CEF2 RailDataFactory

Deliverable 3.1 – Report of bottlenecks data application in rolling stock

Due date of deliverable: 30/09/2023

Actual submission date: 15/11/2023

Leader/Responsible of this Deliverable: Bart du Chatinier (WP 3 lead) / NS

Reviewed: **Y**/N

Document status		
Revision	Date	Description
01	09/03/2023	Document template generated
02	20/07/2023	Content transferred from Confluence
03	28/07/2023	First draft complete
04	28/10/2023	Version submitted to advisory board
05	15/11/2023	Updated version after addressing all advisory board comments
06	15/11/2023	Version submitted to 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: 9 months
(note: amendment request for project extension ongoing)

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
Bart du Chatinier	NS
Philipp Neumaier	DB
Julian Wissmann	DB
Wolfgang Albert	DB
Martin Jungklaus	DB
Philippe David	SNCF
Patrick Marsch (only editorial effort)	DB

Note of Thanks

The authors would like to thank the Advisory Board members for their valuable input to the project!

Disclaimer

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them.

Furthermore, the information in this document is provided “as is”, and no guarantee or warranty is given that the information is fit for any particular purpose. The author(s) and project consortium do not take any responsibility for any use of the information contained in this deliverable. The users use the information at their sole risk and liability.

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 RailDataFactory 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 tools and toolchains, approve 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 RailDataFactory.

In this deliverable of the study, the current bottlenecks of transferring & mutating data within a rail network are studied. The report describes the challenges in data connectivity that are currently present while experimenting and deploying (AI) models within the rail industry. A gap between existing rolling stock and future technological advancements is described and proposals are made how the RailDataFactory can be supplied with a constant flow of data by participating in a pan-European ecosystem.

ABBREVIATIONS AND ACRONYMS

Abbreviation	Definition
AI	Artificial Intelligence
ATO	Automatic Train Operation
ATP	Automatic Train Protection
CS-TSI	Communications, signaling and processing systems within TSI
DAS	Driver Assistance System
ERA	European Union Agency for Railways
ERTMS	Europan Rail Traffic Management System
ETCS	European Train Control System
FRMCS	Future Railway Mobile Communication System
GPS	Global Positioning System
GSM-R	Global System for Mobile Communications - Railways
IM	Infrastructure Manager
LORA	Long Range (communication technology)
OEM	Original Equipment Manufacturer
RU	Railway Undertaking
TSI	Technical Specification for Interoperability
TCMS	Train Control Management System
UIC	International Union of Railways

TABLE OF CONTENTS

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 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 RailDataFactory for the joint development of fully automated driving. The study, being co-funded through HADEA, aims to assess the feasibility of a pan-European RailDataFactory from technical, economical, legal, regulatory and operational perspectives, and determine key aspects that are required to make a pan-European RailDataFactory 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 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 RailDataFactory 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 utilised 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 RailDataFactory. 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 RailDataFactory 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 in 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 aim to reflect the European picture. Within the sector initiative "Digitale Schiene Deutschland", Deutsche Bahn already started to set up some components of the Data Factory in Germany [10].

1.3 AIM AND STRUCTURE OF THIS DELIVERABLE

This current document is the deliverable D3.1 of the CEF 2 RailDataFactory project, covering an analysis of the current challenges of transferring & mutating data in networks of rolling stock. Some bottlenecks for feeding operational data into the RailDataFactory are identified and these findings will be related to existing rolling stock requirements/standards in Europe.

The aim of the document is to obtain early feedback and possible additions from the sector on data acquisition within the railway environment, in order to update the work accordingly and consider the obtained input in the subsequent phases of the project.

The remainder of this document is structured as follows:

- In Chapter 2, stages for maturity in data transfer for European rolling stock are provided;
- In Chapter 3, a short summary of relevant European standards is provided;
- In Chapter 4, existing technologies for train-trackside communication are outlined;
- In Chapter 5, the Data Touch Point, the edge solution developed by DB for transferring large amounts of sensor data from trains to data centers is described;
- In Chapter 6, challenges to the current design for rail connectivity are discussed;
- In Chapter 7, a summary is provided which ends with a conclusion on future train-track interaction for future operational perspectives in train automation.

2 STAGES IN DATA HANDLING ON ROLLING STOCK IN EUROPE

One of the key ingredients for the RailDataFactory is a consistent, quality flow of up-to-date data from European trains and rail infrastructure to the back-end system. Trains, also known as rolling stock, are known to have a wide variety of maturity levels. Every country in Europe has rolling stock with different ages and maturity in technology due to its long lifespan (up to 30-40 years). Railway undertakings (RU) typically replace rolling stock in cycles, using tenders, which are often awarded to one of the European Original Equipment Manufacturers (OEMs). This leads to a mix of new and old rolling stock in operation on the network at the same time. Some rolling stock from RU are very modern and have sophisticated train-to-shore technologies that enable constant data communication. Other rolling stock is older, sometimes manufactured in the 90s, and still in use - despite the absence of train-to-shore technology.

Some types of rolling stock are upgraded with modern technology (also known as refurbishment) to improve performance and extend their lifespan, this is often executed by the OEM. For infrastructure managers (IM), roughly the same pattern can be described. Some parts of European infrastructure are recently built/renovated and are equipped with sensors, the newest train protection and good mobile network coverage. Other parts of Europe are in need of upgrades and are unable to meet the modern standards that new trains require.

This means that in the railway environment, technologies from different decades coexist, all within the same rail network that is often cross-border connected. The European network has a diverse rail network that includes high-speed lines, regional lines, and urban transit systems. Each of these networks have different requirements for rolling stock and infrastructure, which can lead to the use of different technology levels. Each European country can have its own local regulations for rolling stock, beside the European standards, which can lead to different technology being used in different countries.

The heterogeneous technology of rolling stock will be an important challenge while implementing a pan-European ecosystem. Ideally, the RailDataFactory is filled with a wide variety of operational data - whether it's the older locomotives in regional lines or the newest high-speed trains on international high-speed lines. It is expected that the contribution of different types of rolling stock is dependent on their maturity level. Modern, sophisticated trains can provide daily data on a wide range of subsystems - while older trains might need other ways to feed their data into the RailDataFactory. The following stages can be used to identify the maturity levels of how data from trains is handled in European rail:

- **Stage 1 - Offline data retrieval**: Data generated by rolling stock on-board systems can be retrieved manually by technicians during routine maintenance checks. A process can be introduced to upload this data in the RailDataFactory. This process can be time-consuming, prone to errors, and does not provide real-time visibility into the performance of rolling stock.

- **Stage 2 - Onboard data storage**: To address the limitations of manual data retrieval, on-board data storage systems were developed. These systems allow rolling stock to store data on-board and transmit it to a central database when the train returned to the depot. While this is an improvement over manual data retrieval, it still did not provide real-time

visibility into rolling stock performance.

- **Stage 3 - Train-to-trackside**: For train-to-trackside communication several systems were developed to allow rolling stock to communicate with the trackside in real-time or not-real-time. This allows the control center to monitor rolling stock performance and make adjustments to the schedule to optimise performance. In addition, the communication systems allow for remote monitoring systems to provide real-time visibility into the performance of rolling stock. These systems use infrastructure sensors and telemetry to transmit data from on-board systems to a central database in real-time. The data enables maintenance teams to monitor rolling stock performance in real-time and take proactive measures to prevent failures. In addition, the Data Touch Point (see Chapter 5) uses non-real-time communication for transferring sensor data from the train to the trackside and to the data center.
  The technology described above could allow rolling stock to feed shore-based databases in real-time. Data can be from either infrastructure interactions, operational procedures on the train or sensor data from subsystems. This enables RU to optimize performance, reduce downtime, and improve passenger experience. The availability of such a wide range of data also allow capabilities such as condition based & predictive maintenance. It becomes possible to use data analytics and machine learning algorithms to predict when rolling stock components are likely to fail. As a result, maintenance teams can take proactive measures to prevent failures and reduce downtime. Currently, RUs in Europe are exploring the possibilities of using this technology to introduce driver assistance and automatic train operation.

It should be noted that the key points and rough recommendations as documented in RailDataFactory Deliverable 2.3 [11] are conditional for a backbone network that enables the products and processes described above.

3 EUROPEAN STANDARDS FOR DATA COMMUNICATION IN RAIL

In Europe, the use of standardised communication systems in rolling stock are mandated by the Technical Specifications for Interoperability (TSI) issued by the European Union Agency for Railways (ERA). These TSIs establish the technical requirements for train-wayside communication systems in order to ensure interoperability and safety across the European rail network. The TSI for the subsystem "Communications, signalling and processing systems" (CCS-TSI [12]) lays out the requirements for data communication systems in rolling stock, including the use of standardised communication protocols and interfaces.
One of the main standards for train-trackside communication in Europe is the European Train Control System (ETCS [13]). ETCS is a train control system that uses GSM-R (Global System for Mobile Communications – Railways [14]) as the communication infrastructure. It enables the communication of train location, speed, and other operational data between trains and wayside systems, as well as providing Automatic Train Protection (ATP) and possibly Automatic Train Operation (ATO) capabilities. In addition, DB has developed the bbIP (backbone for railway IP communication) as a potential standard for trackside communications for train related data. It provides high-speed data transfer, low latency, and high reliability for applications such as ETCS [15][16].

Another standard in development is the Future Railway Mobile Communication System (FRMCS) [17]. This project aims to develop a standardised wireless communication system for the rail industry, replacing the existing GSM-R technology. FRMCS is being developed by the International Union of Railways (UIC) in collaboration with various stakeholders, including railway operators, manufacturers, and standardization organisations. At the moment of writing, UIC is actively working on defining the technical specifications and requirements for the new system. ERA is actively involved in the FRMCS project, working closely with the UIC and other stakeholders to ensure the new system aligns with the EU's regulatory framework and interoperability requirements.

While these new standards are in development, currently the rail industry faces challenges in implementing proper data communication to transfer data from trains to the trackside infrastructure. Multiple communication protocols are being used but there is no European standard (yet) that has been widely adopted across the entire rail industry. This lack of standardisation has several reasons:

- **Legacy systems**: Many rolling stock types and rail networks have been developed over decades, leading to a diverse range of legacy systems. These systems were built using different technologies and communication protocols, making it difficult to establish a single standardized solution that can integrate with all types of rolling stock operated on the rail infrastructure;

- **Vendor-specific solutions**: Train manufacturers and technology providers often develop proprietary communication protocols for their equipment and systems. As a result, when different vendors' products are used on the same rail network, interoperability issues arise, as these proprietary protocols may not be compatible with each other;

- **Safety and reliability concerns**: The rail industry prioritises safety and reliability, and implementing a new communication protocol on a large scale requires extensive testing and validation to ensure it meets these critical requirements. Developing and adopting a new standard must undergo rigorous scrutiny to minimise the risk of accidents or system

failures;

- **Cost and Investment**: Replacing or upgrading existing communication technology on rolling stock and a complex rail network can be expensive and time-consuming. This cost factor often acts as a deterrent for rail authorities from committing to a specific standard without a clear consensus across the industry.

As described above, various organisations and standardisation bodies are working on developing common protocols. In the meanwhile, the rail industry might rely on intermediate solutions - such as gateway devices that translate data between different protocols to achieve some level of interoperability and data exchange between trains and trackside infrastructure. However, until a comprehensive and widely accepted standard data communication protocol is established, the rail industry will continue to face hurdles in fully harnessing the potential of data-driven technologies for improved efficiency, safety, and passenger experience. In the next chapters, it will be advocated how a pan-European RailDataFactory could help solving this dilemma.

4 EXISTING TECHNOLOGIES THAT ENABLE TRAIN-TRACKSIDE COMMUNICATION

Train-trackside communication has become a crucial aspect of modern rail transportation over time. It started by enabling real time travel information, passenger Wi-Fi and basic services such as geolocation. Nowadays, this same connective technology is applied to enable the exchange of information between trains and trackside systems to ensure the safe and efficient operation of the rail network. The modern train has become part of a sophisticated system where both rolling stock and trackside technologies enable operational performance on the European network. Examples include:

- Sensors on rolling stock are devices that are installed on train subsystems to collect data on various aspects of the train's operation. These sensors can include counters, measurements and events. They are used to monitor the functionality of various subsystems such as doors, traction, and the Train Control and Monitoring System (TCMS). Data feeds from these sensors are used to provide real-time information on the train's performance and condition, which can be used to analyse the functionality of various subsystems. For example, data from door sensors can be used to monitor the operation of the train's doors and identify any issues or malfunctions.
  Data from these sensors can also be used to perform predictive maintenance, which can help to improve the reliability and availability of the train. This can include identifying potential issues with subsystems before they become a problem in operation, and scheduling maintenance and repairs at the most convenient time. By monitoring the train's energy consumption in real-time, it is possible to identify areas where energy consumption can be reduced and make adjustments to the train's operation to minimize energy waste.

- Live positioning services enable train control centers to have real-time information about the location and status of trains on the track. These services use a combination of technologies such as GPS, balises, and other track-side sensors to provide accurate and up-to-date information on the position and speed of trains. By using live positioning services, train control centers can improve the efficiency and safety of train operations. For example, they can use this information to optimise train schedules and routes, ensuring that trains are running on time and avoiding delays. It also allows to have better visibility of the train movements, which in turn can help to make better decisions in case of delays, stranding or incidents.

- Live positioning services can also be used to improve the reliability of train operations by providing real-time information on the condition of the track and infrastructure, identifying potential issues before they become a problem. Algorithms can be created to continuously find deviations in the infrastructure, such as defects on overhead wire or track geometry. Triggered alarms can be monitored for possible development toward higher levels, such as repair or necessary engineering work.

- Driver Assistance Systems (DAS) are designed to support train drivers in making safe and efficient decisions during the operation of a train. These systems use a combination of sensor data, communications, and onboard processing to provide drivers with real-time information and guidance. One of the main functions of DAS is to provide train drivers with information on the current speed and position of the train, as well as the location and status

of other trains on the track. This can include information on train speed limits, signals, and other track-side infrastructure, as well as data from onboard sensors such as cameras and radar.
DAS can also be used to provide drivers with guidance on the best route to take, taking into account factors such as traffic conditions and track conditions. This can include information on train schedules, traffic density, and the location of other trains on the track. Additionally, DAS can be used to support the driver in making decisions related to train operation, such as braking, acceleration, and train control. These systems can use data from onboard sensors, such as cameras and radar, to detect obstacles on the track and provide the driver with guidance on how to safely and efficiently avoid them. DAS is frequently used to improve the efficiency of train operations by providing real-time information on the train's energy consumption, and providing recommendations on how to reduce consumption.

All these examples are applicable for the potential of the RailDataFactory, as with the availability of data from rolling stock all over Europe it will be easier to develop use cases and models. The potential cases are endless, for example a RU can prepare for the introduction of new rolling stock by using data from other countries to develop maintenance models. Or object detection models can be trained by using data from another country, for example to become better in rare weather conditions for that country.

To get data from the train to the shore, there are several existing technologies that enable train-trackside communication, an extensive description can be found in RailDataFactory Deliverable 2.3 [11] in Section 2.2. In a nutshell, some technologies include:

- **GSM-R**: This is a specialised version of the GSM mobile communication standard that is specifically designed for use in the rail industry [14]. It enables the communication of train location, speed, and other operational data between trains and trackside systems. GSM-R is used by almost all countries in Europe for communication with control center;

- **3G/4G/5G**: The use of cellular networks, such as 3G, 4G, and 5G, for train-trackside communications is becoming increasingly common. These networks provide high-speed data transfer, wide area coverage, and support for a range of applications. This technology is often used in European rail for customer Wi-Fi / travel information, etc.;

- **5G campus networks**: A 5G campus network is an exclusive mobile network for a defined local campus. The dedication of the network adds additional security and control features compared to a 5G cellular network and the local campus may span a range of up to a few square kilometers. Next to the security aspects, a campus network offers fast data transfer with low latencies and a high reliability with low energy consumption. Compared to a Wi-Fi 6 connection, the 5G Campus network offers connections with lower latencies, theoretically higher bandwidths and the opportunity to scale to 5G public networks;

- **Wi-Fi 6**: Wi-Fi 6 is the most recent Wi-Fi network protocol. It raises the connection speed, lowers the latency and offers an improved connection quality compared to the Wi-Fi 5 standard. Compared to a 5G Campus network, the proven Wi-Fi technology offers an easy to implement and cheaper solution for network connections but has limitations in regards of connectivity range;

- **Balises**: Balises are transponders placed on the track that communicate with on-board train systems via radio waves. They provide information such as train location, speed and signaling information to the on-board train systems which are further passed on to the trackside systems;

- **LORA communication technology**: This is a Ultra Wide Band communication protocol used in European Rail. It is a low flow protocol which makes it suitable for small data applications such as tracking and monitoring assets in rail systems. Due to the low-power/long-range design, a relative small amount of antennas is required for good coverage.

5 TRACKSIDE EDGE COMPUTING: THE DATA TOUCH POINT

Transferring data from the rolling stock to the data center is one of the big challenges today since massive amount of collected data from the various sensors on the rolling stock overloads the infrastructure that is currently available - Also, the compute power on the trains is not sufficient, and the network connections to the data centers has not the required capability in terms of bandwidth.

To address these needs, an edge solution at the trackside called Data Touch Point is currently being developed. **Fehler! Verweisquelle konnte nicht gefunden werden.Fehler! Verweisquelle konnte nicht gefunden werden.** shows the Data Touch Point (marked in yellow) in the context of the Pan-European Data Factory (see also RailDataFactory Deliverable 1 [5]). The Data Touch Point consists mainly of 3 systems: (1) Communication System, (2) Storage System and (3) Edge Compute System, which are elaborated in the following:

1) The **Communication System** ensures the transfer of the data that was collected and recorded on the rolling stock via suitable technologies like 5G campus networks or Wi-Fi 6 to the Touchpoint (considered an edge location). In delimitation to chapter 2 (Stage 3), the Data Touch Point does not use real-time communication, but fetches the high amount of collected sensor data, while the train stands still in the depot. The Communication System also ensures the data transfer from the Touch Points Storage System to the data center;

2) The **Storage System** in the Touchpoint ensures the caching of the data;

3) An **Edge Compute System** performs data preprocessing, to identify the significant portion and reduce the amount by deleting unimportant parts (see “Data reduction” in **Fehler! Verweisquelle konnte nicht gefunden werden.**).
   The preprocessing of the data consists of the following tasks:
   - Perform sensor specific post processing steps.
   - Perform quality checks to identify unusable data which can be deleted:
     - Check integrity to identify empty or corrupted files
     - Check synchronicity across all sensors
     - Check validity of sensor data
   - Identify suitable data on the basis of metadata and parameters like weather conditions and geo locations to improve data diversity;
   - Run lean AI detectors to identify meaningful objects in the sensor data, enrich the metadata, identify useful data and delete useless parts.
   - Compress data;

Through these steps, the amount of data is reduced significantly, which in return relaxes the bandwidth requirements on the connection to the data centers of the Data Factory.

6 CHALLENGES IN DATA COMMUNICATION IN RAIL

Each of the technologies described have their own strengths and weaknesses, and the choice of which technology to use will depend on the specific requirements of the rail network and the applications that need to be supported. Some of the most significant challenges include:

- **Network coverage**: cellular networks may not be available in all areas where trains operate, particularly in rural and remote areas. This can make it difficult to establish and maintain a reliable connection between trains and shore systems. In addition, since trains are moving in the landscape it has difficulties maintaining a connection with a specific cellular transmission;

- **Bandwidth limitations**: The amount of data that can be transferred over a cellular connection is limited by the available bandwidth. This can be a particular issue for trains that generate large amounts of data, such as trains equipped with sensors and cameras that stream to the trackside;

- **Latency**: The delay between sending and receiving data, also known as latency, can be an issue for train to shore communications. This can be particularly problematic for applications that require low latency, such as train control and command systems;

- **Interference**: Other cellular users in the same area can cause interference with train to shore communications, potentially leading to dropped connections or degraded performance;

- **Security**: Cellular networks are vulnerable to a range of security threats, including attacks on the network infrastructure and data interception. This makes it important to implement robust security measures, such as encryption and secure authentication protocols, to protect sensitive data during transit. Additionally, it is important to have a comprehensive incident response plan in place to quickly detect and respond to security incidents;

To address these challenges, another technology design can be considered. As described in RailDataFactory Deliverable 2.3 [11], in Section 2.3, a new approach would be to set up rail Internet exchange points built by railway infrastructure managers. By using fiber technology, such interconnection points (Rail-IX) would allow connectivity with public networks. This form of edge computing could resolve some of the challenges above.

7 CONCLUSION AND OUTLOOK

Train automation is becoming increasingly relevant for the rail industry given the increasing adaptation of ERTMS and the rapid development of AI. Many RUs and OEMs are investing the development of ATO. Most of the models enabling ATO use AI to revolutionise rail operations, optimising efficiency, safety, and passenger experience. The RailDataFactory could open a new, equal level playing field for data integration, quality, and consistency of rail data. However, the availability of advanced train-track interaction also requires the ability for the train to communicate with trackside systems in real-time. This includes the ability to send and receive data on train position, speed, and other operational parameters, as well as receiving commands and instructions from the trackside systems. As the rail industry continues to evolve and the use of automated trains becomes more widespread, it will be necessary to ensure that these trains can interact seamlessly with other systems and technologies by using European data communication standards. The existence and availability of an advanced backbone is necessary for the safe and effective operation of automated trains in the future.

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. Edge technology could play an important role in improving the performance and efficiency of automated train systems. One of the main benefits is the ability to process large amounts of data generated by trains in real-time. This includes data from sensors, cameras, and other onboard equipment, as well as data from the track-side infrastructure. By processing this data at the edge, it is possible to reduce the latency in the system and make faster decisions. This has also been described in RailDataFactory Deliverable 2.1 [18], Section 3.1. For example, with functionalities like "data preparation" and "data exchange" in edge compute assets.

Edge computing can also help to improve the reliability and availability of train systems. It would become possible to reduce the dependency on centralised systems and reduce the risk of system failures. Another advantage of edge computing in the rail industry is the ability to implement advanced analytics and machine learning algorithms to improve the performance and efficiency of automated train systems. In Deliverable 3.2, it will be described how edge computing with Data TouchPoints (see Data entry points as documented in RailDataFactory Deliverables 2.1 [18] and 2.2 [19]) could make the data transfer to the RailDataFactory and the platform itself an economically beneficial product for European consortium members.

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] CEF-2 RailDataFactory D2.3 – High-speed pan-European Railway Data Factory Backbone Network, see https://digitale-schiene-deutschland.de/en/news/Pan-European-Railway-Data-Factory

[12] CCS-TSI, ERA, 2016, see https://www.era.europa.eu/domains/technical-specifications-interoperability/control-command-and-signalling-tsi_en

[13] ERTMS/ETCS - Set of specifications 3, ERTMS user groups, see https://www.era.europa.eu/era-folder/set-specifications-3-etcs-b3-r2-gsm-r-b1

[14] GSM-R - Functional Requirements Specification/System Requirements Specification, UIC, see https://uic.org/rail-system/gsm-r/

[15] bbIP - bahnbetriebliches IP-Netz, Signon/DB Netz AG, see https://signon-group.com/referenzen/details/db-netz-ag-bahnbetriebliches-ip-netz-bbip-im-vorserienprojekt-dstw-donauworth-meitingen-mertingen-zeitraum-2019-2021

[16] E. Seidler, B. Reichert and C. Kittler, „Das bahnbetriebliche IP-Netz als Schlüssel für die Digitalisierung der Schiene", SIGNAL+DRAHT 12/2021, see https://bit.ly/3OIKcjc

[17] FRMCS, Use Requirements/Use Cases, UIC. https://uic.org/rail-system/frmcs/

[18] CEF-2 RailDataFactory D2.1 – Technical specifications and available solutions for building blocks, components, Cloud / hybrid- Cloud and Edge- Orchestration & Operational concept. https://digitale-schiene-deutschland.de/en/news/Pan-European-Railway-Data-Factory

[19] CEF-2 RailDataFactory D2.2 - Technical specifications and available solutions for Identity Access Management (IAM), Data Management and Transfer and Cyber-Security, see https://digitale-schiene-deutschland.de/en/news/Pan-European-Railway-Data-Factory