CEF2 RailDataFactory

Deliverable 2.1 – Technical specifications and available solutions for building blocks, components, Cloud / hybrid- Cloud and Edge-Orchestration & Operational concept

Due date of deliverable: 30/04/2023

Actual submission date: 30/06/2023

Resubmission date: 11/08/2023

Leader/Responsible of this Deliverable: Julian Wissmann (WP 2 lead) / DB Netz AG

Reviewed: **Y**/N

Document status		
Revision	Date	Description
01	09/03/2023	Document template generated
02	26/05/2023	Content transferred from Confluence
03	26/05/2023	First draft complete
04	06/06/2023	Version submitted to advisory board
05	29/06/2023	Final version after addressing all advisory board comments
06	30/06/2023	Version submitted to the project officer
07	03/07/2023	Adjusted to high resolution images and added header and footer
08	11/08/2023	Disclaimer updated based on the feedback of the granting authority

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: 9 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
Alexander Heine	DB
Jens Dalitz	DB
Julian Wissmann	DB
Philipp Neumaier	DB
Wolfgang Albert	DB
Patrick Marsch (only editorial)	DB

Note of Thanks

We would like to thank our Advisory Board Members Maria Aguado, Saro Thiyagarajan, Oliver Lehmann and Manuel Kolly for the valuable discussion and in particular Xiaolu Rao and Janneke Tax for their thorough reviews of this deliverable and input to this work!

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 Rail 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 focuses 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 Rail Data Factory.

In this deliverable of the study, the high-level architecture and building blocks of the pan-European Rail Data Factory are introduced, considerations on implementation with market available solutions, as well as considerations on operation and orchestration are provided. Altogether, these items serve as a basis for the further work in this study.

ABBREVIATIONS AND ACRONYMS

Abbreviation	Definition
AAD	Azure Active Directory
AI	Artificial Intelligence
API	Application Programming Interface
CEF	Connecting Europe Facilities
CLI	Command Line Interface
ERA	European Union Agency for Railways
FTP	File Transfer Protocol
GoA4	Grade of Automation 4
GUI	Graphical User Interface
HADEA	European Health and Digital Executive Agency
HiL	Hardware in the Loop
HTTP	Hypertext Transfer Protocol
IAM	Identity and Access Management
IM	Infrastructure Manager
JSON	JavaScript Object Notation
ML	Machine Learning
MPLS	Multiprotocol Label Switching
REST	Representational State Transfer
PKI	Public Key Infrastructure
RU	Railway Undertaking
S3	Simple Storage Service
VPN	Virtual private Network

TABLE OF CONTENTS

LIST OF FIGURES

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 localization 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, railway undertakings, 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, in particular, 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 focuses exactly on aforementioned vision of a pan-European Rail Data Factory for the joint development of fully automated driving. The study, being co-funded through HADEA, aims to assess the feasibility of a pan-European Rail Data Factory from technical, economical, legal, regulatory and operational perspectives, and determine key aspects that are required to make a pan-European Rail Data Factory a success. For a better understanding of the studys aim and scope, please see Chapter 1.1 in Deliverable 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 artificial intelligence (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 puts special emphasis on the **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 Data Factory, and also investigates **commercial, legal and operational aspects** that have to be addressed to ensure that the vision of the pan-European Rail Data Factory 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 will be utilized in the Data Factory of Digitale Schiene Deutschland 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 pan-European Rail Data Factory. Here, however, the main focus is on creating first implementations of individual data centers 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 pan-European Rail 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 has its 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", Deutsche Bahn already started to set up some components of the data center in Germany [10].

1.3 AIM AND STRUCTURE OF THIS DELIVERABLE

This current document is the deliverable D2.1 of the CEF 2 RailDataFactory project, covering the architecture and building blocks of the envisioned pan-European Data Factory specifically aimed at analyzing underlying concepts and zones and providing an analysis of required building blocks and potentially available solutions.

The aim of the document is to obtain early feedback and possible additions from the sector on the architecture and building blocks, in order to update the work accordingly and consider the obtained input in the subsequent phases of the project, in which the detailed IAM and data management concept, legal and business aspects will be developed.

The remainder of this document is structured as follows:

- In Chapter 2, the system context is described;
- In Chapter 3, the functional zones of the system are derived and analysed;
- In Chapter 4, building blocks are derived;
- In Chapter 5, implementation considerations and available solutions are provided;
- In Chapter 6, the topics orchestration and operation are further discussed;
- In Chapter 7, a summary is provided.

2 SYSTEM CONTEXT

The system context of the pan-European Rail Data Factory is defined by the connection of multiple data centers. It is important to emphasize, that these data centers do not necessarily consist of exactly the same major subsystems, such as high performance computing or AI processing. A participant may for example decide to not operate its own data center and instead lease that capability from another data center provider of the pan-European Rail Data Factory. Likewise, the number of data sources and Data Touch Points operated may vastly differ based on the railway infrastructure needs. Countries will have various dimensioning requirements, i.e. depending on network size and network usage. Figure 1 gives an overview over the system-context of the pan-European Rail Data Factory, with it's member systems and their subsystems. The dotted lines indicate that a component is optional.

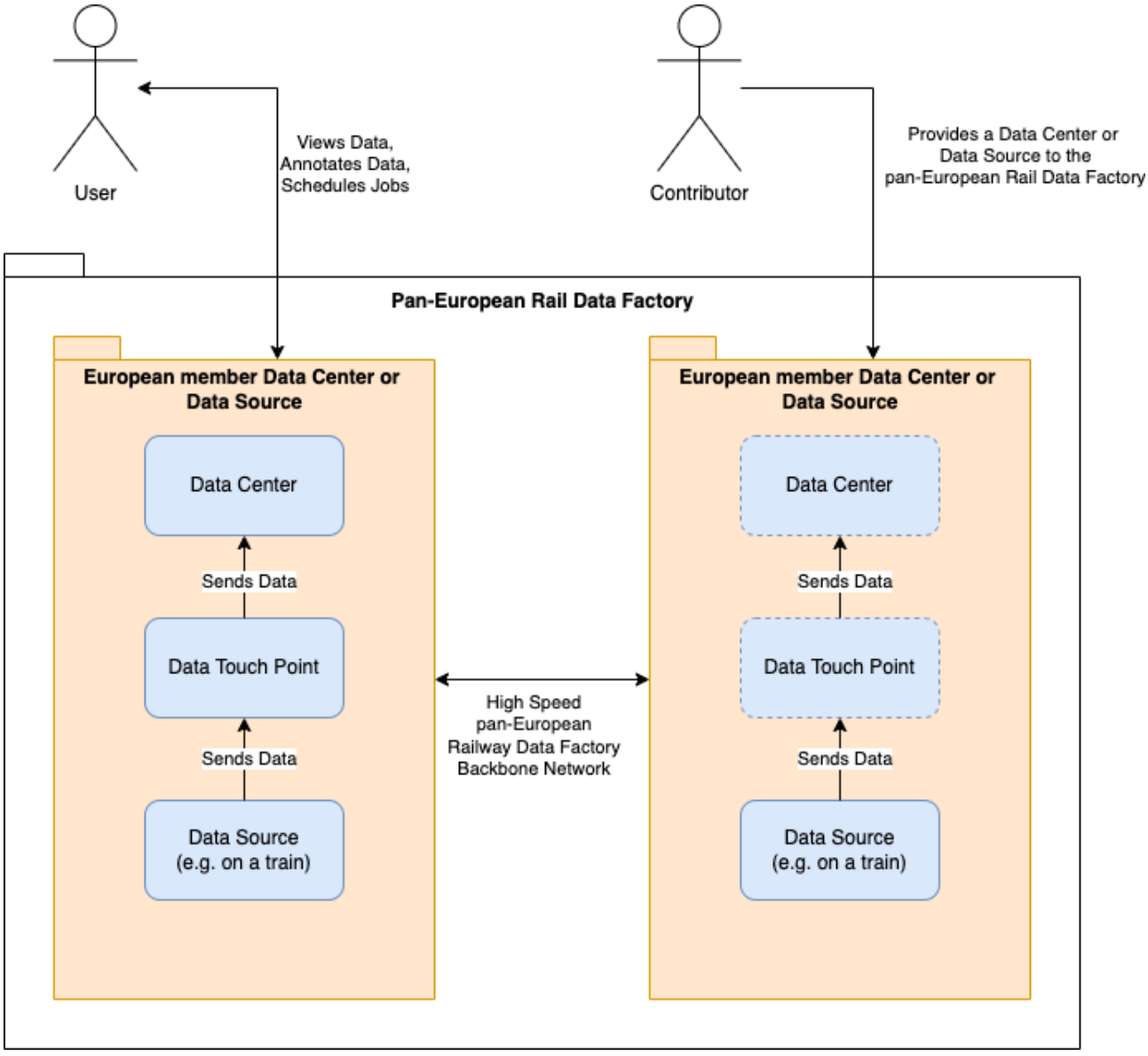

Figure 1: Overview System-Context of pan-European Rail Data Factory (Member Systems and Subsystems).

For differentiation of user and contributor and how to contribute within the pan-European Rail Data Factory, please see Deliverable 1 Chapter 4 [5].

The zones defined in the Pan-European Railway Data Factory can be understood in the following way:

Data Factory

The Data Factory is a linkage of different systems into a processing platform for data. It consists of multiple **Data Sources**, e.g. cameras, lidars or from localisation systems on trains as well of other sources from the trackside depending on the use cases to be implemented, that produce sensor data like pictures or position information. This use case specific data is collected and later on used in AI training in the **Data Center** to create AI models for autonomous driving.

Due to the requirements of partially or fully autonomous driving, the volume of data is huge. The sheer number of high quality sensors required for it, mean that the data collected from the source cannot be transferred to the **Data Center** in real time. Rather, the data must be transmitted in batches from the train, e.g. while it is in the depot, as transmission is only possible to a limited extent while the train is in motion. It also makes sense to carry out initial processing steps, e.g. for data selection and reduction, and thus minimize the costs and effort involved in the transfer. For this purpose, an additional track-side system is provided, the **Data Touch Point**.

Pan-European Rail Data Factory

This system consists of multiple data centers and a high-speed connection between these, the High-speed pan-European Railway Data Factory Backbone Network, and subsystems for an overarching secure access control and data management system. Analogous to the European rail network, it maps the connection of data between the different regions of Europe and supports cross-border data and rail traffic. This will serve as a basis for cross-border development and thus enable seamless transport within Europe.

3 FUNCTIONAL ZONES

To understand the pan-European Rail Data Factory, it is of utmost importance to better understand the context in which the system will be used. According to the European Commission's European Strategy for Data [11], the European Data Space shall support data flows across the sector through data frameworks and data governance. Fragmentation between member states is also identified as a major risk. Within the contextual classification of data and data flows, a functional model emerges. This gives the possibility to generate so-called functional zones that represent contextual data and metadata management in combination with transitions between storage and processing. This concept is the basis for the definition of a pan-European rail data governance and thus the representation and definition of responsibilities. This will make it easier to manage data in a structured way and to determine who is responsible for its management. From a functional point of view, the system is intended to provide an environment for AI training and homologation, that can be used in GoA4 railway operations. Therefore, the typical lifecycle of such an AI system needs to be understood and brought into the broader context described in Chapter 2.

3.1 AI SYSTEM ZONES AND LIFECYCLE PHASES

As the purpose of the system is the development (and homologation) of AI functions for the operation of trains, the system lifecycle can also be defined based on the "framework for artificial intelligence system using machine learning" according to ISO 23053 [12], which DBs internal governing body for IT architecture has also applied to a reference architecture for AI systems.

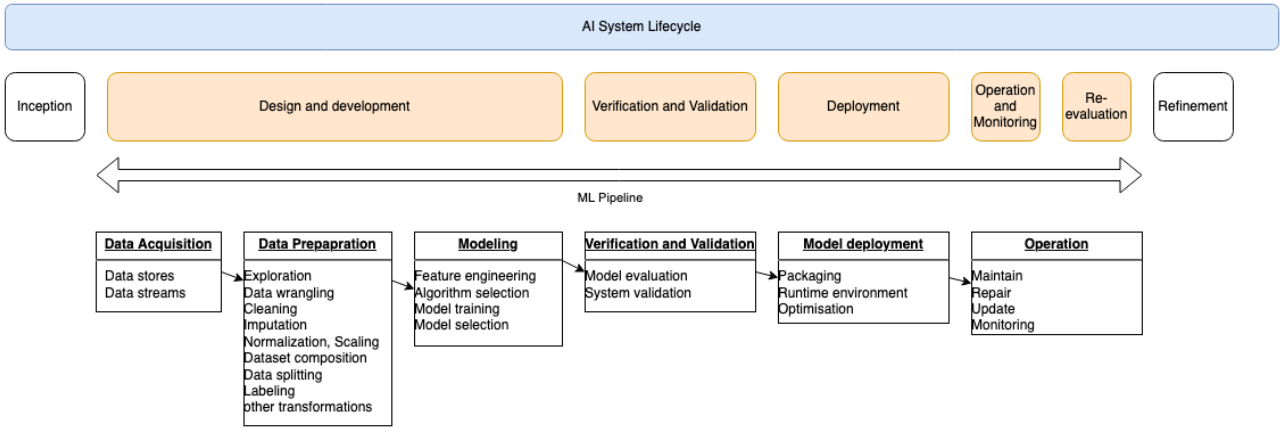

Figure 2: Functional groups of an AI System per ISO 23053, DB Reference Architecture AI.

Phase "Data Acquisition"

Data acquisition is the process of collecting data from various sources and converting it into a format that can be used for analysis or other purposes. This process involves identifying and selecting relevant data sources, collecting the data from those sources, and then transforming and storing the data in a suitable format. The sources of data acquisition can vary depending on the type of data needed, but common sources include sensors, databases, diagnostic platforms, and other types of digital sources. The data acquisition process can be automatic or manual, depending on the nature of the data and the level of precision required.

Overall, data acquisition is a critical step in the data management process as it determines the quality and usefulness of the data that will be used for analysis or decision-making purposes, like the building of AI models. Data acquisition is done either in the context of an individual data center by collecting data from own data sources, or by taking data from other data centers.

Phase "Data Preparation"

Data preparation is the process of cleaning and transforming raw data into a form that can be used for analysis or modelling. It is a critical step in the data analysis pipeline, as the quality and accuracy of the resulting analysis or created models depends heavily on the quality and accuracy of the used data. In this phase, data is being transformed and annotated (labelled), so that the data is enhanced and quality-wise sufficient for an appropriate AI model training. Also grouping of data in data sets is performed in order to have curated and high quality training sources. Should data transformations be necessary in the scope of data acquisition, e.g. compression or encryption/decryption, it should be handled there or in a post-process of the acquisition itself. The same goes for filtering of data, e.g. to filter out data that does not meet quality criteria or which can be identified as unneeded in the acquisition phase.

The data preparation process typically involves the following steps:

- Data enrichment: Gathering additional metadata from various sources, such as infrastructure data, weather data and others;
- Data cleaning: Removing errors, inconsistencies, and duplicates from the raw data. This includes handling missing data, dealing with outliers, and correcting formatting or encoding errors;
- Data anonymization: Removing or obscuring personal identifiable information (PII) from data to protect the privacy of individuals. This involves either removing or encrypting any information that could be used to identify an individual, such as names, faces, body structures, social or ethical characteristics;
- Data transformation: Converting the cleaned data into a structured format that is suitable for analysis. This may include aggregating data, merging multiple data sources, and applying calculations or transformations to the data;
- Data integration: Combining multiple data sources into a single dataset that can be used for analysis or modelling;
- Data reduction: Reducing the size of the data by selecting relevant variables or samples, or by summarizing the data using statistical methods;
- Data annotation: Metadata is added to the raw data to label objects within the raw data;
- Data formatting: Preparing the data in a format that is compatible with the analysis or modelling tools being used;
- Data annotation: Metadata is added to the raw data to describe objects container within.

Overall, effective data preparation is essential for generating accurate and reliable insights from data. It requires careful attention to detail and a good understanding of the data being analysed, as well as the tools and techniques used for data analysis. All these steps happen within a Data Factory and the data touch point.

Phase "Modelling"

AI modelling is the process of building machine learning (ML) models that can learn from and make predictions on data. The goal of AI modelling is to create a model architecture that is capable of accurately predicting outcomes based on patterns and relationships identified in the input data and to train this.

AI modelling depends on several steps, including data preparation, feature engineering, model architecture design or model selection, and training. Feature engineering involves selecting and creating features that will be used by the model to make predictions. Model selection involves choosing the type of machine learning algorithm that is best suited for the problem at hand, such as regression, classification, clustering or even designing a model architecture from scratch.

The final step in AI modelling is training the model on the input data (processed data from the sources plus respective annotations). During training, the model learns to recognize patterns and relationships in the data and adjusts its internal parameters to improve its accuracy. These steps happen within a data center; however, data acquired in another Data Factory may be required in this phase. Once the model is trained, it can be used to make predictions on new, unseen data.

Phase "Verification and Validation"

AI verification and validation 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.

- Verification involves testing the AI system to ensure that it meets the specified requirements and performs as expected. This involves testing the software code and algorithms to ensure that they are correct and meet the design requirements. Verification also involves checking that the system is robust and can handle a range of inputs and conditions;
- Validation involves testing the AI system to ensure that it is accurate and reliable in real-world situations. This involves testing the system with real-world data and scenarios, as well as HIL testing to evaluate its performance and identify any issues or biases that may be present.

In particular, AI verification and validation also involves testing the ethical and social implications of the AI system. This includes ensuring that the AI system does not discriminate against certain groups of people, different railway undertakings or perpetuate biases, and that it meets standards and regulations.

Overall, AI verification and validation is a crucial step in the development of machine learning models and AI systems. By ensuring that these systems are accurate, reliable, and meet ethical and regulatory requirements. This step happens within a data center however data acquisitioned in another data center may be required in this phase.

Phase “Model deployment”

Model deployment 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.

During the packaging stage, the model is optimized for deployment, which involves converting it into a format that can be used by the production environment. This may include converting the model into a compressed format that is optimized for speed and performance or converting it into a containerized format that can be easily deployed in a modular environment, such as the Safe Computing Platform [13] and made accessible to relevant partners.

Once the model is packaged, it is deployed and integrated with the other components of the system. Testing is then performed to ensure that the model is functioning correctly and that it is delivering accurate predictions. This can be done in the data centre in a coupled simulation environment block, as a physical field test or in a Hardware-in-the-Loop (HiL) environment.

Overall, model deployment is a critical step in the machine learning lifecycle, as it is the handover point from which a model can not only be brought into testing environments but also onto rolling stock.

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.

As soon as a model is transferred to productive operation, the performance and behaviour of these AI systems must also be monitored seamlessly in order to ensure safe operation. This means that the necessary measures such as logging and monitoring must also be implemented here and ongoing maintenance to address potential risks and ensure that they operate effectively and responsibly.

The path of an update must also be considered here, because re-training of models will also be required in the future due to external influences and feedback of knowledge gained, and these re-trained models must then be transferred back into productive operation.

Likewise, monitoring and alerting are part of the monitoring methods, so that the metrics of the models can be monitored and analysed and countermeasures applied if necessary.

This model of the ML pipeline does not take supporting processes into account, which is why function groups for data distribution or communications cannot be allocated in it.

3.2 NETWORK ZONES

Network zones refer to an architecture strategy where networks are divided into different zones to improve security and protect data. Each zone has different access rights and security controls to limit the spread of threats within the network and protect sensitive information. By implementing

network zones, the data centers can restrict traffic between different zones and increase security controls and monitoring. This helps to minimise the risk of cyberattacks, data theft, and other threats, and ensure that confidential information remains secure. Further, it supports the data distribution processes regarding how to exchange data and information.

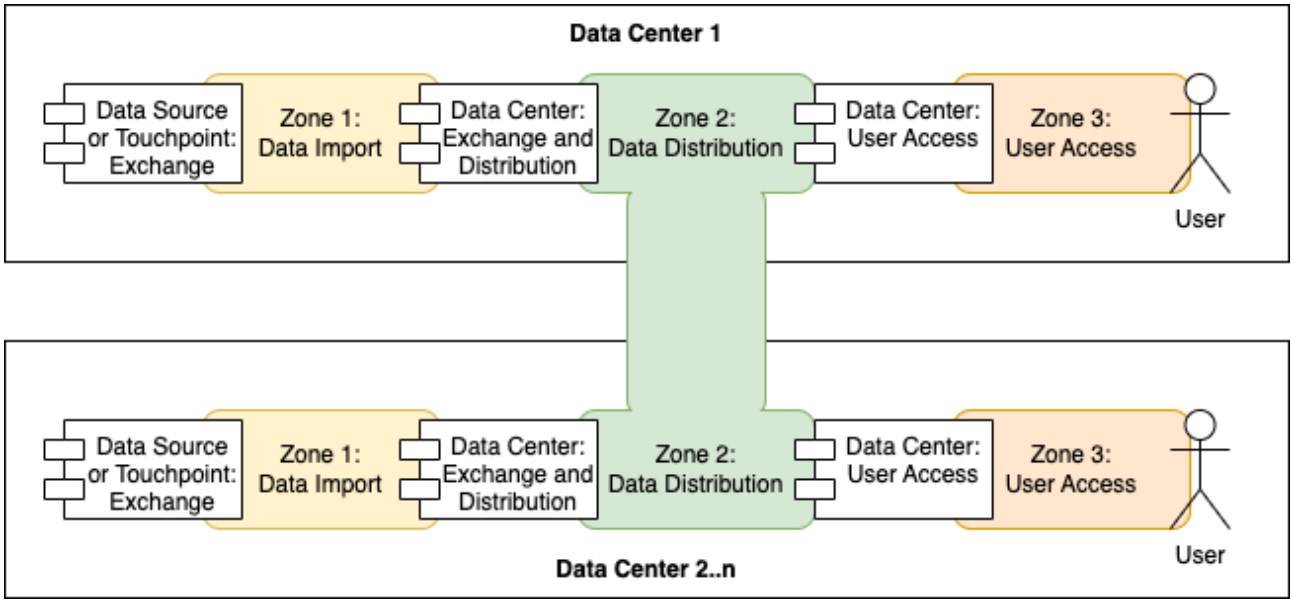

Figure 3: High level network zones of the pan-European Rail Data Factory.

Zone 1 "Data Import"

A data import network zone is a dedicated area within a network architecture that is designed to facilitate the secure transfer of data from external sources into an internal network, such as from a partner or a third-party service. The data import network zone typically includes specific security controls, such as firewalls, intrusion detection, and prevention systems, that are configured to minimize the risk of unauthorized access or cyberattacks. By implementing a data import network zone, organizations can ensure that the flow of data into their systems is managed in a controlled and secure manner, and sensitive information is protected from potential security breaches.

Zone 2 "Data Distribution"

A data distribution network zone is a dedicated area within a network architecture that facilitates the secure distribution of data from a central source to other parts of the network. This zone is designed to provide controlled data distribution, while minimizing the risk of unauthorized access or cyberattacks. In this network zone also the data and information exchange between data centers is handled.

Zone 3 "User Access"

A user access zone is a dedicated area within a network architecture that allows users to access specific resources, applications, and data within a network. Within this zone, the user can access services within the Data Factory through graphical user interfaces (GUIs) or applications. To facilitate management, an administrative backend is usually implemented to enable changes and maintenance tasks. This backend may include tools such as APIs,

command-line interfaces (CLIs), or other administrative interfaces for managing and monitoring the Data Factory services.

Component "Exchange"

In this component, data is exchanged from the source to the data center. This may happen, e.g., through Data Touch Points.

Component "Exchange and Distribution"

In this component, the data exchange between data sources and data centers is handled (Zone 1) as well as the data exchange between all connected data centers (Zone 2), for example to setup a data catalogue or to exchange big amount of sensor data.

Component "User Access"

In this component, user access to a data center is ensured. This means that this component handles the security and network access functions required to let a user of the organization access the data center system.

4 BUILDING BLOCKS

While the functional groups (Fig. 2) provide a neutral representation of an AI system and its functional zones, the pan-European Data Factory as well as the pan-European high-speed network must be aligned with the use cases and requirements defined in D1 [5]. From there, it becomes apparent, that the AI systems data acquisition system is split into a data exchange and an information exchange component as defined in the use-cases CEFT-1 and CEFT-2. Likewise, CEF-5 describes the need for simulation. The following diagram is meant to visualise the resulting building blocks. Furthermore, as the system's aim is to aid in the creation of AI models for fully automated driving, additional steps for certification and approval are required. While these processes resolve mainly around roles and documents, they can be greatly aided by automating the creation of required documents. Therefore, it should also be represented in these high-level building blocks.

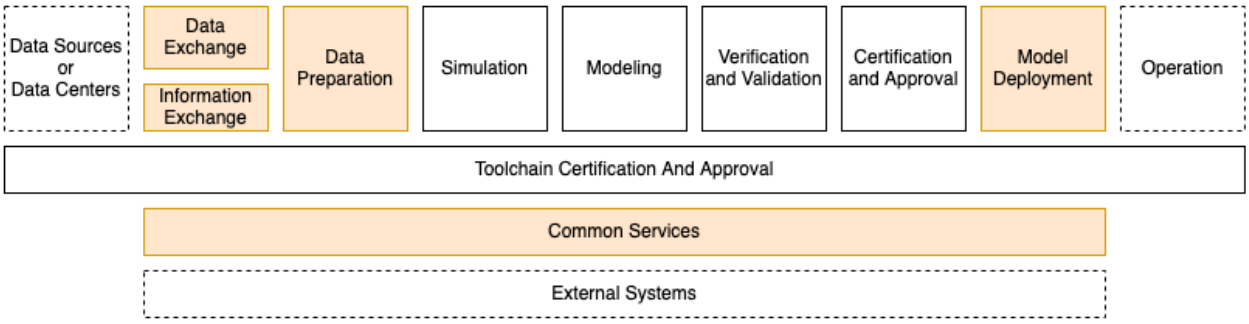

Figure 4: High Level Building blocks as derived from functional zones. Dotted boxes signify components that are part of the overall system but not the data center itself. Orange boxes signify components with required building blocks.

Further analysis of these building blocks is required:

Data Sources

After offloading recorded data from a train, these data have to be transferred to an individual data center. Then these data will be accessible for the pan-European Data Factory.

Data Exchange

The data exchange system is the system enabling the actual data exchange between connected data factories as well as data factories and data touch points. As such it is an important building block for a pan-European Data Factory.

Information Exchange

The purpose of this building block is to exchange 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. As such, it is an important building block for a pan-European Data Factory.

Data Preparation

In the data preparation building block, the key use cases, such as dataset composition (CEF-4) and data searching (CEF-1) can be found. For lack of a better term, this functionality could

be described as being part of the data management functionality. As this functionality needs to interface between data factories, it is an important building block for the pan-European Data Factory.

Modelling

In this building block, data modelling is done, and AI models are trained. It is not directly relevant to the pan-European Data Factory as it has no direct interfaces to other data factories. This however does not mean, that alignment on modelling and training toolchains is not necessary. In order to achieve certifiability, it may make sense to align and find consensus on toolchains to be used in this building block

Simulation

In this building block simulations are done to improve the behavior 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. Verification and validation are essential processes in accordance with EN 50126 (“Railway Applications. The Specification and Demonstration of Reliability, Availability, Maintainability and Safety (RAMS) Generic RAMS Process") for ensuring the safety and reliability of railway systems. 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. It involves conducting tests and simulations to assess the system's behaviour and performance under various operating conditions. By following the verification and validation processes outlined in EN 50126, railway projects can ensure that their systems are thoroughly tested, verified, and validated to meet the required safety standards.

It is not directly relevant to the pan-European Data Factory by itself, but the use cases with a development understand the Data Factory as a tool and thus the requirements of EN 50128 (“Railway applications – Communication, signalling and processing systems”) and EN 50129 apply to the development activities carried out. However, it may make sense to further align on this topic and develop common methods and standards for validation and verification in order to reduce individual efforts at this stage and thus speed up development, save resources and advance an overall European rail system for climate protection.

Certification and Approval

In this building block, the AI system is certified to be used in the railway environment. Depending on the AI systems required safety level, this may involve different procedures and independent entities to do assessments. Once this has been done, the AI system may be used in the railway environment. In order to support this process, required documentation

should be generated automatically, as much as possible. While overall this procedure is defined by EU bodies, its execution lies within the responsibility of national entities.

Model Deployment

In the model deployment building block, models are made available for consumption. As the CEFT-2 requirements state, that models shall be exchangeable between data factories, this is also a required building block for the pan-European Data Factory.

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. Not all of them are required for the pan-European Data Factory. Among others, there are requirements for security, network access and access management in the context of use case CEFT-3. While these are needed in the context of every single Data Factory, the assumption is made that likely any participating organisation already has these services in place. Therefore, it is necessary to define some relevant functions in the realm of common services. These are data import and distribution, user access, but also monitoring which is listed as a requirement in CEFT-3.

Operation

In this building block, the operation of trained models take place. It is not directly relevant to the pan-European Data Factory as it has no direct interfaces to other data factories.

External Systems

This building block provides basic services that are provided by external systems. It is not directly relevant to the pan-European Data Factory as it has no direct interfaces to other data factories.

In practice, though, the building blocks are more fine-granular than described in the section above, as some functionality would, in practice, be split up into multiple subsystems in order to separate concerns as described in Figure 5. This system diagram includes the overall system overview with its components, data flows, building blocks and a functional view.

The system components give an overview over where they are located in the Data Factory. One level down, all currently required high-level building blocks are listed - in orange - and form a grouping within the component of functional zones.

Within the zones, there are in turn various phases that are part of the AI systems lifecycle. For example, the phase of data pipelines, which enable highly automated processing and orchestration, among other things. As well included the AI development process to train and re-train AI models.

A further, deeper gradation releases the functional view of the more detailed building blocks. The functions to be performed are located in these blocks in order to further clean, process, orchestrate, transform, manage, format and prepare the data. It must be noted that this diagram represents the current state of the art for the Data Factory. Therefore, it cannot be considered final as there are still open questions that need to be answered in future research. One open point for example is where to handle data augmentations. While we assume that simple augmentations like resizing of images, changing of colour gradients etc. can be handled by the data curation system, more advanced augmentations like embedding simulated data in images or augmenting rain or snow will likely not be handled by the data curation system.

The black arrows describe how the data is processed through the entire Data Factory and also how a loop has to function during retraining, testing or hyperparameter tuning.

This system diagram also shows how the data flow for annotating data takes place and which functionalities are involved, as well as how a removed model is played back into the train at the end.

Of course, activities must be monitored and recognised at all times via logging and monitoring.

However, the "Data Source Connector" building block must be highlighted for the Connecting European Facilities project, as it is a very important component for the pan-European Rail Data Factory. The Data Source Connector is a crucial component in the field of data integration from other sources and the distribution of data. It acts as a kind of "bridge" or "adapter" that enables data from various sources to be seamlessly merged into a single national Data Factory or a central Data Factory and thus into a common system. The different data sources can originate from diverse systems, technologies, or data formats.
The Data Source Connector must establish a connection to specific data sources, and for this, a necessary API must be defined, or other technologies used so that the different data sources can communicate.
The data should be in a standardised format or the connector must convert it. The Data Source Connector must be capable of converting data into a uniform format so that the data platform can process the data. It should also be capable of meeting the security and privacy requirements of both the data source and the systems that receive and send the data.

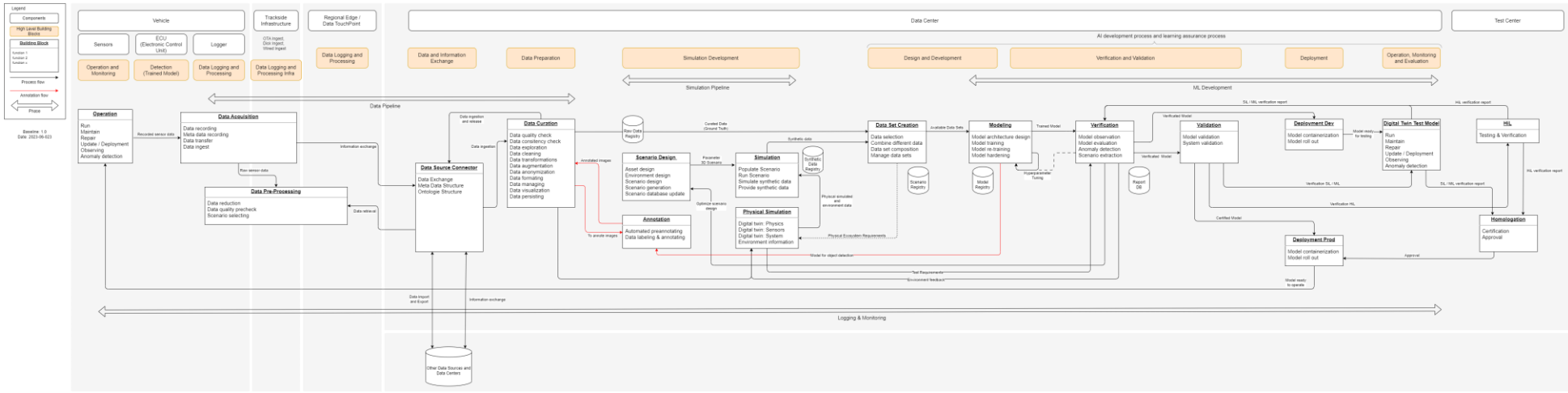

Figure 5: System Diagram.

5 IMPLEMENTATION CONSIDERATIONS AND AVAILABLE SOLUTIONS

Numerous considerations must be taken when implementing the relevant building blocks, which are discussed below.

5.1 DATA EXCHANGE

The data exchange, in particular that between the data factories can be implemented either in a push or a pull model. In the push model, each data center offers a service that a client connects to in order to upload data. In a pull model, the entity serving the data offers a service the client can connect to in order to retrieve the data.

5.1.1 Data Source to Data Center

For the purpose of bringing data from a data source into a data center, it is assumed, that data storage capacity at a data source is limited. Furthermore, it is assumed that data sources are trusted entities. For these reasons, a push-approach should be considered for this transfer as otherwise there is a risk of losing data at the source if a data center does not pull from there in time. Therefore, data sources need write access to a data store in the data center to which they can push data.

5.1.2 Data Center to Data Center

If a pull approach is chosen for data exchange between data centers, the services for providing the data must be located in a network zone separate from the zone for data import. Regardless of the approach chosen, the service handling the data import needs to be able to write data to the backend storage (however in a pull model the ownership of the transfer initiator is the same entity as the owner of the backend storage). It is assumed that the deleting, overwriting or changing existing data on the server side must be prevented to ensure data cannot be modified or manipulated for traceability reasons. An important point to consider is access patterns. Generally, it is assumed that users will request data by linking it in a dataset or even request its creation. This would make a pull based approach quite natural from the perspective of how data is consumed. An additional benefit is, that this would decouple data centers to a larger degree as it would mean a data center cannot influence another one by, e.g., pushing too much data. The access can be limited to read access. A push mechanism would likely require additional security measures.

In the solution space, there is a large choice of existing protocols that can be leveraged for either pushing or pulling data, including among others HTTP(S) or FTP(S). However these protocols do not, by themselves, offer a standardized API for data exchange. The HTTP(S) REST based S3 API which was designed by Amazon for its Simple Storage Service could be leveraged to enable this, in particular for a pull based access. While not an official standard, it has become a de-facto standard for object storage access. It is supported (at least partially) by most large cloud vendors, as well as many solution providers for on-premises storage, including the Ceph Object Storage Gateway [14], NetApp ONTAP or DDN ExaScaler and others. Furthermore, this protocol enables very fine grained control of access as it is based on REST resources. This can, e.g., be leveraged for access protection based on a PKI.

5.2 INFORMATION EXCHANGE

To make data available to other data factories, the public availability of data in a Data Factory must be communicated to other data factories. Like data exchange, this can be realised with different paradigms, either using a client-server pattern with a push or pull strategy, or realised using a queueing system. In addition, each Data Factory could maintain a copy of the grid's entire public data catalogue or serve only as a means to notifying users if each Data Factory maintains only a partial catalogue and searches are disseminated through the grid. Considering the requirement to have data previews available, the latter approach may be more suited to handle such aspects. In addition, this system can be used to publicise the availability of services at affiliated facilities.

5.3 DATA PREPARATION

The central data management of a Data Factory takes place in the data preparation zone. The process of data management is one of the central user tasks within a Data Factory. The system or systems handling this can be understood as the brain of any Data Factory, if not the entire interconnected system. Given the structured or at least semi-structured nature of metadata like labels (annotations), dataset or ontology data, as well other metadata like references to sensor data, usage data, access restrictions, etc. this can be implemented through a CRUD web application backed by a relational, or non-relational (e.g., JSON based) datastore. Of importance are the ability to efficiently search the data catalogue for data of interest across all connected facilities and to put notifications of new data, metadata or changes onto the bus system. For the data to make sense to the users of the pan-European Data Factory, a common ontology and common data formats for labels, datasets, simulation assets etc. are of utmost importance, otherwise data can be exchanged but not easily interpreted.

5.4 MODEL DEPLOYMENT

In order to exchange models between data factories, data exchange mechanisms for models, similar to the information exchange are necessary.

For the exchange of model data, artifact repositories, accessible through REST APIs (Representational State Transfer Application Programming Interface) have become a de-facto standard solution. In addition, however the use of standardized model formats also plays a major role in ensuring that the recipient has a standardized way of using the model. For this purpose, competing standard formats exist. Notably are in particular the NNEF [15] standard published by the Khronos Group, an open specification for network formats, as well as the ONNX toolsuite [16] which is governed by the Linux Foundation. With the latter being backed by some of the largest players in the IT industry, like Microsoft, intel and nVidia, as well as players in the railway market, like Siemens, it may, at this time, be the most relevant tool to do further research on in order to establish the feasibility of its usage in our domain. Models can be converted between the NNEF and the ONNX formats, however it is unknown at this time, if convertibility of models can be assumed for any given format.

5.5 DATA IMPORT

For the import of data from touch points and other data sources, fixed line connections using MPLS leased lines or IPSec tunnels can be considered. Which solution is better suited depends on the size of the data source and implementation cost.

5.6 DATA DISTRIBUTION

Connectivity between data centers requires a high bandwidth. To make this possible and meet the high security requirements, this connectivity should be realised via a fixed line connection that can be implemented using MPLS leased lines from a European network provider. Potentially, dark fibers could also be used for this, however this requires a high degree of coordination between the participating data center operators. A common approach is recommended.

5.7 USER ACCESS

For user access, on the networking level it must be decided, if, for security reasons, access can only be granted from known networks (e.g., fixed connections or IPSec tunnels) or if, based on a growing popularity of modern work patterns, access should also be available via the Internet. Additionally, in order to facilitate user access, additional services are required to facilitate access management. In a grid system this is typically implemented using a federated IAM system.

In the realm of IAM systems, many solutions are readily available today. Multiple protocols come into question for authentication and authorization between data factories and data sources. X.509 client certificates, OIDC or SAML would be viable solutions to solve both machine-to-machine authentication as well as user-to-machine authentication, and could theoretically even be federated, which is desirable in a multi tenancy use-case. While it is important to note that any one of these protocols may be particularly well suited for a particular use-case, it is also common for modern IAM systems to support multiple, if not all, of these protocols. The cloud-based Azure Active Directory (AAD) for example offers support for all of these protocols. Similarly, this can be achieved with open source solutions such as FreeIPA and KeyCloak. However, it is generally assumed that every participant in the pan-European Data Factory has such a system in place already. This topic is discussed in more detail in D 2.2.

5.7.1 Monitoring

The monitoring system’s main responsibilities are the collection of metrices and log data from a Data Factory. This is not only important to ensure the secure and uninterrupted operation of a Data Factory, but also becomes very important in this context as it is a source of data related to the business case of the partners in the connected facility. Implementing a robust monitoring system is crucial for maintaining data quality, with a specific focus on data integrity. Detecting potential data anomalies in a timely manner and responding appropriately requires the implementation of a proactive monitoring system within the network and among partners. Such a system can incorporate advanced data analysis techniques and automated monitoring mechanisms to identify any deviations or irregularities in the data. By promptly identifying anomalies, swift action can be taken

to rectify the situation and restore data integrity to its intended state. This proactive approach ensures the reliability and accuracy of the data, bolstering overall data quality management.

Therefore, common APIs are also important for querying monitoring data and thus enabling different views of the monitoring data in different contexts. For example, the users of the system need to easily track usage or cost across connected facilities, whereby the data security role needs to include the anomaly detection into the system operation centre. A common API could, e.g., be leveraged by employing industry standard, open source monitoring data stores, such as Prometheus.

5.7.2 General Considerations

In order to establish a pan-European Data Factory, interoperability on many levels is key. This not only includes APIs of connected systems as described in this chapter, but, as already implied, also ontologies, in order to have a common understanding of data and metadata and formats for any kind that shall be exchanged. This includes not only data and metadata but also asset data for simulations, e.g. scene descriptions, possibly even sensor models and other data. This topic will be further elaborated on in WP 2.3.

6 ORCHESTRATION AND OPERATION CONSIDERATIONS

Operation and orchestration aspects of pan-European Data Factory incorporate several key topics, such as:

Data Gravity

Given the gravitational pull of data, the compute environment is designed for data-intensive workloads. It prioritises proximity to data sources to minimize 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 organizations to utilise the flexibility and scalability of the cloud while maintaining control over sensitive data. This can be the case for an individual Data Factory, as well as the entire system and should be transparent for the participants.

Workload Scheduling

An intelligent workload scheduler may be employed to distribute workloads across available compute environments, possibly situated in different Data Centers and different Data Factories efficiently. It considers factors like resource availability, data locality, and workload priorities to optimize resource utilization and minimize training time. This requires APIs to be present in data factories so that tenants can schedule workloads, automatically. These APIs must not necessarily be the same as typical market available workload schedulers are quite flexible in this regard, however a minimum set of requirements should be defined.

Multi-Tenancy

The system supports multi-tenancy, allowing multiple users or organizations 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 utilized. These appliances consist of pre-configured software stacks, libraries, and frameworks optimized, e.g., for deep learning and AI workloads or taking over management and IT operations tasks in a Data Center backed by cloud services.

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 Data Factory tenants. In order to enable participants to use cloud resources, without discrimination, it must be possible to interact with data center cloud systems or even individual tenants, so that participants can freely choose between them.

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. Whether dedicated connections, such as direct links or virtual private networks (VPNs) are required to satisfy the bandwidth needs and to ensure reliable and secure data transfer is a case-by-case decision that must be made in the context of specific Data Centers. The secure data transfer is discussed in Deliverable 2.3 Chapter 4.

Multi-Site

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 minimizing the impact of potential outages. In the context of the pan-European Data Factory, this must be transparent for tenants or other parties requesting access to data.

Maintenance

The pan-European Data Factory needs to incorporate mechanisms to announce maintenance in connected systems in order to ensure smooth operations. Each Data Factory must facilitate patch management, software upgrades, and hardware maintenance without disrupting ongoing workloads. As the system would allow the setup of multitenancy, as well as loosely coupled, federated instances, it is up to a data centers operator to schedule and negotiate maintenance windows mostly with its own users as impact to other data centers will be limited in a federated approach.

Environmental Considerations

The data center operators need to take into account environmental considerations, such as power efficiency and cooling requirements. The computational requirements for AI training and high-performance storage inside of a data center, as well as high bandwidth requirements come with an ever increasing demand for power. Therefore, optimizing power usage effectiveness (PUE) through energy-efficient hardware selection and intelligent power management techniques should be considered by every participant in the pan-European Data Factory who opts to operate their own infrastructure.

7 CONCLUSION AND OUTLOOK

In the first deliverables D 1.1, D 1.2 and D 1.3 [5] of the CEF2 RailDataFactory study, the vision and concept of a pan-European Data Factory has been introduced, including the definition of terminology and of roles, and the identification of use cases and their related requirements.

In this deliverable, an architectural foundation has been laid out that needs to be carefully extended in the future. This includes high level concepts and building blocks for the pan-European Data Factory as well as a System Diagram containing a detailed view on the components, building blocks and processes. Based on this, considerations are presented that need to be evaluated in depth in developing the concept further in future work.

Deliverable 2.2 will provide concepts for the IAM (Identity and Access Management) system, which is as well is a key part for a pan-European Rail Data Factory as well as concepts for data management and data security. Not only will it take care of the topic of data governance and how IAM can work in a federated mode, but also the security aspect is dealt with there, including what needs to be considered when transferring data. In addition to presenting these concepts it is defined how they will fit into the complete ecosystem. Following this, Deliverable 2.3 will provide concepts and implementation proposals for the pan-European Rail Data Factory backbone network.

In future studies or a follow up project, concrete technological implementations need to be evaluated, and more detailed concepts for governance and data protection need to be developed.

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, May 2023. [Online]. Available: https://digitale-schiene-deutschland.de/Downloads/2023-04-24_RailDataFactory_CEFII_Deliverable1_published.pdf

[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/news/en/Data-Factory

[11] COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS "A European Strategy for Data", 19.02.2020, https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1593073685620&uri=CELEX%3A52020DC0066

[12] ISO/IEC 23053:2022 "Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)", 06.2022, https://www.iso.org/standard/74438.html

[13] RESEARCH REPORT SIL4 CLOUD, 14.09.2022, https://digitale-schiene-deutschland.de/Downloads/Report%20-%20SIL4%20Cloud.pdf

[14] Ceph Object Gateway S3 API, see https://docs.ceph.com/en/latest/radosgw/s3/

[15] NNEF Specification, see https://www.khronos.org/nnef

[16] ONNX Project, see https://github.com/onnx/onnx