Japan Bilingual Publishing Co.

Transportation Development Research

https://ojs.bilpub.com/index.php/tdr

RESEARCH ARTICLE

European Common Data Management Platform Definition for Railway AI Function Development

Mikel Labayen[1,2*] **Daniel Ocho de Eribe**[1] **Ander Aramburu**[3] **Marcos Nieto**[4] **Naiara Aginako**[2]

1. Autonomous Vehicle Department, CAF Signalling, Donostia, 20018, Spain
2. Computer Sciences and Artificial Intelligence Department, University of the Basque Country, Donostia, 20018, Spain
3. R&D Department, CAF, Beasain, 20200, Spain
4. Connected & Cooperative Automated Systems Department, Vicomtech Research Centre, Donostia, 20009, Spain

ARTICLE INFO

Article history

Received: 2 January 2024

Revised: 20 January 2024

Accepted: 26 January 2024

Published Online: 12 March 2024

Keywords:

Common data management platform

Artificial intelligence

AI training and testing

Autonomous vehicle

ABSTRACT

Digitalisation and automation of operations in the railway industry include the use of Automatic Train Operation systems that provide automated functions to reach different levels of automation, known as the Grade of Automation (GoA) levels. These levels go up to GoA4 in which the train is automatically controlled without any staff on board. Artificial intelligence has emerged as technology that can substitute humans in certain driving tasks, in GoA3 (driverless) and GoA4 (unattended) modes. AI capabilities include perception, decision-making, precise positioning, or optimization of communications. The success of AI models depends on the quality and diversity of the data used for training, along with the set-up of a data life-cycle framework that covers creation, training, testing, deployment and monitorisation. The management of training datasets implies both expensive and time-consuming data gathering, labelling, curation and formatting efforts, potentially hindering the development of reliable AI systems. This paper presents a Common Data Management Platform developed by a consortium of European railway stakeholders, devised to efficiently manage data for AI training, and which is demonstrated in two different Proofs of Concept.

1. Introduction

Nowadays, autonomous driving functions in railway operations are based on developments in the automotive sector. For example, Advanced Driver Assistance Systems (ADAS), which come directly from the automotive sector, have been implemented in several demonstrators and showcases in the railway domain. These solutions have something in common: they are all based on artificial sensing.

Artificial sensing permits gathering information from

*Corresponding Author:
Mikel Labayen,
Autonomous Vehicle Department, CAF Signalling, Donostia, 20018, Spain; Computer Sciences and Artificial Intelligence Department, University of the Basque Country, Donostia, 20018, Spain;
Email: mlabayen@cafsignalling.com

DOI: https://doi.org/10.55121/tdr.v2i1.143

the environment, which becomes a key factor when talking about enabling autonomous operation for railway transport optimization. Furthermore, autonomous driving requires the implementation of new on-board functions that complement Automatic Train Protection (ATP).

Some of these functions rely on the perception of both indoor and outdoor environments. Sensing the outdoor environment offers driving clearance (no obstructions, signal status on green...), speed supervision (signals for speed limits), or vehicle localization. Moreover, sensing the indoor environment will be primarily necessary in GoA4, where automated event detection and quicker reaction times will increase operational safety, on-board security, and overall service quality (maintenance).

In railway scenarios, the perception layer based on Computer vision (CV) and Artificial Intelligence (AI) technologies including sensor fusion provides the required environment understanding. CV and AI technologies are essential for providing situational awareness, or the assessment of events, objects, and their relevance around the vehicle. However, in order to achieve AI solutions adapted to the railway domain, massive volumes of high-quality data are required. AI training and testing needs pre-recorded and synthetic scenes, data processing tools that imply complex computations and storage infrastructures. Needless to say, this huge task requires the collaboration of all the stakeholders in the rail sector.

Therefore, the creation of a complete and comprehensive Common Data Management Platform (CDMP) will benefit the whole railway ecosystem: from suppliers, who will reduce needs for initial investments, to clients, who will benefit from shorter time-to-market and better products. A platform of this magnitude would avoid AI caused drift, making it possible to move towards the goal of achieving sufficient operational maturity to enable autonomous operation. This maturity serves as the foundation for obtaining the needed certification procedures as well as guaranteeing predictable behaviour.

However, this is a huge and costly task and it would not be efficient to start from scratch without studying and absorbing the advances already proven in the automotive sector with many years of successful experiences. Although the two sectors do not share the same operating environment, there are many similarities to be found in the technology, platforms, methodologies and other tools necessary for the development of environmental sensing techniques.

The work presented in this article faces this challenge, and it proposes a stakeholder-agreed (European main railway players; operators, infra-managers, suppliers...) solution to the problem. This new contribution focuses on the definition of the Common Data Management Platform for artificial sense training, testing and certification and its validation through different Proofs of Concept (PoC).

This paper reports the results of that work: Section 2 resumes some pertinent previously published works from the automotive industry, first database attempts of the railway world and analysis about reusability/exportability of automotive experiences to railway. Section 3 provides the identified new and most relevant use cases to be covered by the CDMP. Section 4 describes the high-level platform definition based on identified requirements and the description of the main modules of the platform (data acquisition/generation, data labelling and training/testing). Section 5 focuses on data management (including the functional architecture that is suggested) to guarantee the system's scalability, modularity, and interoperability. Section 6 summarises the criteria of data protection and anonymisation. Finally, the PoCs, which are a reference guide describing how the data management platform was created for a particular use-case, are presented in Section 7. Section 8 draws the conclusions of this work and adds some suggestions for future work.

2. Related Work

2.1 MLOps Platforms

The convergence of advances in Deep Learning (DL), Big Data (BD) and High-Performance Computing (HPC) technologies has created a fruitful technology ecosystem that leverages the creation of effective AI systems in many sectors. In the context of smart mobility, AI has proven a technology that enables advanced functions such as sensorial perception, decision making, route optimization, and, eventually, automated driving functions.

The fuel of AI is data, and, as a consequence, many efforts are devoted to creating technologies to support data creation, management, processing and monitoring. As technology develops, a vast amount of libraries, platforms, applications, standards and initiatives are emerging and thriving in the Machine Learning Operations (MLOps) landscape. The MLOps concept extends the DevOps (Development and Operations) from the SW industry by adding data and ML-specific applications [1], and therefore includes a wide range of tools and perspectives. Some examples are versioning (DVC, Liquidata), labeling (Scale, OpenLabel), processing (Spark, Dagster), exploration (dbt, Rapids, pandas), data lakes/warehouse (snowflake, databricks), sources (S3, Parquet, Postgres), training (Pytorch, fast.ai, RAY, Hugging face), resource management (slurm, Docker), SW management (git, visual code), experiment management (mlflow, tensorboard, neptune, comet), hyperparameter tuning (sigopt, tune), monitoring (fiddler, grafana), edge (TensorRT, Onnx, TensorFlow Lite), Web (Kubernetes, Lambda, Seldom), CI/testing (Jenkins, circleci, buildkite), etc.

All-in-one MLOps solutions with integrated services already exist, mostly promoted by large cloud vendors, such as FBLearner, Google Cloud AI Platform, or AWS SageMaker. Other options are FloydHub, Paperspace, Gradient, Neptune or Domino Data Lab, to name a few. These offers include managed Platform-as-a-Service solutions that simplify technology choices and accelerate time-to-market development of ML solutions, at the cost of limiting the portability of the project, adhering to private formats, and elevating cloud infrastructure costs.

2.2 Data Sharing

Data sharing has become one of the main pillars of the EU Digital Strategy, with the publication of the European Strategy for Data, which includes the EU Data Act [2], that joins other regulatory initiatives that cover privacy such as General Data Protection Regulation (GDPR) or Artificial Intelligence development (EU AI Act).

As a response, Open Data initiatives, such as the International Data Space Association (IDSA) [3], Gaia-X [4], or the EU Open Data Portal [5] have been created. They establish principles, guidelines, reference architectures and guidance to standardisation, industry, academia, legal entities and national regulatory bodies.

European projects funded by the Horizon Europe programme are requested to produce Data Management Plans (DMP) that address FAIR data (Findable, Accesible, Interoperable and Reusable) strategies. These include identification of data and metadata types, mechanisms for publication, interoperability, clear licensing options, security, ethics and privacy preservation.

Standardisation of data formats (captured/generated data, metadata, pre-trained models…), data structures (labelled data, configuration files of sensors…), execution processes, and communication protocols between repositories or access policies are the biggest challenges facing these types of common and interoperable databases.

2.3 Automotive/Railway Databases

In parallel to the MLOps frameworks and the European-level regulatory framework, a large number of datasets are being released and made openly available (mostly for research purposes) for AI training and testing, containing millions of images, point clouds and other data from a variety of sensors. They are mostly used to train AI models for object detection/identification, navigation/positioning or environment monitoring applications. These datasets can be seen as practical exercises to effectively structure and distribute data with the purpose of being used for benchmarking purposes (e.g., KITTI [6], Virtual KITTI [7], nuScenes [8], Apollo [9]), to gain prestige (e.g., Audi A2D2 [10], Waymo [11], Ford [12], Lyft5 [13]), or pioneering in specific application domains (e.g., Woodscape [14] on fish-eye segmentation, DMD [15] on driver monitoring). Their estimated volume (at Q1 2023) sums up to more than 12TB of data, and more than 3800 hours of driving.

However, the lack of standardized data formats, the heterogeneity of the purpose-specific vehicle set-ups, and customized annotation models imply that different data parsers must be developed in order to test, train or validate algorithms or models for each data source.

Apart from perception-related datasets, platforms for scenario-based testing have been created (but not openly), such as Safety Pool [16], Streetwise [17], Scenius [18], AD-SCENE [19], or PEGASUS [20], containing scenario descriptions and tools to run virtual testing.

In the railway domain, the number of datasets is way more limited, and, to the best of our knowledge, no scenario-based initiative exists so far. Some datasets for camera-based AI training exist, such as RailSem19 [21], a dataset with more than 8500 images of rail traffic semantic annotations on rail scenes, and FRSign [22], which contains labeled images of French railway traffic lights. Datasets for semantic segmentation use LiDAR-based set-ups to produce point-clouds of railroad environments [23–25] or thermal images [26].

2.4 Analysis of the Reusability/Exportability of AI-related Technologies to Railway

The automotive sector has led AI-based pioneering advances in autonomous mobility with solutions for perception, decision-making, or navigation functions. In other domains, such as agriculture or railway mobility, the industry doesn’t aim to reinvent the wheel and adapt such advances to their own specificities.

In particular, railway and automotive operations have strong similarities. Use cases are often aligned: obstacle detection, traffic sign recognition, or vehicle localisation. Enabling technologies and scopes are also equivalent: (a) perception based on camera and range (RADAR, LiDAR) sensors, (b) functions for detection, identification, tracking and distance estimation, (c) digital map infrastructures and services, (d) vehicle-to-anything wireless communications, and (e) similar validation and verification technologies (data-driven, virtual testing, scenario-based evaluation, etc.).

Nevertheless, the gap exists, and differences need to be highlighted to specialize the AI-related technological choices. In some cases, differences impose additional challenges: (a) larger vehicle sizes imply more complex sensor set-ups, or (b) heavy dynamics imply longer sensing distances to actuate preventive braking maneuvers. However, in general, railway operation simplifies some of the road-level dimensions: (c) simpler motion dynamics (longitudinal paths), (d) higher levels of automation operation, (e) limited or pre-fixed driving tracks (more controlled infrastructure monitoring, preidentified risk areas such as level crossings), (f) more restricted environments and behaviour of other actors, or (g) less power/size limitations for AI equipment.

3. Use Cases

This section identifies use cases to be considered in the development of a common data platform. They contain key aspects enabling an effective usage of the CDMP to search, share and combine data from different sources where all participants can benefit from each other. Moreover, the cases suggest a framework to mutually improve the datasets and models through the possibility of reporting data/model analysis and completeness issues or detecting errors to the owners. In addition, it provides the developers with the ability to evaluate their models on predefined datasets for specific tasks with several use cases. For a clear overview, the various use cases are divided into five different groups:

- Platform workflow: The use case describes the life cycle of the data in the platform; from the collection and upload to the deployment of an individual dataset until its discard e.g., upload captured data and create dataset.
- Platform management: These use cases describe the process of, a) receiving access to the platform with specific rights (access and rights management), b) uploading new raw data to the data management platform and the ability to update each dataset (upload and update a dataset), c) ensuring the safety of the platform (ensuring data protection) and finally d) requesting specific additional data samples e.g., snowy environment, within the platform (request of additional data).
- Data quality: These use cases describe the process of, a) evaluating the completeness of a given dataset (analysis of dataset completeness), b) evaluating the quality of the dataset by considering the accuracy and correctness of a dataset by examining different aspects (analysis of dataset accuracy and correctness) and finally c) an uploaded raw dataset and its further supplementation which requires a careful pre-analysis (Data preparation and supplementation) e.g., cross-validate one of the uploaded dataset (for third party body) according to agreed standard.
- Data traceability: These use cases describe the process of, a) querying the CDMP for desired task-related datasets (searching the dataset in the platform) and b) discovering issues in a given dataset from the platform and reporting them to the dataset owner e.g., inform data owner with badly labelled data.
- AI model development: These use cases describe the process of, a) an AI model by taking or requesting desired datasets from the platform and splitting them into training, validation, and testing sets (training of AI model and model registry), b) evaluating a trained AI model in the CDMP for a defined task (testing a trained AI model) and finally c) specifying testing procedures and datasets for AI models solving defined tasks in railway (testing data management and specification) e.g., create an AI model from labelled samples and validate it with validation dataset.

4. Common Data Management Platform Definition

This section details the most relevant points when defining a data management platform: the requirements that it has to meet, both functional and operational, and the main core modules description that comprise the CDMP (data acquisition/generation, labelling and training/testing).

4.1 High-level Requirements Overview

This section overviews the fundamental requirements that establish the scope, functionality and expected methods to utilize the platform. Functional requirements determine what the platform is supposed to do and the operational ones define how to build and/or run the system. In general, operational requirements can be also understood as those non-functional requirements that determine other aspects such as performance expectations or standards to be used. The identified high-level general requirements are:

General functional: The platform shall be a container of data intended for its use in AI-related processes (training, re-training and testing) and editable by multiple users simultaneously. It shall enable Create, Read, Update, Delete (CRUD) operations, permit metadata to be contained, guarantee traceability, allow back-up/archiving options and also options to categorize/organize content according to use cases, domains or relevant tags.

General operational: As a general rule, the platform shall enforce the utilisation of standard file formats for sensor data, metadata and annotations. In addition, the platform shall be deployable in any local or cloud environment, be accessible via programmatic interfaces and expose callback entry points for networking protocols. It shall also include authorisation mechanisms to determine the level of access of the users.

Moreover, application-related requirements focus on AI-related applications (training, re-training and testing applications that provides the platform) must be considered. Finally, the specific utilization of the platform for AI applications mandates the definition of content-related requirements that specify characteristics of the content itself.

- Application-Related: The platform shall contain all data and metadata needed to feed an AI-related application:
- Training: The platform shall enable a neural network to be trained using a prepared dataset (data plus labels) to produce a model that can later on be used to predict labels on new data.
- re-Training: The platform shall enable the model to be re-trained (or produce an updated version of the model) using existing models and newly prepared datasets (using techniques such as data augmentation, filtering, and grouping...) and with the ability to measure the gain in performance, quality or any other key Performance Indicator (KPI). All of this, make use of incremental learning strategies (without repeating the training processes of previous steps).
- Testing: The platform shall provide mechanisms and tools to be able to design, define and execute AI model tests as well as to save and compare the results of each test for continuous testing.
- Visualization: The platform shall provide test visualization mechanism and inspection routines to analyze information about the dataset or model, e.g., balance of labels, subset KPIs, statistics of re-trained models, and analysis of extracted features...
- Content-Related: The platform shall contain training datasets in the form of data (multi-sensor recordings) plus annotations, trained models and hyper-parameters to configure all application processes (e.g., learning rate, initial weights, batch size, etc.). The platform shall also contain functional scenario descriptions for scenario-based testing, logical and specific scenario descriptions and real-world routes which can be matched with scenario tags to perform real-world tests.

4.2 Main Modules

A representation of all the logic modules that can be found in the platform proposed in this work. The following ones are considered as the most important core modules inside the whole platform.

Data Acquisition and Generation

The data that will populate the datasets on the platform can be captured in a real environment or can also be generated in a synthetic environment. A common understanding in the AI community is that AI models are as good as the dataset used to train them. For this reason, data representing the expected Operational Design Domain (ODD) with high fidelity, the situations under which the model is assumed to operate correctly, is a key factor.

In order to create the most realistic dataset of the environment, the most traditional method is to acquire data adding sensing capacity to the train or infrastructure. The most traditional options use different types of sensors (Cameras, RADAR, LiDAR, IMU...). According to the functionality that the perception system wants to cover, there shall be a sensor or group of sensors that fulfil the requirements. Table 1 makes the relation between the correct sensor and the most relevant use cases of the future autonomous train.

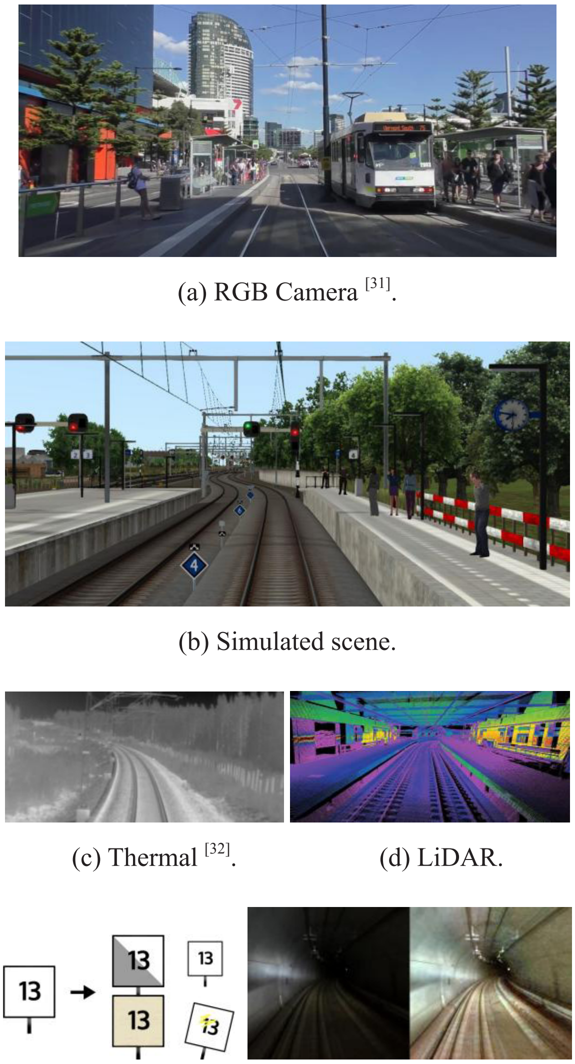
Table 1. Railway use-case vs needed sensors.

Use-cases/Sensor	RGB CAM	IR CAM	RAD.	LiD	Aud.	Odo.
Obstacle detection	x	x	x	x		
Sign/signal recognition	x			x		
Switch & path monit.	x		x	x		x
Train localization	x		x	x		
Ext. environment monit.	x	x	x	x	x	
Infrastructure supervision	x			x		
Platform monitoring	x	x		x		
Rolling stock monit.	x	x			x	x
Passenger’s supervision	x	x		x	x	

Note: RGB and IR cameras, RAD.—RADAR, LiD.—LiDAR, Aud.—Audio and Odo.—Odometry.

However, creating a dataset with real images and covering all operational conditions might be unaffordable, extremely expensive or very difficult to manage. In addition, assuming the recordings can be created, labelling them is usually the main bottleneck. These limitations focus the attention on synthetic data generation, either by creating limited discrete samples (created from canonical images or from generative deep learning) or sequences creating a completely new scenario from a virtual simulator (from simulator engines). The first approach implies the utilization of data augmentation and DL techniques such as EnlightenGAN [27] to increase the variability of the resulting dataset modifying seed examples and creating modifications (shadow, color, size, rotation, lightning...).

The second method simulates an entire recording campaign, running a simulator engine that creates a virtual world where the elements of interest are represented naturally and thus shows all the required variability. Modern simulation engines (CARLA [28], LGSVL [29], Prescan [30]) can be used via programmatic interfaces to produce virtual scenes with high-fidelity sensors, environments and behavioural models, etc. From these simulators, data can be gathered, manipulated and batch processed as desired. Figure 1 shows several examples of different types of data captured in real environments, as well as synthetic data created from simulators.



(a) RGB Camera [31].

(b) Simulated scene.

(c) Thermal [32].

(d) LiDAR.

(e) Augmentation.

(f) E-GAN.

Figure 1. Different data examples; a), b) and c) represents data acquired in real world using different sensor and at different railway environment; d), e) and f) are generated by simulators, data augmentation techniques and deep learning algorithms (E-GAN).

Data semi-automatic Labelling

The annotation files describe the organised rich description of the scene. These files contain labels for the scene's objects as well as other information such as sensor meta-data, encoding schemes for various geometries, connections to ontologies, knowledge repositories, and other external resources. In order to build databases that are shareable and inter-operable, these labels, containing the relevant information, should be stored and organized using a common/standard format. On the other hand, the annotation criteria which comprise the guidelines to be followed while making the annotations in order to prevent the annotator's personal interpretations, should be also agreed in order to achieve the objective.

Although there are previous formats, such as JSON schema or Google Protocol Buffers, that allow comprehensible annotation both for computers and people, the automotive sector has been inclined to propose and define a standard (unique international standard for multi-sensor labelling by now) for the raw content labelling for the training and testing of AI models. ASAM OpenLABEL [33] proposes a univocal procedure for classifying and describing the many elements/objects of the driving environment. Furthermore, it may be adapted to fit into the taxonomy requirements of a particular user or company as it does not define how to describe the real world (taxonomy). Because of this, OpenLABEL may be a reliable standard that applies to the railway industry.

Model Training and Testing

According to the specifications, the platform contains tools for training and testing AI models. Regarding training operations, the user will be able to submit his own code, establish a specific and private architecture, or choose the various well-known state-of-the-art training methods available in the platform. On the other hand, regarding testing processes, the validation and testing module of the platform will be able to execute inference batches and compare, in an automatic way, the output data using the standard AI validation metrics or custom metrics established by each user. For both case the previously labelled data will be used as input data for training, and the ground truth for testing. The generated models (well-tagged) could be stored in the same platform.

It is quite challenging to keep track of all these processes, especially if the user wants to compare different training and testing sessions and manage the built-in models throughout the different deployment phases. This platform will be able to warrant model traceability and continuous monitoring of the AI model performance in order to enable continuous integration and deployment (CI/CD) pipeline.

5. Data Management

Once the high-level requirements were gathered, it was possible to define the functional architecture of the CDMP, as shown in Figure 2. This architecture not only gives an idea about what components are the building blocks of the envisioned CDMP, but also about how data is expected to flow through them.

First of all, data are acquired by different kinds of sensors (e.g., cameras) in the Data Acquisition Unit (DAU) (also compatible with the necessary parameters to boost synthetic-data-generation processes) and sent to the Data Anonymisation and Provisional Data Storage Unit (DA-PDSU), where they are made complaint with GDPR requirements and safely stored until they are requested and sent to the Data Labelling Unit (DLU). In the DLU, users can make annotations on objects and enrich the scenes by giving context or adding information about actions. The data labelling stage shall be made compliant with the latest ASAM initiatives, such as ASAM OpenLABEL and ASAM OpenScenario. After this, both the anonymised data and the created labels are sent to the Main Data Storage Unit (MDSU), where they are stored and assessed by means of the Data Analysis Unit (DAnU). From this point, authorised users have access to validated data through the Data Downloading Unit (DDU) under reasonable request.

Given the mechanisms available nowadays to automatise the whole AI development pipeline, it was decided to provide the platform with additional, model-oriented functionality. This functionality starts in the Data Pre-Processing Unit (DPPU), where data are prepared for training according to the selected use case. This unit also splits the data into the training and validation datasets, which flow to the Model Development Unit (MDU), and a good, sterile testing dataset which flows to the Validation and Verification Unit (V & VU). After AI models are trained and optimised by means of the validation dataset, the best performing ones are sent to the V & VU, where they are evaluated against the test dataset. Models which are compliant with the V & V requirements are eventually transferred to the Model Registry (MR) and stored there so that authorised users can retrieve and utilise them.

Even if the certification process and the eventual model deployment and monitoring are not within the scope of this research work, they are represented in Figure 2 so that the CI/CD pipeline can be fully conceived. Table 2 contains some of the open-source technologies which could be utilised for some of the described steps.

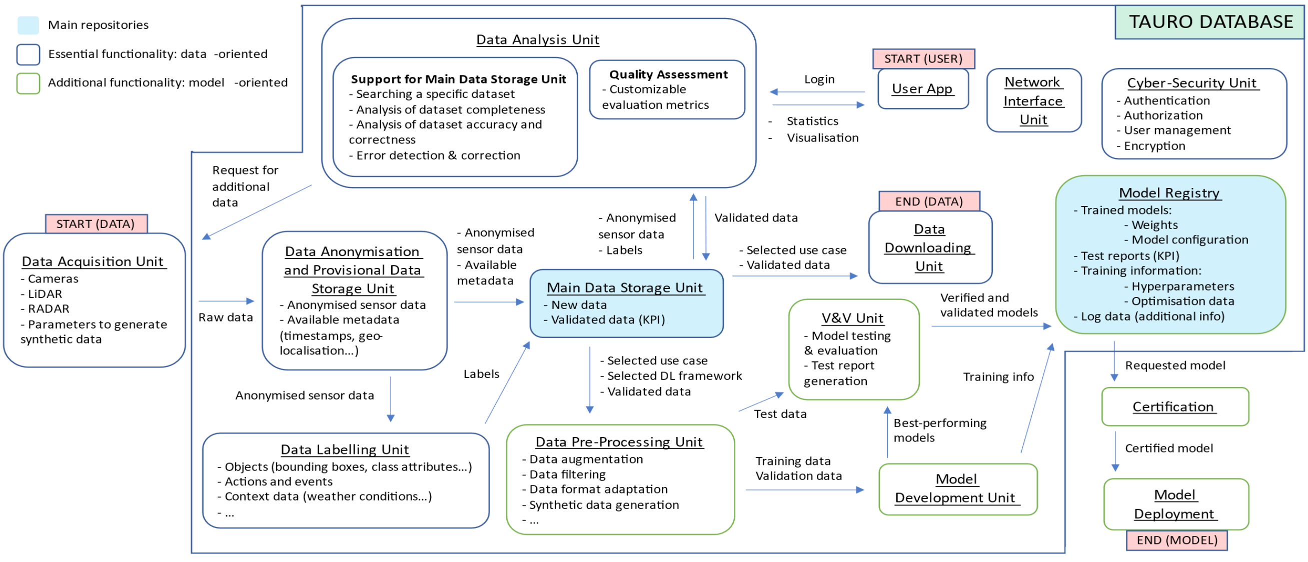


Figure 2. Functional architecture of the CDMP.

Table 2. Open-source technologies proposed for each pipeline stage.

Process	Technology
Raw data ingestion	HDFS, HDF5
Data cleaning + enrichment	Spark, Hadoop
Model training and tracking	MLFlow, Comet, TensorBoard
Scene Detection (ASAM	MongoDB + Elasticsearch,
Data Catalog	Neo4j, GraphDB
CI/CD	Gitlab CI/CD, Jenkins

6. Data Protection and Anonymisation

The collected data stem from different sensor types. Gathering data using cameras, audio or laser sensors is defined as a controlled activity by different governments and, consequently, these data are submitted to country-specific GDPR, especially where data collection includes personal information. In particular, the GDPR restricts the transfer of personal data to countries outside of the EU that do not have an equivalent level of protection. For this reason, the common data management platform must provide specific tools for data protection and anonymisation depending on the specific country laws. Detecting and blurring faces or texts (i.e., car number plates) are some of the particular scenarios. Moreover, customers and contributors shall comply with the local and global data protection standards and all parties should follow secure methods of data transfer to contribute to the CDMP. Finally, the data management platform shall permit storage of data in geo-specific locations to comply with GDPR policies.

There are two main issues that should be taken into account when analyzing privacy when data is stored in the cloud. Data collected by sensors that may have contained personal information is one of them. This data must be managed according to GDPR regulations, inaccessible to unauthorised users, and adequately safeguarded against data theft. The second issue is about user privacy using the cloud. It must also be protected with solutions [34], which is based on a non-bilinear group signature system, and can be used to provide anonymous authentication, where personal attributes can be proven without revealing the identity of the users.

7. PoC: Implementation and Case study

The proposed data management platform can be implemented with different flavors. On the one hand, the traditional on-premise solution has been joined by cloud and also hybrid options. On the other hand, the entire solution can be developed using different-level implementations. For instance, low level implementations offer a deeper understanding of processes, greater control, and lower costs. However, they also require a certain level of knowledge and expertise in different areas, and their configuration can be time-consuming. In contrast, high-level solutions, such as fully managed ML services, are available for those who are not capable of building a proprietary methodology.

In order to validate the proposed solution, two real-use cases were addressed of detecting 1) railway traffic lights and signals and 2) switches using YOLO models. For both cases, the data are annotated RGB images and a specific sub-pipeline was built in batch mode using microservice technology (from data gathering to model training and testing stages). However, the implementation is different depending on the use case. Cloud and on-site infrastructures are used for the former and latter cases, respectively.

7.1 Use Case 1 (Cloud): Railway Traffic Light and Signal Detection

In the first use case, the entire AI pipeline is performed on the AWS cloud. First, the training, validation and testing datasets were uploaded to the MDSU (S3 bucket). Secondly, the images were pre-processed in the DPPU: 1) filtered in order to obtain homogeneous traffic signal classes among the training, validation and test sets and 2) resized to fit the You Only Look Once (YOLO) [35] model. A few thousand images were manually annotated using Video Content Description (VCD) format [36] in the DLU. Next, the DL training and inference phase models were packaged as docker images allocated on Amazon Elastic Container Registry (ECR) for posterior training and inference tasks in the MDU and V&VU (Amazon Sagemaker), achieving a mAP@0.5 of 34.9% (AP@0.5 of 44.6% for traffic lights and AP@0.5 of 25.2% for traffic signs). Finally, the trained model was stored in the MR (S3 bucket). Figure 3 shows inferences made by trained AI models.

The total cost for the training process (100 epochs) on the most expensive tested option (ml.p3.2xlarge) was less than 1.59$ for a total computation time of 1874 seconds. Depending on the necessity, the solution is easily scalable to the required instance. During the training process, we incurred a fixed cost of 60$ month in terms of availability and maintenance of the AWS account and in additional costs of storage (docker images and image datasets), Virtual Private Cloud (VPC), maintenance and other costs that were negligible.

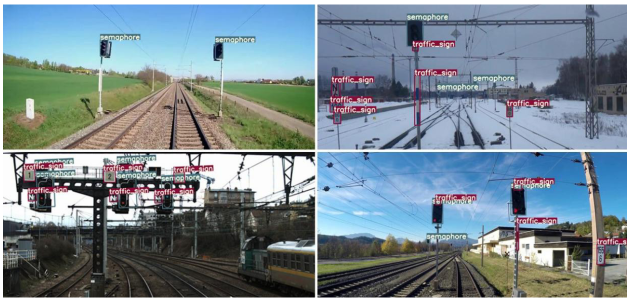


Figure 3. Railway signal & signs detection examples (FR-Sign database).

7.2 Use Case 2 (On-site): Railway Switch Detection

In the second use case, the entire AI pipeline was performed on-site. The objective of this use case was to detect railway switches and to determine whether they were open to the left or to the right based on the ego-perspective of a rail vehicle. Given that part of Railsem19's dataset is aligned with this purpose, this dataset was stored on a local MDSU. To complement Railsem19 and showcase a local implementation of the DLU, some additional frames were extracted from a private dataset belonging to CAF Signalling. Switches appearing on them were labelled by means of the Computer Vision Annotation Tool (CVAT). The labelling process consisted of drawing bounding boxes around the identified switches and naming them according to the classes they belonged to: "switch-left", "switch-right" and "switch-unknown". After that, all these data (images and labels) were downloaded in YOLO format and stored locally in the MDSU.

It is worth highlighting here that it was intended to split this use case into two consecutive functionalities. The first of them would locate all possible switches on an image under the general "switch-all" class and the second one would classify the status of these switches. The main reason behind this decision was if the different switch classes were very similar, it would actually make it possible to utilise visual features learnt from "switch-left" objects to locate "switch-right" objects and vice versa. Afterwards, a simpler DL model could be trained to differentiate between these two classes.

Considering this strategy, several data-preparing operations were made in the DPPU. First, Railsem19's data was filtered so that the switch-related data could be used for training. Then, Railsem19's labels were converted to YOLO format so that they could be merged with the self-annotated ones. After that, the data was prepared to train the two different kinds of models needed: a YOLO model for the switch location functionality and a custom Convolutional Neural Network (CNN) for the switch classification functionality. For both kinds of models, some data augmentation operations were also performed.

Finally, the AI models were trained on an NVIDIA GeForce RTX 2080 Ti and optimised over PyTorch (switch detector) and TensorFlow (switch classifier) in the MDU and evaluated in the V & V Unit, achieving the switch detector an mAP of 38.8% and the switch classifier an accuracy of 76% on a well-balanced test set. Both models were stored locally in the MR. Figure 4 shows inferences made by trained AI models.

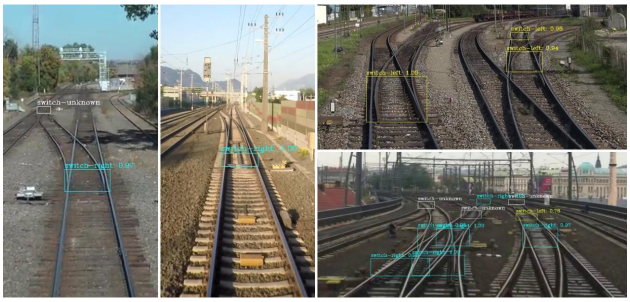


Figure 4. Railway switch detection examples (Railsem19 database).

7.3 Comparison

After performing both experiments, we highlight the benefits of having an entire solution hosted and managed in the cloud compared with an on-site solution. Developing and maintaining a local instance of the platform is very costly as the required back-end technology is resource-consuming. On the other hand, keeping the platform accessible with all the security and performance guarantees only increases this effort. Cloud solution is a better option in terms of availability, shareability, scalability and security. In addition, a microservices-based solution encourages versatility, efficiency, and low maintenance and finally high-level frameworks offer managed tools for boosting traceability (i.e., tracking datasets and model lineage). A summary of the proposed pipelines is shown in Figure 5, considering both the cloud and the on-site approaches.

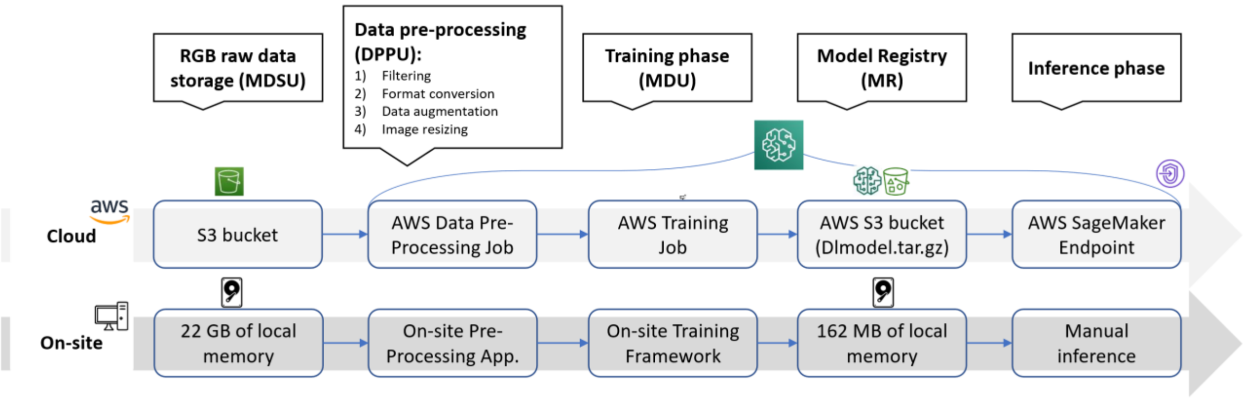


Figure 5. Proposed pipelines.

8. Conclusions and Future Work

This work presents a description of a common data management platform designed for the European railway sector based on agreed requirements and specifications among the main stakeholders. This platform enables developing perception systems that are robust and safe enough to make autonomous rail operation possible. To this end, the main tasks reported in this work are: an analysis of the state-of-the-art databases from the automotive sector as a very valuable input; identification of the most relevant platform use cases; definition of the envisioned CDMP; consideration of other aspects of data management; the establishment of access and contribution policies, and the development of the first instances of the data management platform as PoC based on the addressed case studies, and has complied with the most critical requirements. As a result, this platform has shown the capability of performing the key steps of the entire AI pipeline: data ingestion, data filtering, data labeling and AI model training and testing phases.

Considering the results obtained in the PoC, the cloud solution can be concluded to be a better alternative compared to an on-site solution for the data management platform construction in terms of availability, shareability, scalability and maintainability. In addition, the microservices strategy leads to a language agnostic pipeline that accelerates deploying AI models and building the CI/CD pipelines.

We can therefore conclude that the two major contributions of this work focus on the definition of the platform (from specifications to logical architecture, including the definition of use cases and security protocols) and the tests that validate its viability.

Furthermore, although approaches like the one described in this work are becoming de-facto standard in the automotive sector, the railway sector is still evolving to adopt AI-centered methodologies. At the time of writing this article, there are only a few remarkable open datasets related to AI for the railway domain. This work tries to bridge this gap, by designing and defining a common data management platform which could be better known as a common data platform.

This work will continue in the work forum within the R2DATO project of the European Europe's Rail programme, where the recently created "Data Factory" group will deepen each of the sections described here. The future iterations will eventually converge to the large-scale platform implementation and a scenario where all the interested parties in the EU share data and benefit from the data shared by others. They will also meet all the requirements and further discuss who will manage and host the platform, who will contribute, who will be able to use it and under what conditions.

Author Contributions

Mikel Labayen is the main author of the work. He has led the Work Package of the TAURO European project where the research has been carried out, participating in each and every one of the sections (from the analysis of the state of the art, to the concept tests, including the definition of the platform and use cases). He has also been in charge of leading the writing of this article. Daniel Ochoa de Eribe has contributed to the architecture proposal and proof of concepts (Sections 5–7), Marcos Nieto to the state-of-the-art analysis and platform definition (Sections 2–3), Ander Aramburu to the proofs of concept (Section 7) and Naiara Aginako to review the work done and the article written.

Conflict of Interest

NO conflict of interest.

Funding

This work was developed as a specific task within the TAURO research project [37] inside the Shift2Rail framework. This project has received funding under grant agreement No 101014984.

References

[1] Machine Learning Operations [Internet]. [cited 2023 Aug 10]. Available from: https://ml-ops.org/

[2] Data Act: Commission Welcomes Political Agreement on Rules for a Fair and Innovative Data Economy [Internet]. [cited 2023 Aug 10]. Available from: https://ec.europa.eu/commission/presscorner/detail/en/ip_23_3491

[3] International Data Spaces [Internet]. [cited 2023 Aug 10]. Available from: https://internationaldataspaces.org/

[4] Gaia-X [Internet]. [cited 2023 Aug 10]. Available from: https://www.data-infrastructure.eu/GAIAX/Navigation/EN/Home/home.html

[5] European Data [Internet]. [cited 2023 Aug 10]. Available from: https://data.europa.eu/

[6] Geiger, A., Lenz, P., Stiller, C., et al., 2013. Vision meets robotics: The kitti dataset. The International Journal of Robotics Research. 32(11), 1231–1237.
DOI: https://doi.org/10.1177/0278364913491297

[7] Cabon, Y., Murray, N., Humenberger, M., 2020. Virtual kitti 2. arXiv preprint arXiv:2001.10773.
DOI: https://doi.org/10.48550/arXiv.2001.10773

[8] Caesar, H., Bankiti, V., Lang, A.H., et al. (editors), 2020. nuscenes: A multimodal dataset for autonomous driving. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 11621–11631.

[9] Huang, X., Wang, P., Cheng, X., et al., 2019. The apolloscape open dataset for autonomous driving and its application. IEEE Transactions on Pattern Analysis and Machine Intelligence. 42(10), 2702–2719.
DOI: https://doi.org/10.1109/TPAMI.2019.2926463

[10] Geyer, J., Kassahun, Y., Mahmudi, M., et al., 2020. A2d2: Audi autonomous driving dataset. arXiv preprint arXiv:2004.06320.
DOI: https://doi.org/10.48550/arXiv.2004.06320

[11] Sun, P., Kretzschmar, H., Dotiwalla, X., et al. (editors). 2020. Scalability in perception for autonomous driving: Waymo open dataset. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 2446–2454.

[12] Agarwal, S., Vora, A., Pandey, G., et al., 2020. Ford multi-AV seasonal dataset. The International Journal of Robotics Research. 39(12), 1367–1376.
DOI: https://doi.org/10.1177/0278364920961451

[13] One Thousand and One Hours: Self-driving Motion Prediction Dataset [Internet]. Available from: https://arxiv.org/pdf/2006.14480.pdf

[14] Yogamani, S., Hughes, C., Horgan, J., et al. (editors), 2019. Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Korea. p. 9308–9318.

[15] Ortega, J.D., Kose, N., Cañas, P., et al., 2020. Dmd: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis. Computer Vision-ECCV 2020 Workshops. Springer: Cham. pp. 387–405.

[16] The Global Initiative for Certifiable AV Safety [Internet]. [cited 2023 Aug 10]. Available from: https://www.safetypool.ai/

[17] Streetwise: Accelerating Automated Driving with Advanced Scenario-based Safety Validation [Internet]. [cited 2023 Aug 10]. Available from: https://www.tno.nl/en/digital/smart-traffic-transport/smart-vehicles/streetwise/

[18] AVL SCENIUS [Internet]. [cited 2023 Aug 10]. Available from: https://www.avl.com/-/scenius

[19] A Path to a European Scenarios Database for ADS and ADAS Specification, Validation, and Homologation [Internet]. [cited 2023 Aug 10]. Available from: https://www.vvm-

projekt.de/fileadmin/user_upload/Mid-Term/Presentations/VVM_HZE_EmmanuelArnoux.pdf

[20] PEGASUS [Internet]. [cited 2023 Aug 10]. Available from: https://www.pegasusprojekt.de

[21] Zendel, O., Murschitz, M., Zeilinger, M., et al., 2019. Railsem19: A dataset for semantic rail scene understanding. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; 2019 Jun 16–17; Long Beach, CA, USA.

[22] Harb, J., Rébéna, N., Chosidow, R., et al., 2020. Frsign: A large-scale traffic light dataset for autonomous trains. arXiv preprint arXiv:2002.05665.
DOI: https://doi.org/10.48550/arXiv.2002.05665

[23] Cserep, M., 2022. Hungarian MLS point clouds of railroad environment and annotated ground truth data. Mendeley Data.
DOI: https://doi.org/10.17632/ccxpzhx9dj.1

[24] Lamas, D., Soilán, M., Grandío, J., et al., 2021. Automatic point cloud semantic segmentation of complex railway environments. Remote Sensing. 13(12), 2332.
DOI: https://doi.org/10.3390/rs13122332

[25] Yu, X., He, W., Qian, X., et al., 2022. Real-time rail recognition based on 3D point clouds. Measurement Science and Technology. 33, 105207.
DOI: https://doi.org/10.1088/1361-6501/ac750c

[26] Yuan, H., Mei, Z., Chen, Y., et al. (editors), 2022. RailVID: A dataset for rail environment semantic. ICONS 2022: 17th International Conference on Systems; 2022 Apr 24–28; Barcelona, Spain.

[27] Jiang, Y., Gong, X., Liu, D., et al., 2021. Enlightengan: Deep light enhancement without paired supervision. IEEE Transactions on Image Processing. 30, 2340–2349.
DOI: https://doi.org/10.1109/TIP.2021.3051462

[28] Dosovitskiy, A., Ros, G., Codevilla, F., et al. (editors), 2017. CARLA: An open urban driving simulator. Proceedings of the 1st Annual Conference on Robot Learning; 2017 Nov 13–15; California, USA.

[29] SVL Simulator by LG [Internet]. [cited 2023 Aug 10]. Available from: https://www.svlsimulator.com/

[30] Simcenter Prescan Software [Internet]. [cited 2023 Aug 10]. Available from: https://www.plm.automation.siemens.com/global/en/products/simcenter/prescan.html

[31] Melbourne Tram Drivers View [Internet]. [cited 2023 Aug 10]. Available from: https://www.youtube.com/watch?v=lMx1Bx2Ei08abchannel=Schony747

[32] Ristić-Durrant, D., Franke, M., Michels, K., 2021. A review of vision-based on-board obstacle detection and distance estimation in railways. Sensors. 21(10), 3452.
DOI: https://doi.org/10.3390/s21103452

[33] ASAM OpenLABEL V1.0.0 [Internet]. [cited 2023 Aug 10]. Available from: https://www.asam.net/project-detail/asam-openlabel-v100/

[34] Malina, L., Hajny, J., Dzurenda, P., et al., 2015. Privacy-preserving security solution for cloud services. Journal of Applied Research and Technology. 13(1), 20–31.
DOI: https://doi.org/10.1016/S1665-6423(15)30002-X

[35] Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M., 2020. Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.
DOI: https://doi.org/10.48550/arXiv.2004.10934

[36] Nieto, M., Senderos, O., Otaegui, O., 2021. Boosting AI applications: Labeling format for complex datasets. SoftwareX. 13, 100653.
DOI: https://doi.org/10.1016/j.softx.2020.100653

[37] Technologies for Autonomous Rail Operation [Internet]. [cited 2023 Aug 10]. Available from: https://cordis.europa.eu/project/id/101014984

[38] R2DATO [Internet]. [cited 2023 Aug 10]. Available from: https://projects.rail-research.europa.eu/eurail-fp2/

[39] Europe's Rail [Internet]. [cited 2023 Aug 10]. Available from: https://rail-research.europa.eu/