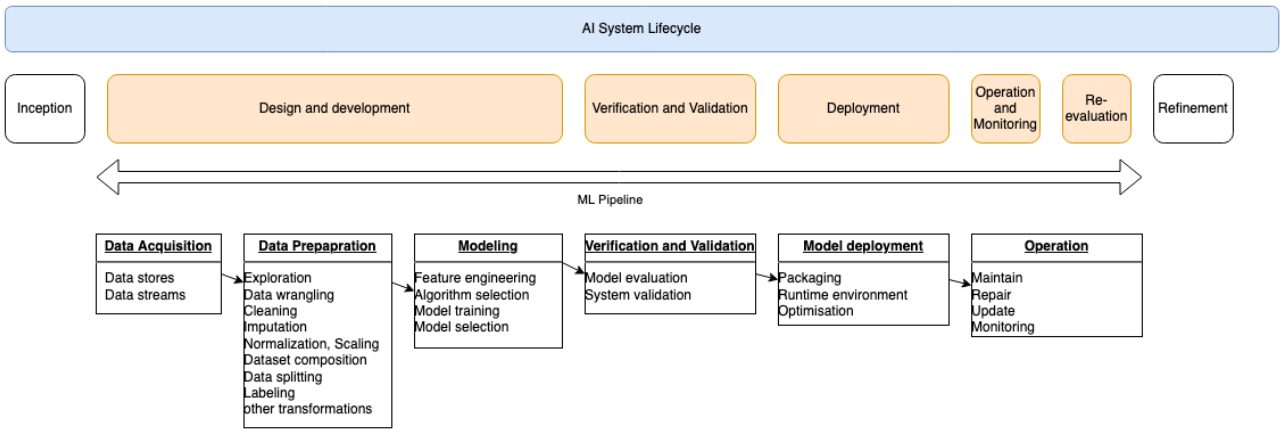

Figure 3. Functional groups of an AI System per ISO 23053, see also D 2.1 [15].

In short, the life cycle encompasses the following phases, as detailed in Deliverable D 2.1 [15]:

1) **Phase "Data Acquisition":** Data acquisition is the process of identifying, selecting and collecting data from various sources such as onboard or trackside sensors or diagnostics systems and converting it into a format that can be used for analysis or other purposes. The data acquisition process can be automatic or manual, depending on the nature of the data and the level of precision required;
2) **Phase "Data Preparation"**: Data preparation is the process of grouping, cleaning, transforming and raw data into a form that can be used for analysis or modelling. This phase also includes steps such as the filtering of data according to quality criteria, annotation (i.e. the addition of metadata with object labels), or data anonymisation;
3) **Phase "Modelling":** AI modelling is the process of building machine learning (ML) models that can learn from and make predictions on data. AI modelling involves several steps, including data preparation, feature engineering, model architecture design or model selection, and the actual training of the model on the data from the previous phase;
4) **Phase "Verification and Validation"**: This is the process of testing and evaluating machine learning models to ensure that they are accurate, reliable, and meet the specified requirements. This is an important step in the development of AI systems, as it helps to identify and correct errors and biases that could negatively impact the performance of the model when applied to actual rail operation scenarios. AI verification and validation also involves testing the ethical and social implications of the AI system, for instance ensuring that the AI system does not discriminate against certain groups of people or different railway undertakings, and that it meets standards and regulations;
5) **Phase "Model deployment"**: This is the process of integrating a trained machine learning model into a production environment, so that it can be used to make predictions on new, unseen data. The deployment process involves several steps, including selecting an appropriate deployment method, preparing the model for deployment by packaging it, and testing the model to ensure that it is functioning as expected;
6) **Phase "Operation"**: In this phase, the certified AI model is operated on rolling stock. Information is looped back into the ML pipeline and affects the data-acquisition and processing. This phase also involves the seamless monitoring of the performance and behaviour of the AI system to ensure safe operation.

3.1.2 Building Blocks of the Pan-European Data Factory

Based on the exemplary data life cycle for AI training shown in the past section, and on an elaborate identification and analysis of use cases in Deliverable D 1 [5], the key building blocks of the PEDF have been identified as shown in Figure 4, see also Deliverable D 2.1 [15].

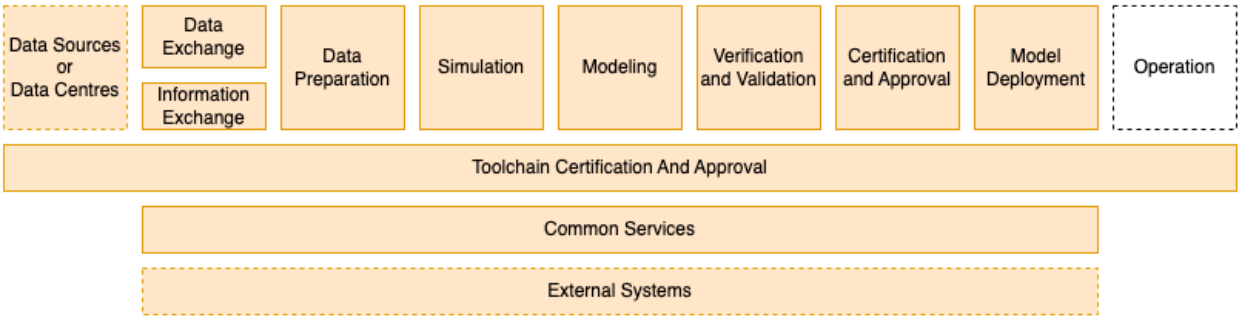

Figure 4. Identified high-level building blocks of the PEDF. “Operation” is display differently, as it is typically not seen as part of the PEDF itself.

These building blocks can be described as (with more details in Deliverable D 2.1 [15]):

- **Data Sources**: After offloading recorded data from a train, this has to be transferred to an individual Data Centre. Then, this data will be accessible for the PEDF;
- **Data Exchange**: The data exchange system enables the actual data (content) exchange between connected Data Factories as well as Data Centres and data Touch Points;
- **Information Exchange**: The purpose of this building block is to exchange (meta) information, e.g., on the availability of new data or the flagging of data. This can, e.g., be realised through a bus system in which all participating facilities are connected;
- **Data Preparation**: This building block takes care of aspects like dataset composition and data searching, and could also be more broadly described as data management functionality;
- **Modelling**: In this building block, data modelling is done, and AI models are trained. This block may be specific to individual Data Factories. Nevertheless, alignment on modelling and training toolchains throughout the PEDF appears highly beneficial;
- **Simulation**: In this building block, simulations are done to improve the behaviour of the AI system by training scenarios that cannot easily be implemented in real world tests. Furthermore, this allows for such trainings to be executed in parallel or in an accelerated manner, thereby allowing to accelerate the training and re-training of the AI;
- **Verification and Validation**: In this building block, model validation is ensured, for instance in accordance with EN 50126 (“Railway Applications. The Specification and Demonstration of Reliability, Availability, Maintainability and Safety (RAMS) Generic RAMS Process") [16]. Verification involves checking and assessing whether the requirements and specifications of the system have been properly implemented. It focuses on confirming that the system design and implementation align with the defined safety goals and standards. On the other hand, validation is concerned with demonstrating that the system meets the intended operational needs and performs its functions correctly;
- **Certification and Approval**: In this building block, the AI system is certified to be used in the railway environment. Depending on the required safety level, this may involve different procedures and independent entities to do assessments;
- **Model Deployment**: In the model deployment building block, models are made available for consumption. A key aspect here is that models shall be exchangeable between Data Factories;
- **Toolchain Certification and Approval**: Additionally, certification and approval may be required for toolchains and building blocks used in the entire process as described in the building block description for verification and validation;
- **Common Services**: A wide range of functionalities can be offered in the form of common services, for instance related to security, network access and access management, but also data import and distribution. It is expected that individual Data Factories may have specific own such services in place, nevertheless some should be defined on the level of the PEDF;

- **Operation**: In this building block, the operation of trained models take place. It is not directly part of the PEDF, but anyway listed here for completeness;
- **External Systems**: This building block provides basic services that are provided by external systems.

For each of the aforementioned building blocks of the PEDF, a more fine-granular breakdown has been performed, and the data flows in between these functions has been analysed, as detailed in Deliverable D 2.1 [15].

Further, for the majority of the building blocks, it has been investigated how these could be implemented, and which existing protocols or solutions could potentially be reused. For instance, for the Data Exchange building block, different push and pull options have been investigated, together with specific protocols and APIs that could be used. The reader is referred to Deliverable D 2.1 [15] for details.

It has to be noted that stated considerations on building blocks and in particular the referenced details on implementation options can only be seen as hypotheses, and are subject to open questions, such as where certain detailed data processing functions should take place in the PEDF.

3.1.3 Orchestration and Operational Considerations

To realise the aforementioned functions and building blocks in an efficient, scalable and future-proof way throughout the PEDF, the following considerations on PEDF orchestration and operation have been formulated, with details in Deliverable D 2.1 [15].

- **Data Gravity**: Given the gravitational pull of data, the compute environment in the PEDF should be designed for data-intensive workloads. It should prioritise proximity to data sources to minimise latency and optimise performance, where suitable. Given a multi-tenant approach, it may be possible to schedule trainings at locations where the required data is situated, thereby reducing data transfer demands;
- **Hybrid Cloud**: The system can use a hybrid Cloud approach, combining both on-premises infrastructure and public Cloud resources. It enables participating organisations to utilise the flexibility and scalability of the cloud while maintaining control over sensitive data;
- **Workload Scheduling**: An intelligent workload scheduler may be employed to distribute workloads across available compute environments, possibly situated in different Data Centres and different Data Factories, based on factors like resource availability, data locality, and workload priorities to optimise resource utilisation and minimise training time;
- **Multi-Tenancy**: The system supports multi-tenancy, allowing multiple users or organisations to securely share resources while maintaining isolation. It provides dedicated compute instances, storage, and network resources for each tenant, ensuring data privacy and resource allocation fairness;
- **Cloud Appliances**: To simplify deployment and configuration, pre-packaged Cloud appliances tailored for specific use cases may be utilised. These appliances consist of pre-configured software stacks, libraries, and frameworks optimised, e.g., for deep learning and AI workloads;

- **Dynamic Scaling to Cloud Resources**: The individual Data Factories may seamlessly scale compute resources based on workload demands. They intelligently provision additional cloud resources as needed by leveraging auto-scaling capabilities provided by the Cloud provider, ensuring cost-efficiency and flexibility based on the policies and requirements of the PEDF tenants;
- **Connectivity**: In order to facilitate the usage of Cloud resources and scaling, high-speed, low-latency connectivity between on-premises infrastructure and public cloud resources would be required;
- **Multi-Site Capabilities**: To address geographical redundancy and enable disaster recovery, Data Factories may incorporate multi-site capabilities. This is an important factor when it comes to replicating critical data and workload instances across geographically distributed sites, ensuring business continuity and minimising the impact of potential outages;
- **Maintenance**: The PEDF needs to incorporate mechanisms to announce maintenance in connected systems in order to ensure smooth operations. Each individual Data Factory must facilitate patch management, software upgrades, and hardware maintenance without disrupting ongoing workloads;
- **Environmental Considerations** The Data Centre operators need to take into account environmental considerations, such as power efficiency and cooling requirements, for instance by optimising power usage effectiveness (PUE) through energy-efficient hardware selection and intelligent power management techniques.

3.2 High-Speed Pan-European Backbone Network

The proposed High-Speed Pan-European (Railway Data Factory) Backbone Network is envisioned to be a transformative infrastructure designed to revolutionise data exchange across Europe's railway sector, in particular connecting individual Data Factories, Data Centres and Touch Points across Europe, but also stations, operational centres and maintenance facilities. This network, aimed at enabling efficient and secure data communication to power advanced machine learning and AI applications, is crucial for the digital transformation of European railways. It refers to a robust and ultra-fast telecommunications network that spans across Europe, connecting various Data Centres and railway operational networks.

Two approaches for a backbone network were presented in Deliverable D 2.3 [17]:

1. The construction of a network involves the use of existing dark fibre infrastructures of the railway infrastructure operators, such as those offered by SNCF or DB InfraGo AG (formerly DB Netz AG). The direct use of dark fibre can be cost-efficient and technically advantageous, as no civil engineering work is required, for example. Dark fibre offers high bandwidth, low latency and improved security, which are crucial for safety-critical applications in the context of GoA4. In addition, Dense Wavelength Division Multiplexing (DWDM) could be used to significantly increase bandwidth, albeit at a higher cost. As part of the digitisation of the rail network, infrastructure managers often need to build such fibre optic networks to connect their field elements to the control infrastructure;
2. Another approach would be to set up one or more internet exchange points, which are built by infrastructure managers who operate extensive fibre optic networks in the field. These Rail-Internet-Exchanges, referred to as Rail-IX, would serve as hubs for the railway networks

and the connection to the Internet Service Providers (ISPs). This facility would enable integration with public networks and allow connections from organisations such as industry partners. The implementation of Rail-IX was illustrated with several steps and components.

The backbone network is conceptualised with a set of foundational architectural principles, as detailed in Deliverable D 2.3 [17], and an advanced Identity and Access Management (IAM) framework as detailed in Section 3.3.2. The architectural principles guide the overall design and functionality of the network, ensuring it is efficient, secure, adaptable, scalable, and capable of meeting the diverse needs of the European railway sector.

In terms of physical implementation, the three variants "Utilisation of existing dark fibres", "Use of service providers" and "New network connection construction" were compared.

In the context of responsibility and organisational setup of a backbone network for the PEDF, three different approaches were considered from the perspective of the organisational network: the centralised approach, the federated approach and the hierarchical approach.

- In the **centralised approach**, all users work together in a single PEDF. Functions such as storage, computing resources and identity access management are centralised. A high-speed backbone network connects the data sources and ensures efficient data transfer. Although this approach enables centralised data management, it raises concerns about governance, ownership and organisational challenges. It can also limit data management and support for new or experimental data types;
- The **federated approach** breaks down data silos and merges existing Data Centres into a common PEDF. Members can use their own Data Centres or connect to existing participants, with data catalogues ensuring data sovereignty and data protection. A Pan-European backbone network facilitates data exchange and enables connection to other data sources. Members retain ownership and control of their data, following their own policies and frameworks;
- The **hierarchical approach** connects a network of Data Centres with a central data repository. Relevant data is consolidated for the pan-European system and the Data Centres perform AI tasks. While centralisation enables efficient data sharing, it can also incur costs for duplication and distribution. Only agreed data of sufficient quality is stored centrally, but members have the freedom to experiment with the data in their own systems.

Each approach involves trade-offs in data management, control and flexibility. The choice depends on the specific needs and considerations when implementing the PEDF.

In the context of establishing a backbone network for the PEDF, which facilitates seamless data exchange among various railway companies such as DB, SNCF, and NS/ProRail, the incorporation of a Zero Trust Architecture (ZTA) and Software-defined Networking (SDN) are of particular importance. Drawing upon insights gleaned from Deliverable D 2.2 [18] foundational network literature exemplified by Andrew S. Tanenbaum's seminal work "Computer Networks" [19], and the NIST Special Publication 800-207 “Zero Trust Architecture” [20], this architectural framework seeks to eschew implicit trust and instead places a stringent emphasis on robust authentication and authorisation mechanisms.

SDN, as a key component, empowers centralised network control, offering the agility, scalability, and flexibility required for dynamic network configuration. The synergy of SDN with Zero Trust augments the dynamic network management capabilities necessary for the effective implementation of Zero

Trust security protocols. This includes elements such as network segmentation, adaptive adjustment of access rights and continuous network monitoring.

Key technical architecture principles for the PEDF backbone network include:

1. **Decentralised Data Processing**: The network should utilise decentralised processing nodes for more efficient and proximate data handling (Data Touch Points);
2. **Security-Centric Network Design**: Utilising Zero Trust principles, the network should be designed without implicit trust. Every access point in the network must undergo authentication and authorisation processes;
3. **Network Segmentation and Microsegmentation**: These techniques enhance security and network access management by allowing fine-grained access control;
4. **Elastic Network Infrastructure**: Applying SDN principles allows for dynamic adjustments of the network to meet changing requirements;
5. **Interoperability and Standards**: Adhering to industry standards and protocols is critical for ensuring interoperability between different systems and organisations;
6. **Continuous Monitoring and Analysis**: Ongoing network monitoring is necessary to identify unusual activities and respond quickly to potential security threats.

These architectural principles are essential for constructing a robust, secure, and future-proof backbone network, meeting the demands of a modern, pan-European railway traffic system, which serves both fully automated driving and other railway-related services.

The high-speed network is essential for enabling real-time data analysis and decision-making, which is crucial for operational efficiency and safety in railway operations. It plays a pivotal role in facilitating cross-border railway operations by providing a consistent and reliable data communication infrastructure across Europe.

3.3 SECURITY FRAMEWORK AND IAM

The PEDF is expected to incorporate a robust security framework including a sophisticated Identity and Access Management (IAM) system. These elements are crucial for ensuring the secure and efficient handling of vast quantities of data across the European railway sector.

The IAM system defines clear roles, responsibilities, and access privileges for various stakeholders, ensuring that data access is tightly controlled and aligned with the specific needs of each user. Advanced authentication and authorisation mechanisms safeguard against unauthorised access.

In parallel, the security architecture of the PEDF adheres to stringent data protection regulations, including GDPR compliance, and employs a comprehensive suite of cybersecurity measures. These measures include encryption, intrusion detection, and continuous monitoring, crucial for protecting sensitive data from cyber threats. Additionally, the system encompasses risk management practices, incident response protocols, and disaster recovery strategies, ensuring data integrity, business continuity, and minimal service disruption.

Together, the security framework and IAM system form a critical foundation for the PEDF, enabling it to operate as a secure, compliant, and efficient data management hub within the European rail network. This integrated approach to identity management and security is pivotal in managing the

complex challenges of a large, decentralised, and highly regulated environment like the European railway sector. In the following sections, some more details on the aforementioned points are provided.

3.3.1 Overarching Security Framework

As introduced in Deliverable D 2.2 [18], the Cybersecurity framework for the PEDF is made-up of three pillars:

- An **organisational framework** defining security roles, processes, guidelines, management of cybersecurity risks, compliance and legal aspects;
- An **engineering process framework** defining a security process over the whole life cycle of the systems / products and components. As part of this framework, the connection and relationships between safety and security should also be addressed;
- **Continuous cyber security activities** such as security monitoring, continuous risk identification / mitigation and incident response / Business Continuity.

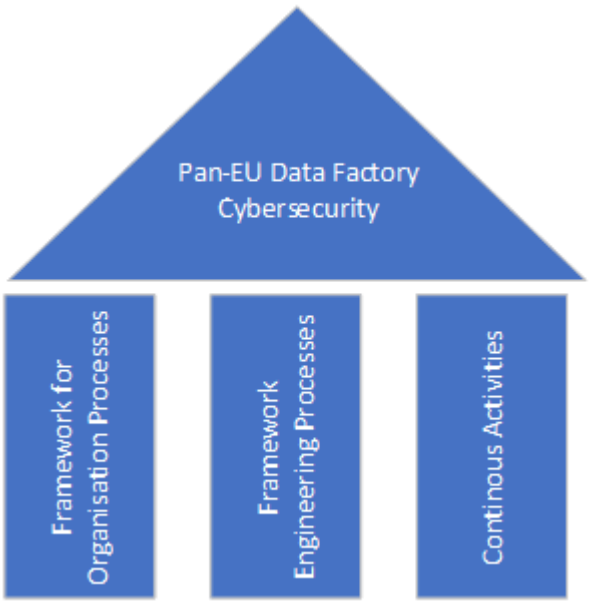

Figure 5. Cybersecurity framework of the PEDF [18].

The taxonomy of operational security and privacy considerations for the PEDF consists of five categories based on NIST SP 1500-4r2 “NIST Big Data Interoperability Framework: Volume 4, Security and Privacy” [21] and described in more detail in Deliverable D 2.2 [18]:

- **Risk Management:** A uniform methodology to evaluate the business risk level by conducting regular risk assessments shall be identified and applied throughout the PEDF. The purpose of this is to identify and evaluate threats exploiting technical and/or procedural vulnerabilities leading to the compromise of the confidentiality, integrity and availability of infrastructure, applications and/or data. A high-level threat analysis that has been conducted along these lines is described in Deliverable D 3.3 [22];

- **Infrastructure Management:** This involves security and privacy considerations related to hardware / software / network operation and maintenance, including threat and vulnerability management, monitoring and alerting, malware resilience, system redundancy and recovery;
- **Device & Application Management:** Devices and applications (incl. tool chains) shall be registered, and their configuration shall be managed along their whole lifecycles;
- **Data Governance** refers to the overall management of the availability, usability, integrity, and security of the data employed in the PEDF, as detailed in Section 3.4.
- **Identity and Access Management**, as stated before, is a cornerstone to provide secure access to the data and services of the PEDF, as detailed in Section 3.3.2**.**

3.3.2 Identity and Access Management (IAM)

Given the scale of the envisioned PEDF infrastructure, and the wide array of different users and stakeholders as listed in Section 2.2.1, the design and consequent application of an appropriate identity and access management is vital.

After a comparison of different approaches conducted in the project, it is recommended that the PEDF pursues a **federated IAM approach**, allowing different organisations to manage their own user identities while enabling interoperability and secure data sharing across the network. This appears best suited to recognise the diverse nature of the railway ecosystem, and the fact that the PEDF will likely build upon the connection of existing individual Data Factories, naturally already bringing along individual IAM approaches. In Deliverable D 2.2 [18] a potential federated IAM approach for the PEDF is investigated in detail, including detailed role definitions, considerations on a federated identity model, possible specific authentication and authorisation protocols, considerations on the life cycles of identities and the IAM as such, an analysis of potential IAM solutions, and detailed IAM requirements.

While the reader is referred to Deliverable D 2.2 [18] for details, some key paradigms of the envisioned IAM approach are listed here:

- **Roles and Responsibilities:** IAM in the PEDF is structured to define clear roles and responsibilities for various stakeholders, including data producers, data consumers, and governance bodies. This structure ensures that each entity involved in the data ecosystem understands its duties and the extent of its access permissions;
- **Access Controls and Privileges:** The system implements robust access controls, restricting data access based on user roles and responsibilities, and adhering to the principle of least privilege. This means users are granted only the access necessary to perform their job functions, minimising the risk of unauthorised data exposure;
- **Authentication and Authorisation:** The IAM framework employs advanced authentication and authorisation mechanisms. These include multi-factor authentication and role-based access controls, ensuring that only authorised personnel can access sensitive data and systems.

Overall, the IAM system ensures precise control over data access through clearly defined roles and responsibilities, coupled with advanced authentication and authorisation protocols. This setup guarantees that data access aligns with each user's specific role, enhancing security and operational efficiency.

3.4 DATA ARCHITECTURE, MANAGEMENT AND GOVERNANCE

The proposed data architecture and management of the PEDF can be described as a comprehensive, integrated system designed to optimise the handling, processing, and utilisation of railway data across Europe. Data management refers to the comprehensive process of collecting, storing, protecting, and responsibly sharing data with others. The concept of connected Data Centres presents a strategic approach for managing and utilising data that is stored in multiple Data Centres. Key aspects of this architecture are listed in following, while the reader is referred to Deliverable D 2.2 [18] for details.

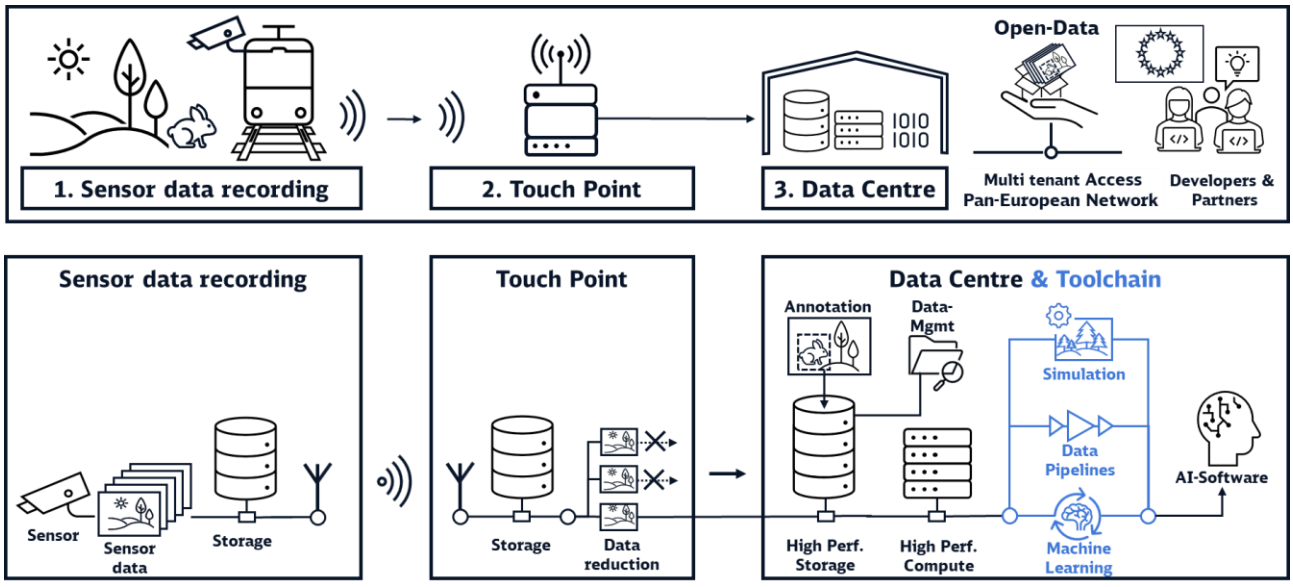

Figure 6. Flow of data from sensor recordings to trained AI models.

Data Collection on Train and Trackside

The architecture is built to accommodate data from a wide range of sources, including data from trains and infrastructure, diagnostic and environmental data. As also shown in Figure 6, the data from sensors on train and trackside is recorded and transferred to the Data Touch Point for further processing, once the initial data quality is checked and verified the data is transferred to the Data Centre.

A crucial aspect of the architecture is the standardisation of data formats. This ensures that data interoperability can be achieved.

Centralised and Decentralised Data Processing

The architecture employs both centralised and decentralised data processing. Centralised processing allows for comprehensive analysis and data analytics, while decentralised processing (at the edge) enables real-time, local data analysis. The edge components, such as Data Touch Points, adhere to a specific set of processes to assess data quality and the degree of encryption and decryption. This evaluation occurs prior to the data being transmitted to the Data Centre.

By harnessing the power of AI and machine learning, the system can conduct intricate analytics such as simulating the train drive and identifying objects on the tracks. The detailed information for data processing is also covered in Deliverable D 2.2 [18].

Data Storage and Accessibility

Considering the vast amount of data, scalable storage solutions are integral to the architecture. This includes Cloud-based storage and physical Data Centres distributed across Europe. These Data Centres are linked by a backbone network to facilitate seamless data exchange and transfer.

The architecture ensures efficient data access and retrieval mechanisms, allowing various stakeholders to access the data they need promptly and securely. For the detailed data categorisation, the reader is referred to Deliverable D 2.2 Section 4.4 [18].

Data Governance

A robust data governance framework is established to manage data ownership, access rights, and usage policies, ensuring compliance with regulations such as GDPR. A governance framework has been outlined in Deliverable D 2.2 Section 4.8 [18], providing a comprehensive overview of how data governance for the project will be implemented. An illustration of this is the Internal Governing Body and Central Governing Body at the consortium level.

Quality Control

Ensuring high data quality is paramount. The architecture includes mechanisms for data validation, cleansing, and standardisation to maintain data accuracy and reliability. These measures are defined on system level as well as sub system levels.

Security and Privacy

Given the critical nature of railway data, the architecture incorporates strong cybersecurity measures, including encryption, intrusion detection systems, and continuous monitoring.

In compliance with privacy regulations, the architecture incorporates features for anonymising data and ensuring the secure handling of personally identifiable information. Actively categorising the data and applying appropriate measures, including geofencing and similar methods, to bolster the security of the mentioned data.

Interoperability and Standards

The architecture is designed for cross-border compatibility, facilitating seamless data exchange and collaboration between different European countries and railway companies.

Aligning with international and European standards, the architecture adheres to interoperability and data exchange protocols, ensuring smooth integration with existing systems.

In essence, the data architecture and management system constitute a comprehensive and cutting-edge framework. Its design aims to fully leverage the potential of railway data, promoting improved operational efficiency, safety, and customer satisfaction throughout the European rail network.

4 ANALYSIS OF THE PAN-EUROPEAN DATA FACTORY FROM DIFFERENT PERSPECTIVES

Beside developing a possible technical design of the Pan-European (Railway) Data Factory (PEDF), the project has also investigated this from a broad array of non-technical perspectives.

For instance, the **operational challenges** that RUs are facing, e.g., related to the retrieval of onboard data, have been investigated in Deliverable D 3.1 [23]. The PEDF relies on a consistent flow of up-to-date data from European trains and its rail infrastructure. The diversity in maturity levels of rolling stock poses a challenge, as different countries use trains with varying ages and technology. Railway undertakings replace rolling stock through tenders, resulting in a mix of new and old trains in operation. Some rolling stock is equipped with advanced train-to-shore technology, while older ones lack such capabilities. Infrastructure also varies, with some areas having modern features like sensors and good mobile network coverage, while others need upgrades. The coexistence of technologies from different decades in the cross-border European rail network poses a challenge for the PEDF. Given the known bottlenecks of data application in rolling stock it is has been suggested to research the applicability of edge computing in the rail industry to resolve the known limits of current data communication in rail. Edge computing, in particular the currently developed Data Touch Point is a technology that enables data processing and analysis to be performed closer to the source of the data, rather than in a centralised location.

The potential **economic benefit** of an open data infrastructure as envisioned for the PEDF, was described in Deliverable D 3.2 [24]. The conclusion was that the PEDF has the potential to revolutionise data processing and sharing in the EU rail industry. Multiple factors need to be further investigated for creating an economic benefit: standardisation between rolling stock and wayside hosting, promotion of collaboration between various stakeholders, the possibility to scale the PEDF by design, to be aware of regulatory compliance and finally securing adequate funding options such as public-private partnership, grants and European funding.

A detailed **Cybersecurity risk analysis** for the PEDF has been conducted, as documented in Deliverable D 3.3 [22], following the STRIDE and Bowtie Risk Model approaches. Beside various measures related to network and application security, it has become apparent that data management is crucial for the PEDF, e.g.,

- **Universal Standards & Agreements** concerning data ownership, definitions, data formats, and data exchange protocols among all stakeholders to ensure seamless integration and cooperation;
- **Data Security & Risk Management** to address and mitigate data security threats, risks, and vulnerabilities associated with sophisticated models and large-scale data applications;
- **Regulatory Compliance & Risk Mitigation** to leverage applicable European standards and regulations to mitigate identified risks and enhance data exchange security.

In a legal study of the PEDF captured in Deliverable D 3.4 [25], typical data pathways from the recording of sensor data to the deployment of AI models in rail operation have been analysed w.r.t. applicable European or national regulations and laws. As Figure 7 shows, aspects like data protection are obviously relevant in the initial steps of data recording and annotation, but certification and liability play a role throughout the whole data pathway. Especially liability of course poses a challenge: If there is an accident in rail operation based on an AI model that was created with data and toolchains from different stakeholders, who is to be held liable? These and other legal aspects will be continued to be investigated after the project.

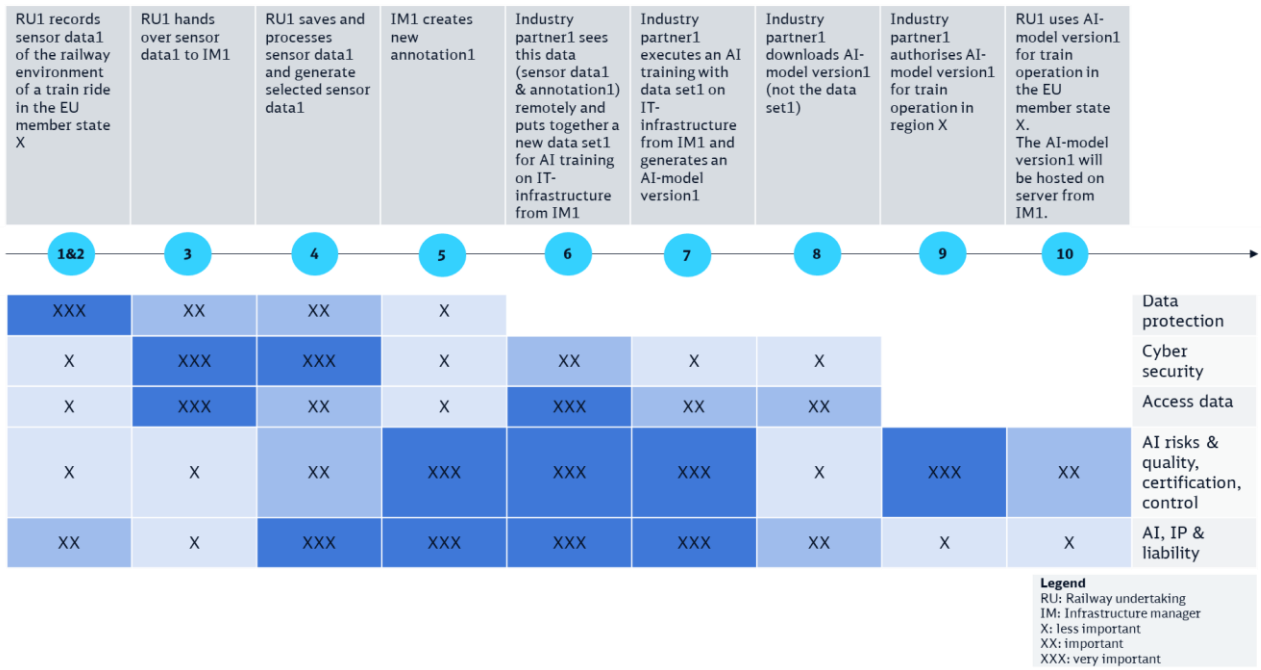

Figure 7. Analysis of the legal and regulatory aspects to be considered by the PEDF.

5 POSSIBLE DEPLOYMENT STRATEGY FOR THE PAN-EUROPEAN DATA FACTORY

This section outlines the strategic deployment plan for the Pan-European (Railway) Data Factory (PEDF), a key initiative aligned with the European Union's goals for digital transformation and innovation in the railway sector.

The roll-out strategy for the PEDF, as detailed in Deliverable D 4.2 [12] has been divided into three parts (short, medium and long term, see Figure 8), each addressing different aspects of its development and integration.

In the **short term**, the focus is on immediate needs, i.e., customised technical and legal solutions for individual (R&D) projects within the individual (somewhat "national") data factories. This means that those in need will seek or develop their solutions while mapping the contractual obligations to ensure that rights, liability, intellectual property and data ownership are delineated for the stakeholders in these projects.

In the **medium term**, the strategy aims to bring the ongoing activities and developments of the individual data factories in line with the standardisation requirements of the PEDF. This will be achieved through a transition that gradually aligns compliance with the standards and protocols agreed to be used in the PEDF, ensuring the eventual integration and seamless functioning of the various data factories within the PEDF framework.

Part of the long-term strategy is the comprehensive coordination of standardisation efforts with a focus on data quality, formats, interfaces and interconnectivity via a High-Speed Pan-European Backbone Network, as well as a uniform toolchain. In addition, an organisation responsible for these aspects should be created, ideally a European, independent and non-profit standardisation body.

Proposed participation paths for future PEDF members are the interface pillar and the toolchain pillar.

- The **interface pillar** emphasises the flexibility needed for the many different use cases and allows members to use their parameters to facilitate data exchange between PEDF members by common interfaces and standards. This approach is based on the coordination of data formats, data organisation, data quality, annotations, model architectures, data anonymisation and data privacy ensurance, sensors and data collection and ensures a smooth exchange of data between members and data factories;
- In contrast, the **toolchain pillar** aims to achieve harmonisation of the entire toolchain. While it offers versatility and the potential for full interoperability in data collection, processing, quality assurance, transfer, access and simulation, as well as training and evaluation of ML models, its rigid specification may only partially fulfil specific wishes of individual members.

The strength of the strategy lies in the pragmatic step-by-step development of an efficient cooperation and integration, which can develop the PEDF into an integrative, versatile and effective pan-European initiative.

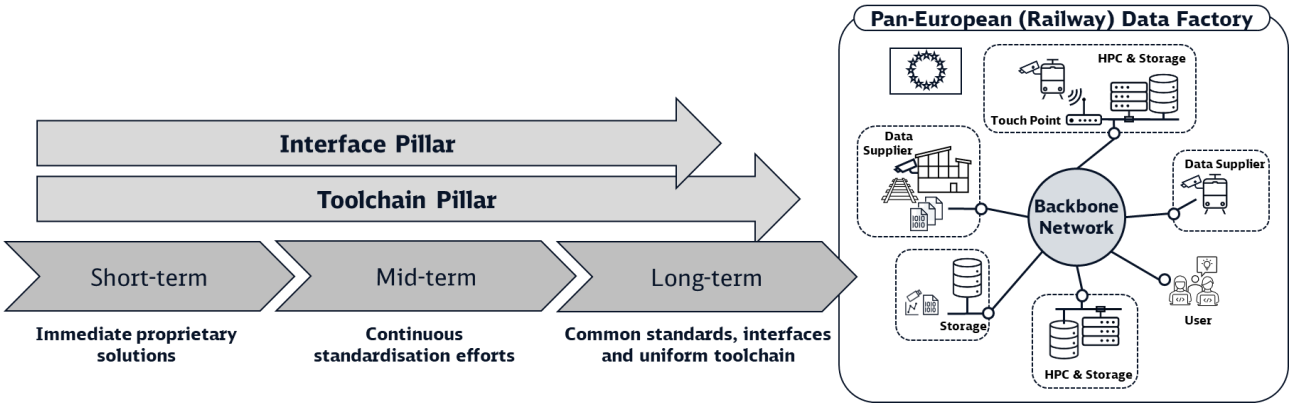

Figure 8. Short-, mid- and long-term strategy for setting up the PEDF supported by Interface- and Toolchain pillar.

In Deliverable D 4.2 [12], the envisioned phases of the deployment have also been further detailed, as shortly listed here:

Phase 1: Initial Analysis and Framework Development

- Needs Assessment: An exhaustive assessment to understand existing data management practices within the European railway sector, identifying key areas for improvement.
- Framework Formulation: Development of an overarching framework, addressing technical, operational, and regulatory dimensions, ensuring compliance with EU standards and regulations.

Phase 2: Pilot Implementation and Iterative Development

- Pilot Projects: Execution of targeted pilot projects across selected regions to evaluate practical viability and operational impact.
- Feedback Integration: Incorporation of feedback from pilot phases to refine the system, ensuring alignment with user needs and operational realities.

Phase 3: Infrastructure Development and Standardisation

- Network Infrastructure: Establishment of a high-speed backbone network, crucial for seamless data transfer and compliance with EU digital infrastructure goals.
- Data Centres: Deployment of strategically located Data Centres, both centralised and decentralised, equipped to handle large-scale data processing.

Phase 4: Stakeholder Engagement and Collaborative Frameworks

- Inclusive Involvement: Active participation of diverse stakeholders, including railway operators, technology providers, and policy makers, fostering a collaborative approach in line with EU principles of stakeholder engagement.
- Harmonisation Efforts: Efforts to harmonise operational procedures, ensuring interoperability and standardisation across member states.

Phase 5: Training, Capacity Building, and Compliance

- Skill Development: Comprehensive training initiatives to enhance user competencies and familiarity with the PEDF system.
- Regulatory Adherence: Ensuring strict adherence to EU regulations, particularly GDPR, reinforcing the commitment to data protection and privacy.

Phase 6: Gradual Rollout and Scalable Expansion

- Controlled Deployment: A phased rollout strategy to manage scalability and ensure a controlled expansion of the PEDF's capabilities.
- Adaptive System Design: Designing the system to be inherently adaptable, allowing for future technological integrations and sectoral evolutions.

Phase 7: Performance Monitoring and Continuous Enhancement

- Impact Assessment: Ongoing monitoring of the system's performance against predefined KPIs, ensuring alignment with EU objectives for digital innovation in transportation.
- Dynamic Evolution: Commitment to continuous improvement, reflecting the dynamic nature of technological advancements and sector-specific needs.

It should be noted that in the project also detailed thoughts were put into the potential process of adding new members to the consortium, as detailed in Deliverable D 4.1 [11].

The proposed deployment strategy for the PEDF is designed to align with the European Union's strategic objectives of enhancing digital infrastructure, fostering innovation, and improving operational efficiency in the railway sector. This multi-phased approach ensures a meticulous and inclusive process, paving the way for a transformative impact on Europe's railway systems.

6 OPEN POINTS

While the CEF2 RailDataFactory project could foster a common understanding on a Pan-European (Railway) Data Factory (PEDF), establish a basic architecture and clarify many details, many points were identified throughout the project that require further analysis and discussion, as elaborated in this chapter.

Network Implementation Options

1. Exploration of Network Models: Various network implementation options, including centralised, federated, and hierarchical models, have been identified, as mentioned in Section 3.2. Each presents unique advantages and challenges in terms of data management, governance, and flexibility. Deciding on the most suitable model requires careful consideration of these factors.

2. Security and Data Sovereignty: Ensuring data sovereignty and security is paramount, especially in the federated approach which currently appears most realistic for implementation. The network should be designed for compatibility and interoperability with initiatives like Gaia-X [26] and Catena-X [27], enhancing data sovereignty and transparency.

Technical and Operational Considerations

1. Infrastructure Choices: The options for building backbone networks, such as utilising existing dark fibres, contracting service providers, or constructing new network connections, need thorough evaluation. The decision hinges on factors like specific requirements of the railway infrastructure and the balance between control, security, flexibility, and scalability.

2. Operational Networks Governance: Further investigation is needed in the governance of operational networks, particularly concerning data management and exchange within these networks. This includes establishing clear protocols and procedures for data sharing, ensuring data integrity and compliance with security standards.

Commercial and Legal Aspects

1. Cost Analysis: A comprehensive cost analysis is essential to evaluate the economic viability of different network implementation options. This includes considering the initial investment, ongoing operational costs, and potential long-term savings or benefits.

2. Legal and Regulatory Compliance: There is a need for an in-depth legal assessment, especially concerning cross-border data sharing and compliance with EU-wide regulations like GDPR. This encompasses understanding the legal implications of different network models and ensuring that the chosen model adheres to EU standards.

Security and Privacy

1. Zero Trust Concepts in Network Design: Implementing zero trust concepts in the network, particularly in the context of a federated approach, is vital. This involves ensuring secure data transmission and handling without necessarily establishing trust relationships between all network participants.

2. Balancing Security and Accessibility: The security architecture must balance robust protection with the accessibility needs of various stakeholders, requiring a nuanced approach to network security and data privacy.

Further Research and Development

1. Continuous Technological Assessment: Ongoing evaluation of emerging technologies and network solutions is necessary to ensure that the PEDF remains cutting-edge and adaptable to future advancements.

2. Stakeholder Engagement: Engaging with key stakeholders to gather insights and feedback on network implementation, operational challenges, and governance structures is crucial for the project's success.

These open points highlight the need for continued research, development, and collaboration to address the complexities and challenges of implementing the PEDF. Addressing these points will be essential for aligning the project with European strategies and objectives, particularly in terms of innovation, digital transformation, and cross-border collaboration in the railway sector.

7 CONCLUSION

As we synthesize insights from the comprehensive RailDataFactory CEF2 project documents, it becomes unequivocally clear that the development of the envisioned Pan-European (Railway) Data Factory (PEDF) is not just a progressive step but a monumental leap forward for the European railway system. This initiative stands at the vanguard of innovation, marking a pivotal moment where Europe can assert its position as a global leader in technological advancement and market strength within the rail industry.

Innovation and Market Leadership

The PEDF embodies the essence of European ingenuity. By harnessing cutting-edge technologies, sophisticated data management, and collaborative frameworks, this initiative is set to transform the European railway landscape. Its successful implementation will propel Europe to the forefront of railway innovation, showcasing a model of efficiency, safety, and reliability that can be emulated globally. The PEDF is not just an infrastructural project; it's a testament to Europe's commitment to leading the digital revolution in transportation.

Catalyzing Economic Growth and Competitiveness

The economic implications of the PEDF are profound. It promises to streamline operations, reduce costs (compared to the case where stakeholders like IMs, RUs, suppliers, etc., all establish independent Data Factories without exploitation of synergies among these), and open new avenues for business models and services within the railway sector. This initiative is a critical driver for the European market, bolstering competitiveness and fostering economic growth. It will enable European railway businesses to offer unparalleled services, enhancing customer satisfaction and cementing Europe's status as a hub of commercial excellence in rail transport.

Advancing Climate Policy and Sustainable Transportation

In the face of global climate challenges, the PEDF aligns perfectly with the European Union’s climate policy goals. By optimising railway operations and resource utilisation, it contributes significantly to reducing the carbon footprint of transportation. The PEDF supports sustainable mobility solutions, furthering the EU's commitment to environmental stewardship and green transportation initiatives. The PEDF is not just about modernising railways; it's about shaping a sustainable future with new functionalities and enabler of European data collaboration for coherent transportation system in Europe.

The need for Europe to support and drive forward the PEDF is clear and requires a joint project involving a whole range of European organisations, including policy makers, industry pioneers, innovators and the general public. The commitment to this initiative goes beyond simply improving railway infrastructure; it symbolises Europe's determination to establish itself as a leader in technological innovation, economic progress and environmental protection.

Moving toward fully automated Rail Transport in Europe

In this report, the PEDF proves to be a key element of the strategic development towards efficient and ecological fully automated rail transport in Europe. This will sustainably stimulate the economic dynamics in railway development, which among other things will serve as a linchpin for sustainable transport.

This initiative should not only be seen as an infrastructural improvement of the railway sector, but rather represents a profound investment in the future of Europe. A future in which Europe is not just a participant, but a pioneer and a role model on the world stage, especially in the field of rail excellence. Now is a crucial time for Europe to take the initiative and help itself achieve unprecedented levels of innovation and sustainability in the rail sector.

REFERENCES

[1] Shift2Rail program, see https://rail-research.europa.eu/about-shift2rail/

[2] Europe's Rail program, see https://projects.rail-research.europa.eu/

[3] Sensors4Rail project, see "Sensors4Rail tests sensor-based perception systems in rail operations for the first time," Digitale Schiene Deutschland, 2021. [Online]. Available: https://digitale-schiene-deutschland.de/en/Sensors4Rail

[4] DIRECTIVE (EU) 2016/797 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL, see https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016L0797

[5] CEF2 RailDataFactory Deliverable 1, "Data Factory Concept, Use Cases and Requirements", Version 1.1, April 2023. [Online]. Available: https://digitale-schiene-deutschland.de/en/news/2023/Pan-European-Railway-Data-Factory

[6] Shift2Rail TAURO project, Horizon 2020 GA 101014984, see https://projects.shift2rail.org/s2r_ipx_n.aspx?p=tauro

[7] P. Neumaier, "First freely available multi-sensor data set for machine learning for the development of fully automated driving: OSDaR23", 2023. [Online]. Available: https://digitale-schiene-deutschland.de/en/news/OSDaR23-multi-sensor-data-set-for-machine-learning

[8] Open Sensor Data for Rail 2023, 2023. [Online]. Available: https://data.fid-move.de/dataset/osdar23

[9] R2DATO project, see https://projects.rail-research.europa.eu/eurail-fp2/

[10] P. Neumaier, "Data Factory - "Data Production" for the training of AI software," Digitale Schiene Deutschland, 2022. [Online]. Available: https://digitale-schiene-deutschland.de/en/news/2023/Pan-European-Railway-Data-Factory

[11] CEF2 RailDataFactory, D 4.1 – "Deployment activities description for a pan-European Railway Data Factory", December 2023. [Online]. Available: https://digitale-schiene-deutschland.de/en/news/2023/Pan-European-Railway-Data-Factory

[12] CEF2 RailDataFactory, D 4.2 – "Pan-European Railway Data Factory deployment planning and strategy proposal", December 2023. [Online]. Available: https://digitale-schiene-deutschland.de/en/news/2023/Pan-European-Railway-Data-Factory

[13] FRMCS, see https://uic.org/rail-system/telecoms-signalling/frmcs

[14] ISO/IEC 23053:2022, "Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)", 2022

[15] CEF2 RailDataFactory D 2.1 – "Technical specifications and available solutions for building blocks, components, Cloud / hybrid-Cloud and Edge-Orchestration & Operational concept", August 2023. [Online]. Available: https://digitale-schiene-deutschland.de/en/news/2023/Pan-European-Railway-Data-Factory

[16] EN 50126, "Railway Applications. The Specification and Demonstration of Reliability, Availability, Maintainability and Safety (RAMS) Generic RAMS Process", 2018

[17] CEF2 RailDataFactory D 2.3 – "High-speed pan-European Railway Data Factory Backbone Network", August 2023. [Online]. Available: https://digitale-schiene-deutschland.de/en/news/2023/Pan-European-Railway-Data-Factory

[18] CEF2 RailDataFactory D 2.2 – “Technical specifications and available solutions for Identity Access Management (IAM), Data Management and Transfer and Cyber-Security”, August 2023. [Online]. Available: https://digitale-schiene-deutschland.de/en/news/2023/Pan-European-Railway-Data-Factory

[19] A. Tanenbaum and D. Wetherall, “Computer Networks”, Prentice Hall, Boston, 5th edition, 2011

[20] NIST, “Zero Trust Architecture”, NIST Special Publication 800-207, August 2022, see https://doi.org/10.6028/NIST.SP.800-207

[21] NIST, “NIST Big Data Interoperability Framework: Volume 4, Security and Privacy”, NIST SP 1500-4r2, September 2019, see https://doi.org/10.6028/NIST.SP.1500-4r2

[22] CEF2 RailDataFactory D 3.3 – “Description of cybersecurity vulnerabilities, threat scenario’s and usable standards to mitigate associated risks”, November 2023. [Online]. Available: https://digitale-schiene-deutschland.de/en/news/2023/Pan-European-Railway-Data-Factory

[23] CEF2 RailDataFactory D 3.1 – “Report of bottlenecks data application in rolling stock”, November 2023. [Online]. Available: https://digitale-schiene-deutschland.de/en/news/2023/Pan-European-Railway-Data-Factory

[24] CEF2 RailDataFactory D 3.2 – “Business case whether open data infrastructure would be attractive for European rail”, November 2023. [Online]. Available: https://digitale-schiene-deutschland.de/en/news/2023/Pan-European-Railway-Data-Factory

[25] CEF2 RailDataFactory D 3.4 – “Legal and regulatory assessment catalogue”, November 2023. [Online]. Available: https://digitale-schiene-deutschland.de/en/news/2023/Pan-European-Railway-Data-Factory

[26] GAIA-X, see https://gaia-x.eu/

[27] CATENA-X, see https://catena-x.net/en/