UNICO I+D Project
6G-DATADRIVEN-02

6G-DATADRIVEN-02-E8

Draft System architecture

Abstract

This document presents the initial system design of the 6G-DATADRIVEN-02 architecture. It defines the main entities and building blocks present in the system architecture to achieve an automated and zero-touch deployment using AI/ML mechanisms in Industry 4.0 scenarios. The document considers a pool of factory floors assisted by a central cloud to perform Industry 4.0 related tasks with the help of AI/ML and an autonomous orchestration loop. This deliverable focuses in defining the AI processing, data sources, interconnectivity, and process automation.

Financiado por la Unión Europea NextGenerationEU

GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES

Plan de Recuperación, Transformación y Resiliencia

Document properties

Document number	6G-DATADRIVEN-02-E8
Document title	Draft System Architecture
Document responsible	Carlos J. Bernardos (UC3M)
Document editor	Jorge Martin-Perez, Carlos J. Bernardos (UC3M)
Editorial team	Jorge Martin-Perez, Carlos J. Bernardos (UC3M)
Target dissemination level	Public
Status of the document	Final
Version	1.0
Delivery date	30-11-2022
Actual delivery date	30-11-2022

Production properties

Reviewers	Antonio de la Oliva (UC3M)

Disclaimer

This document has been produced in the context of the 6G-DATADRIVEN Project. The research leading to these results has received funding from the Spanish Ministry of Economic Affairs and Digital Transformation and the European Union-NextGenerationEU through the UNICO 5G I+D programme.

Contents

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

List of Figures

Financiado por la Unión Europea NextGenerationEU

GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES

Plan de Recuperación, Transformación y Resiliencia

UNICO I+D

List of Acronyms

5GC: 5G Core

AI: Artificial Intelligence

API: Application Programming Interface

ARIMA: Autoregressive Integrated Moving Average

B5G: Beyond 5G

DT: Digital Twin

IoT: Internet of Things

LSTM: Long Short Term Memory

ML: Machine Learning

NR: New Radio

RU: Radio Unit

SFC: Service Function Chain

SDN: Software Defined Network

SQL: Structured Query Language

PNF: Physical Network Function

VLAN: Virtual Local Area Network

VNF: Virtual Network Function

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

Resumen Ejecutivo

Este documento proporciona un diseño inicial de la arquitectura del proyecto 6G-DATADRIVEN-02, caracterizada por el uso de la inteligencia artificial para gestionar la industria conectada. La arquitectura propuesta integra herramientas de inteligencia artificial y colección de datos para automatizar las tareas de mantenimiento y producción en entornos de industria conectada.

Los principales resultados descritos en este entregable son:

- el diseño de una entidad de inteligencia artificial en industria conectada;
- el diseño de una entidad destinada a la colección de métricas en entornos industriales;
- la descentralización del procesado de métricas para garantizar la privacidad de datos;
- el entrenamiento federado/centralizado de modelos de inteligencia artificial para mejorar el rendimiento de las tareas de producción y mantenimiento;
- el registro automático de dispositivos para integrarlos en el proceso de producción; y
- la automatización del despliegue de servicios de industria conectada.

En línea con la arquitectura propuesta en el presente documento, se ha llevado a cabo investigación relacionada con la industria conectada usando inteligencia artificial. En concreto, se ha publicado:

- una solución que permite usar inteligencia artificial para mitigar la interferencia inalámbrica en el control remoto de un brazo robótico (Milan Groshev J. M.-P., 2022) (Milan Groshev J. S.-P., 2022); y
- la formulación del problema de despliegue de servicios de robots para entornos de industria conectada (Applications, 2022).

El resto del documento está redactado en inglés, de cara a maximizar el impacto del trabajo realizado en este proyecto.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

Executive Summary

This document details the initial system design of the architecture for 6G-DATADRIVEN-02, which relies on artificial intelligence to manage the connected industry. The proposed architecture integrates artificial intelligence and data collection tools to automate maintenance and production tasks in connected industry environments.

The main results described within the deliverable are:

- the design of an artificial intelligence entity for connected industry;
- the design of an entity that collects metrics in an industrial environment;
- the decentralization of the metrics' processing to ensure data privacy;
- the training of federated/centralized models of artificial intelligence to boost the production and maintenance tasks;
- the automatic registration of devices to integrate them in the production pipeline; and
- the automation of connected industry service deployment.

Inline with the proposed architecture within this document, we have done research related in the connected industry area using artificial intelligence. Specifically, these are the corresponding publications:

- a solution to mitigate the Wireless interference of remotely controlled robotic arms (Milan Groshev J. M.-P., 2022) (Milan Groshev J. S.-P., 2022); and
- the formulation of the problem related to the deployment of robotic services in connected industry (Applications, 2022).

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

1. Introduction

This deliverable aims to define the overall system architecture of 6G-DATADRIVEN-02. The scope is to establish the necessary entities and building blocks to allow an autonomous and zero-touch deployment of AI/ML services for Industry 4.0.

The overall architecture considers a pool of factory floors connected to a central entity that coordinates their operation. Also, the designed architecture considers local operation at the factory floor Edge premises. Hence, this document defines a cooperative Edge and Cloud architecture that resorts to 5G and B5G connectivity to stitch the Industry 4.0 devices with the AI/ML and data storage entities within the Edge and Cloud.

The present document defines the AI processing architecture that allows both federated learning and centralized approaches to derive AI/ML models necessary in Industry 4.0 tasks. The document specifies the pipeline and different components required to process the data coming from the sources, and identifies the technologies that are available to perform such tasks.

Additionally, this document explains how the proposed architecture follows an automation design to achieve zero-touch deployment of Industry 4.0 services, and AI-based orchestration. As a result, the proposed architecture autonomously allocates computing and network resources to run the trained AI/ML models that performs the Industry 4.0 tasks. Moreover, upon the authentication of new devices in the factory floor, the proposed architecture detects it and triggers re-orchestration to include new devices in the production pipeline.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

2. General system architecture

In this section we describe a general system architecture to exploit the usage of distributed data in industrial environments.

We consider a pool of *factory floors*, each of them with IoT and industrial devices as actuators, surveillance cameras, sensors, robots, etc. Each factory floor is provided with a 5G/B5G setup that provides connectivity to the *IoT* devices. Through the *5G/B5G connectivity*, IoT devices exchange data related to the production process to the Edge premises within (or not) the factory floor. It is at such Edge premises, where the IoT data is gathered to assess the pertinent data processing or management through *AI* methods. For example, the sensor data from a rotator may indicate that the rotator is vibrating more than usual, and an AI method/algorithm may forecast its incoming break, thus, indicating the machine to stop. Additionally, the collected data could be fed into the Digital Twin (*DT*) of the factory floor to enhance the production monitoring.

Each factory can belong to a *pool of factories* that exchange their data to enhance the production process, e.g., by sharing vibrating patterns of machines that resulted into malfunctioning in other factory floors, thus, enhancing the preventive maintenance procedure. The inter-factory connectivity allows the data exchange so as the coordination between local DTs and AI algorithms at each Factory Floor.

Additionally, every factory floor within the pool shares data with a *Cloud central entity* that has a global view of the Factory pool. The Cloud central entity is useful to do a global monitoring of a company/cooperative that controls a pool of factory floors. As the Edge premises of each factory floor, the Cloud central entity has a DT that has a global view of the factory pool status; and it also has a holistic AI that can cooperate with the AI of each factory floor to take smart decisions within the whole factory pool.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

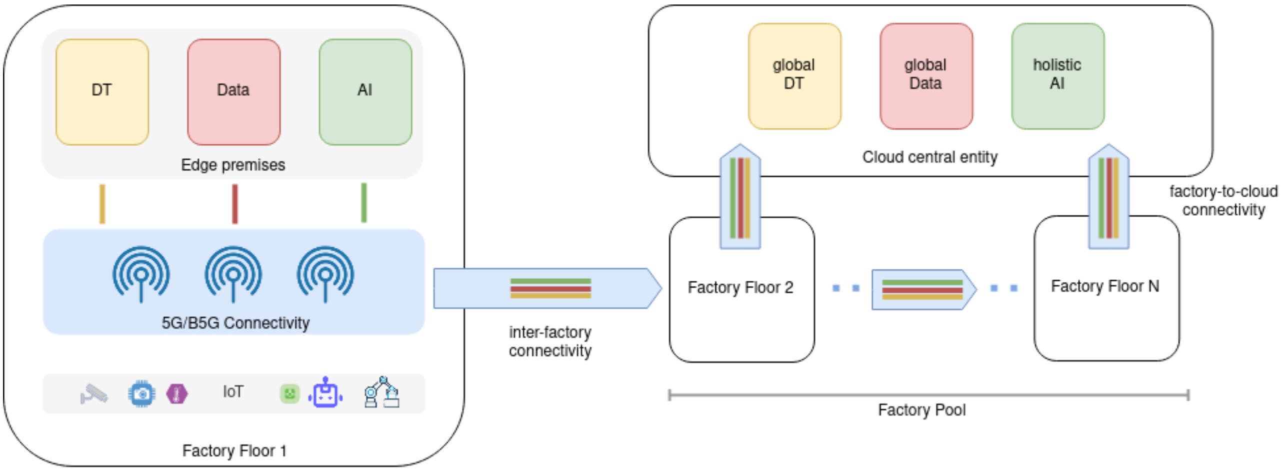

FIGURE 1: GENERAL SYSTEM ARCHITECTURE

Figure 1 provides an overview of the general system architecture, that is, the main building blocks that compose the architecture for the exploitation of distributed data in industrial environments. In the following we enumerate and detail each element/entity within the general system architecture:

- *Factory floor*: we refer to a physical building where the machinery of a factory plant is.
- *IoT pool*: each factory floor hosts a set of IoT devices as cameras, sensors or actuators. The industrial machinery also belongs to the IoT pool, for it also belongs to the set of devices that produce data and information to recreate the status of the factory floor.
- *5G/B5G connectivity*: we refer to the set of networking devices that connect the IoT pool to the internet. In particular, it consists of a set of radio devices as NR antennas with their respective 5G Core (5GC) stack to process the incoming traffic. Additionally, the 5G/B5G connectivity comprises all the switches, routers, Ethernet, and fiber links that interconnect the 5G/B5G stack with the Edge premises and Cloud central entity.
- *Edge premises*: it is the set of servers and NFS that hosting the DT, Data, and AI functionality. The Edge premises could either be located within the factory floor (On premise Edge), at the operator premises (Fat Edge), or at any other levels (local, regional, etc.). The aim of the Edge premises is to provide computational capabilities to tackle the data handling and AI tasks to automate the factory floor tasks.
- *DT* entity: it is the set of software entities that reproduce a DT of the Factory Floor. It ranges from simulation software, numerical models, and visualization tools that mimic the behaviour of the IoT pool elements performing industrial tasks, e.g., a robotic arm.
- *Data entity:* we refer to the data storage software and files that contain historic and up-to-date information of the reports produced by the IoT pool within the factory floor. As an example, the Data entity may contain a database storing the number of goods produced by a factory line, and what is the health-status of the robotic arms involved in the process.
- *AI* entity: it contains the whole pipeline to perform AI tasks, ranging from data pre-processing scripts, AI/ML models for different tasks, and training environments to increase the AI/ML

models accuracy. For example, the AI entity may host the pipeline to perform classification tasks that help to decide whether the produced goods pass or not the quality control check.

- *Inter-factory connectivity*: it comprises both the physical and logical internet connectivity between the factory floors. Specifically, it is the set of fiber links, switches, routers, VLANs and IPsec tunnels that allow the connectivity between factory floors within the same pool,yet preserving the data privacy and integrity throughout the communication. Thanks to the inter-factory connectivity is possible to establish a connection to exchange information/data among the DT, AI and Data entities of different factory floors.
- *Factory-to-cloud connectivity*: same as the inter-factory connectivity, but stitching the Factory floor and Cloud central entity elements. Consequently, the exchange of information over it allows the factory floor to feed the cloud central entity with reports on the data, DT and AI models' status.
- *Cloud central entity*: it is the set of servers, and data storage resources that contain an holistic and general view of the factory pool. The cloud central is typically a pool of computing and data resources that are located at some cloud facility with elastic capabilities that can grow according to the computational needs of the factory pool.
- *Global DT entity*: it contains a synced DT of the factory pool, i.e., the entity hosts also the set of simulation, models and visualization tools that sync with the Dts of each factory floor to produce a global DT. The Global DT can mimic the current and even future behaviour of the whole factory pool, hence, resulting in a real-time estimation of the performance of the industrial facilities.
- *Global Data entity*: as its factory floor counterpart, the Global Data Entity is the set of data storage mechanisms and files that gather the IoT reports of all the factory floors. The Global Data entity may be a mirror of the whole factory floor Data entities, or contain only the information of interest for both the global DT and holistic AI.
- *Holistic AI entity*: as the AI entity within the factory floor, the Holistic AI is the pipeline that allows the data pre-processing, defines the set of models, so as the training tools. The Holistic AI can communicate with the factory floor AIs through the Factory-to-cloud connectivity, and use their partial knowledge to assess active learning and federated learning procedures. Hence, it takes advantage of partial learning stages of the factory floors to boost the performance of a Holistic AI that would foresee a vast casuistic consisting on the aggregated knowledge of all the factory floors AI.

The aforementioned general system architecture gives an overview of the high level entities involved in the 1st draft of our system architecture. Figure 1 illustrates the different entities and their interaction, so as the hierarchy of all components.

In the following sections we enter in detail into the description of the AI processing architecture, and the Network architecture. Both architectures detail how the AI entity and 5G/B5G connectivity stitch the entities within the general system architecture.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

3. AI processing architecture

In this section we describe the AI processing architecture running at the Edge premises of a factory floor. The AI entity of each factory floor retrieves the data coming from the IoT devices to extract, process, and exploit useful information for the factory floor maintenance and automation tasks. The main goal of the AI entity is to assess the tasks that it has designated within the factory floor. Such tasks have an associated a target like detecting a defect in the factory lane, and a metric to measure the accuracy of how the AI performs such task.

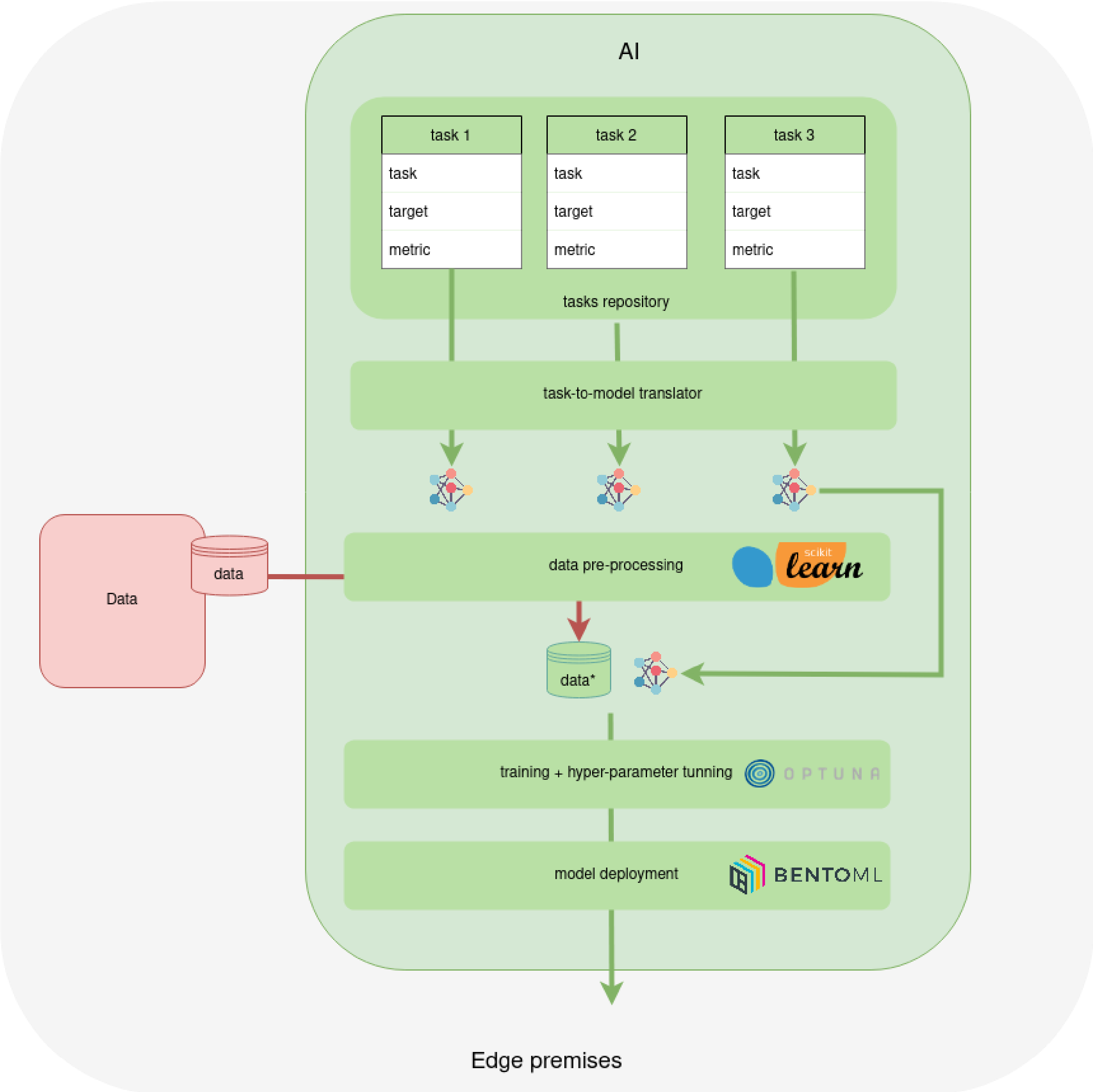

FIGURE 2: AI ENTITY ARCHITECTURE AT THE EDGE PREMISES OF A FACTORY FLOOR

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

Figure 2 illustrates the AI processing architecture building blocks. These are the tasks repository, task-to-model translator, data pre-processing, training+hyper-parameter tuning, and model deployment stage. These building blocks provide an end-to-end pipeline that takes as input data and tasks definitions, and deploys models to perform AI tasks in the factory floor.

In the following we describe each building block of the AI processing architecture:

- *Tasks repository*: it stores a list of tasks, each of them defined in a descriptor file (e.g., a YAML). The factory floor owner onboards in the tasks repository the list of tasks to perform at the factory floor. A task may be to do the quality check of the goods produced along a factory lane.
- *Task-to-model translator*: its duty is to translate a task into an AI/ML model that can perform such task. It achieves such translation by checking a list of pre-stored AI/ML models that suit the task to perform. For example, for a predictive maintenance task it will output a model as ARIMA or an LSTM neural network.
- *Data pre-processing*: prior to the execution/training of the AI/ML model it is necessary to do data pre-processing. This entity uses the scikit learn library to do data normalization, NaN filtering, and train/test splits. As a result, it produces a dataset that is ready to be fed to the AI/ML model.
- *Training+hyperparameter tunning*: it takes as input the pre-processed dataset and AI/ML model, and performs the corresponding training. Namely, it uses the optuna library to do hyperparameter tunning of the model. Optuna receives the metric to be optimized within the task, e.g., the F1-score to check whether the quality control accurately classified a piece with a problem within a production lane. Then, Optuna performs multiple training sweeps modifying hyperparameters of the chosen model, e.g., changing the number of LSTM neurons within a hidden layer of a neural network. The advantage of using Optuna is that it works with multiple AI/ML libraries for neural networks and statistics as pytorch, chainer, tensorflow, MXNet or Scikit-learn. Consequently, we can achieve an automated training process for a wide variety of AI/ML models for the factory floor.
- *Model deployment*: after the AI/ML model is trained inside the Training+hyperparameter tuning, the model deployment entity uses BentoML to put in production the trained model. Namely, a predictive maintenance ARIMA model is deployed by BentoML and receives HTTP requests with historic data to tell whether there will be an imminent failure or not.

Note that the AI architecture in Figure 2 shows the different steps within the Edge premises of the factory floor. However, Figure 1 illustrates that AI/ML entities of different factory floors may interact among them, and the cloud central entity have a holistic AI/ML that works over the pool of factory floors. In the following we detail both the interaction between AI/ML entities in the pool factory floors, and the holistic AI/ML in the cloud central entity.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

3.1. Federated learning between factory floors

This section specifies how to the AI/ML of different factory floors interact to achieve federated learning. By using federated learning it is possible to preserve the privacy of the data produced by each factory floor, yet cooperating with others to achieve a common AI/ML model using the knowledge of all factory floors.

Figure 3 illustrates how the federated learning would work between the edge facilities of two factory floors. Both, factory floor 1 and factory floor 2, belong to a common pool for the exchange of information. For example, both factory floors may belong to an automotive consortium.

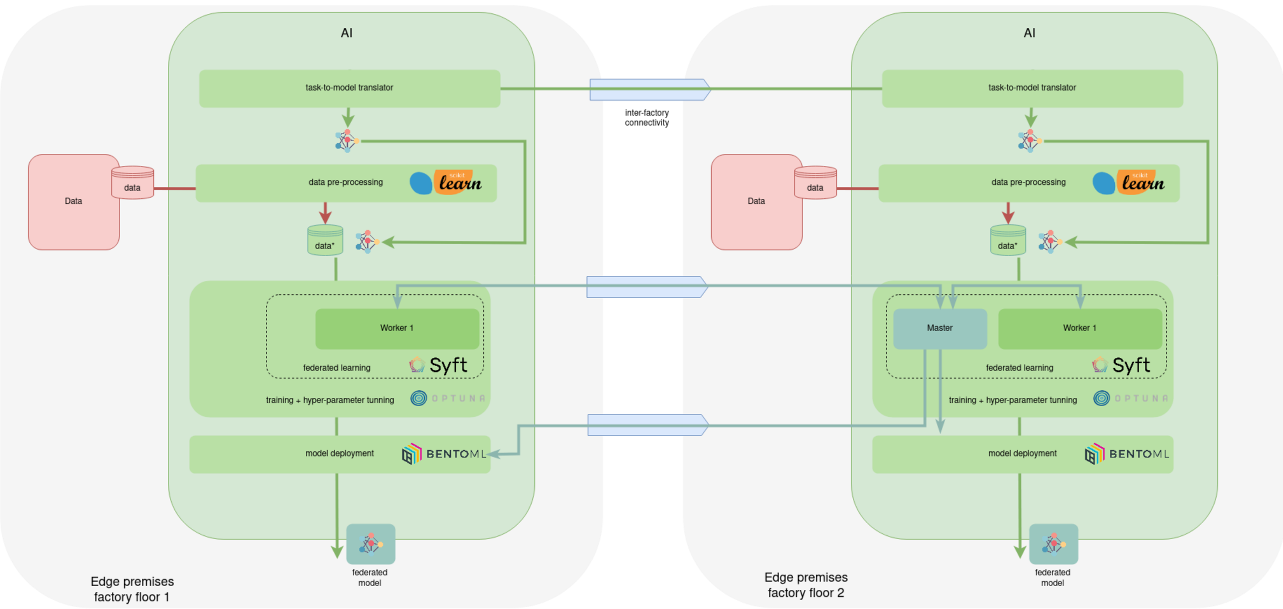

FIGURE 3: FEDERATED LEARNING BETWEEN FACTORY FLOORS FLOORS

In the federated learning paradigm there are multiple workers that train the AI/ML model with local data, i.e., each factory floor trains at its edge premises a model using its data. The training of each factory floor results in a local model that coordinates with other factory floors' models to share the learning. As a result, the federated learning yields a federated AI/ML model that is common to all factory floors although, yet with each being agnostic to others' data.

Fig. 3 illustrates the federated learning with two factory floors involved in the procedure. We now enumerate all the steps (arrows) within Figure 3, which explain the training stage:

1. the *task-to-model translators* of each factory floor agree on a common AI/ML model to use for a specific task. Such agreement happens over the inter-factory connectivity;
2. each AI entity uses the data locally stored within the factory floor, and selects the common AI/ML model agreed by both factory floors;
3. one of the factory floors is selected as the master in the *federated learning* training procedure (factory floor 2), and triggers the training stage at the *training+hyperparameter tuning* entity using the Syft library;

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

4. the master gathers the weight updates of each worker at every epoch, and average them to send weight updates to each worker. Note that at each epoch of the training stage optuna keeps track of the model performance to assist the master in the hyperparameter tuning;
5. once the training finished, the master forwards the common federated model to the *model deployment* entity of each factory floor, which deploys with BentoML and HTTP API to interact with the resulting AI/ML model.

Although Figure 3 only illustrates the training for two factory floors, it is worth remarking that the procedure applies for as many factory floors as we want. As long as all factory floors belong to the same pool, the master factory floor can synchronize with all of them to issue the federated training. The only difference is that it would have N workers, rather than the 2 workers of Figure 3.

3.2. Cloud central entity holistic AI

The cloud central entity has a holistic AI that serves for the purpose of training AI/ML models for all the factory floors. Note that this initial architecture does not enter into the details on where the holistic AI is hosted, contrary to the factory floor counterparts, which are equipped with Edge facilities. The cloud central entity may count with a pool of computed resources hosted on premises or not, as the case of a remote cloud provider.

The cloud central entity connects with each factory floor via the factory-to-cloud connectivity. Such connection is necessary for the exchange of data between the central cloud entity and factory floors, and also to exchange AI/ML related information as trained models, or training interaction.

The holistic AI trains a central AI/ML model using a centralized learning procedure, or a federated learning approach; as detailed in the next subsections.

3.2.1. Centralized AI/ML training

If the AI/ML model is trained in a centralized fashion, as depicted in Figure 4, each factory floor should exchange its own data with the central cloud entity. Such data may be reported:

1. directly by the Data entities of each factory floor (Figure 4); or
2. by the data pre-processing entities of each factory floor.

The former case is the most straightforward for it does not need a prior pre-processing, and leaves such task to the data pre-processing entity within the central cloud entity holistic AI. However, the absence of a pre-processing stage at the factory floor may lead to the transmission of useless data that could have been filtered out, or even the exposure of privacy-sensitive information. Hence, going for option 2. is of special interest when privacy and few data transfer is desired.

Either 1. or 2. is selected to transfer the information to the Data entity of the central cloud entity, the holistic AI follows the same training procedure as in Figure 2. That is, it takes all the data gathered

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

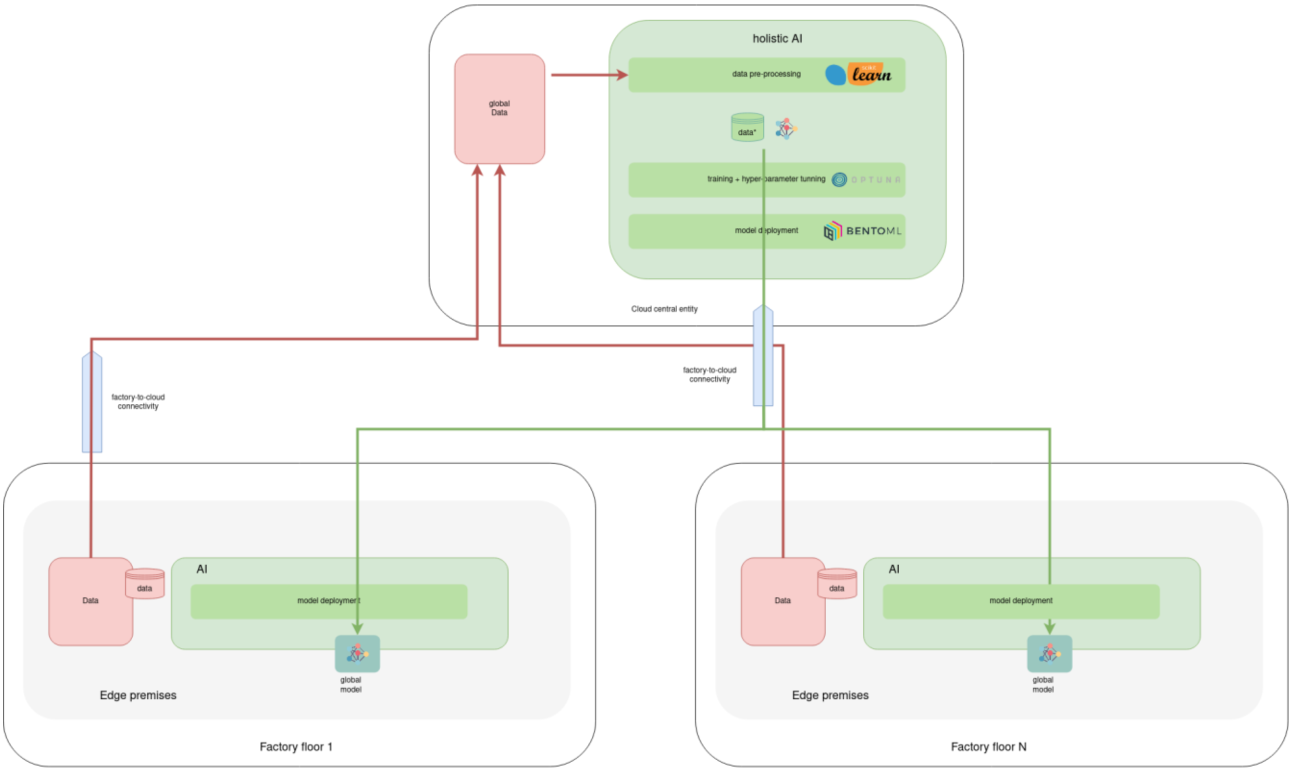

FIGURE 4: CENTRAL AI/ML TRAINING PROCEDURE

from every factory floor within the pool, to perform filtering and training, and hyperparameter tuning. As a result, the holistic AI model deployment forwards the trained AI/ML model to the deployment entity of each factory floor. Finally, every factory floor ends up with an AI/ML model trained by the holistic AI – see Figure 4.

3.2.2. Federated AI/ML training at the central cloud entity

In case the central cloud entity wants to carry out a federated learning procedure, it can replicate the procedure among factory floors that coordinate to carry out a federated learning. Contrary to such case (Figure 3), now the central cloud entity has the master within the holistic AI.

First of all, the holistic AI agrees with the AI entities of each factory floor which is the AI/ML model that they will use within the federated learning stage. Differently than in Figure 4, the central cloud entity does not require any data to perform the training. Its master node interacts with the workers of each factory floor to synchronize at the training stage. It exchanges with all of them the averaged weights received from the AI/ML models trained locally within each factory floor, and produces a federated AI/ML model that passes to the model deployment entity.

Finally, the BentoML at the central cloud entity exchanges the trained AI/ML model to the model deployment entities of each factory floor, so each requests the necessary tasks to the trained model, e.g., forecasts about imminent failures.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

Note that this procedure presents two advantages with respect to the centralized training of the prior section. The first one is that the data is locally pre-processed and only resides at each factory floor, thus, preserving the privacy. Secondly, the data exchange with the holistic AI only carries weight updates, which will be presumably smaller than exchanging their whole datasets.

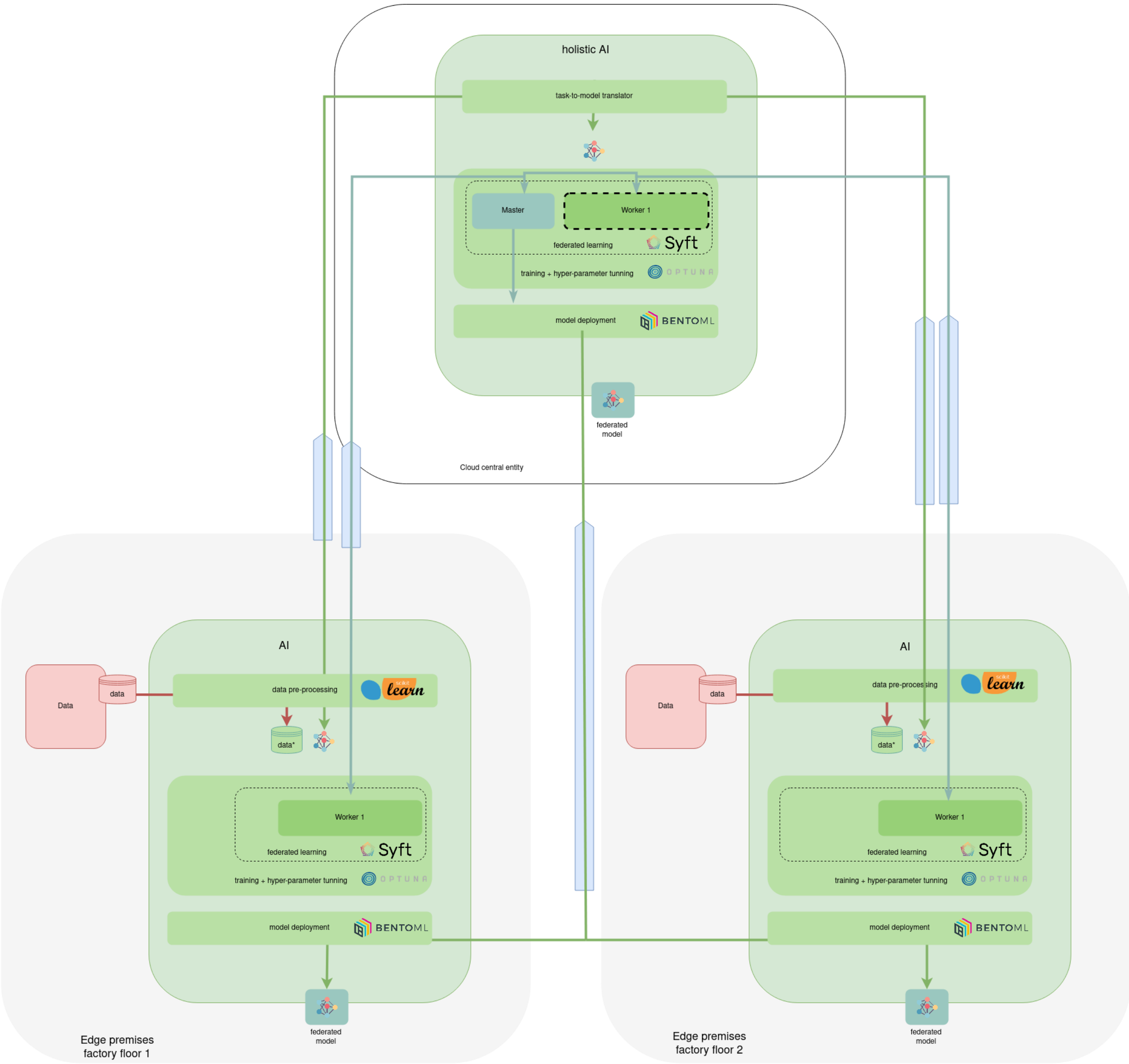

FIGURA 9: FEDERATED AI/ML TRAINING

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

4. Data sources

The factory floor collects data of the IoT devices as thermometers, state of robotic arms, etc. All these metrics are reported to the Data entity within the factory floor for internal uses as predictive maintenance tasks, error recovery, etc. Additionally, the Data entity of the factory floor keeps track of the devices available at its own premises: the network and IoT devices. To prevent non-authorized devices from trying to maliciously connect to the factory floor, there is an Authentication step within the factory floor.

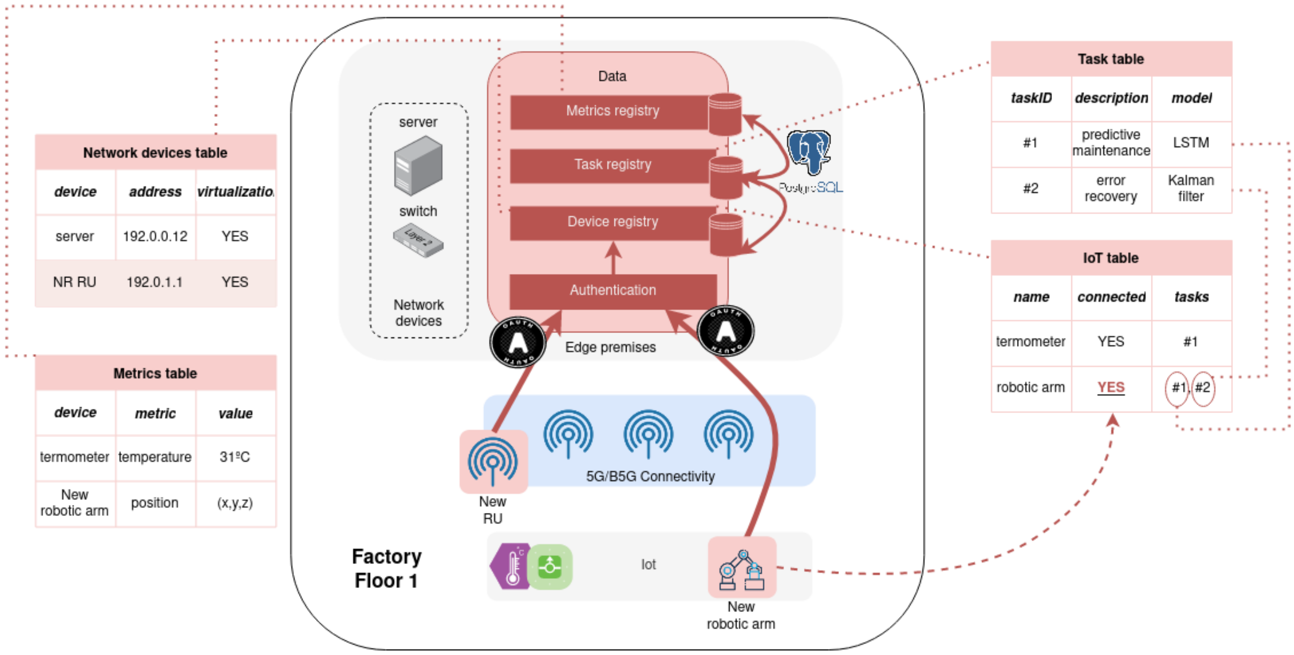

FIGURE 6: DATA SOURCES WITHIN THE FACTORY FLOOR

Figure 6 illustrates the data sources available within the factory floor, all of them captured by SQL tables stored in PostgreSQL at the factory floor premises. In the following we detail both the Authentication building block within the Data Entity, so as the registries holding the factory floor data sources:

- *Authentication*: this building block is the entry point for new devices attaching to the factory floor premises. Either if they are network or IoT devices (as the New RU and New robotic arm in Figure 6, respectively), they should do an OAuth challenge with the Authentication building block, so the latter verifies that they can operate at the factory floor.
- *Device registry*: this building block keeps track of the IoT and Network devices tables that store the factory floor devices. The network and IoT devices will only appear at such tables if the Authentication building block succeeded in the OAuth challenge. With the IoT table and Network devices tables it is possible to know which are the available devices to connect the IoT devices with the related AI/ML models as, e.g., the predictive maintenance. Furthermore,

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

upon the attachment of new devices (as the New robotic arm in Figure 6) it is possible to report the orchestration AI/ML model (see Next section) the availability of a new IoT device to consider it for a given task, e.g., the predictive maintenance. Regarding the network devices table, it contains records on the addresses of each device, and whether it has virtualization capabilities or not, so it is possible to deploy components on top of them in the orchestration stage.

- *Tasks registry*: this building block records the existing tasks within the factory floor, e.g.: the predictive maintenance, or error recovery shown in Figure 6. This information is kept within the tasks table, which also stores the AI/ML model used to perform the task. For example, the predictive maintenance model may be an LSTM neural network that was trained and deployed as specified in Section 3. It is worth mentioning that each task within the tasks table has a taskID that is associated to each IoT device in the IoT table. In such a manner, it is possible to have a one-to-many mapping of the task and associated devices, respectively. Specifically, Figure 6 shows that the New robotic arm is associated to the predictive maintenance, and its successful authentication leads to considering itself at the execution of the predictive maintenance task.
- *Metrics registry*: this building block keeps track of the metrics reported by the IoT devices at the Edge premises of the factory floor. It stores information related to the temperature, location of the robotic arms manipulators, etc. It is the content of the Metrics table of Figure 6 which keeps track of historic data used by AI/ML models to execute their tasks. For example, the temperature report may help a predictive maintenance AI/ML model to infer whether some IoT pieces may break soon.

It is worth mentioning that the Data sources at the factory floor may also exchange the data with the cloud central entity. In such a case, the global Data entity (see Figure 1) also has a metrics, tasks, and devices registry building block that collects data from the respective building blocks at the factory floor facilities. Consequently, the AI/ML models at the holistic AI can process global data to take better decisions of certain tasks.

Note that the data entity of each factory floor is isolated and only may share data with the central cloud entity. Therefore, the data is not exchanged among factory floors, which prevents data privacy leakages.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

5. Automation

One of the main goals of the present document is to achieve automation within the factory floors through the proposed architecture. With automation we refer to zero-touch deployment and AI-based orchestration. The former is understood as the possibility of deploying services in a plug-and-play fashion, i.e., without the necessity of a manual configuration of the service to be deployed (e.g., have a new robotic arm performing the necessary tasks automatically right after it authenticates in the factory floor). The latter refers to the AI/ML management/orchestration of network and computational devices (servers or IoT devices with storage or processing capabilities) to host the services running at the factory floor premises. For example, an AI/ML model can decide if that the predictive maintenance model is hosted at a server within the Edge premises of a factory floor, and allocate the necessary bandwidth resources to interconnect the robotic arm to the server.

With the help of Figure 7, in the following sections we explain the components involved in the AI-based orchestration and zero-touch deployment.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

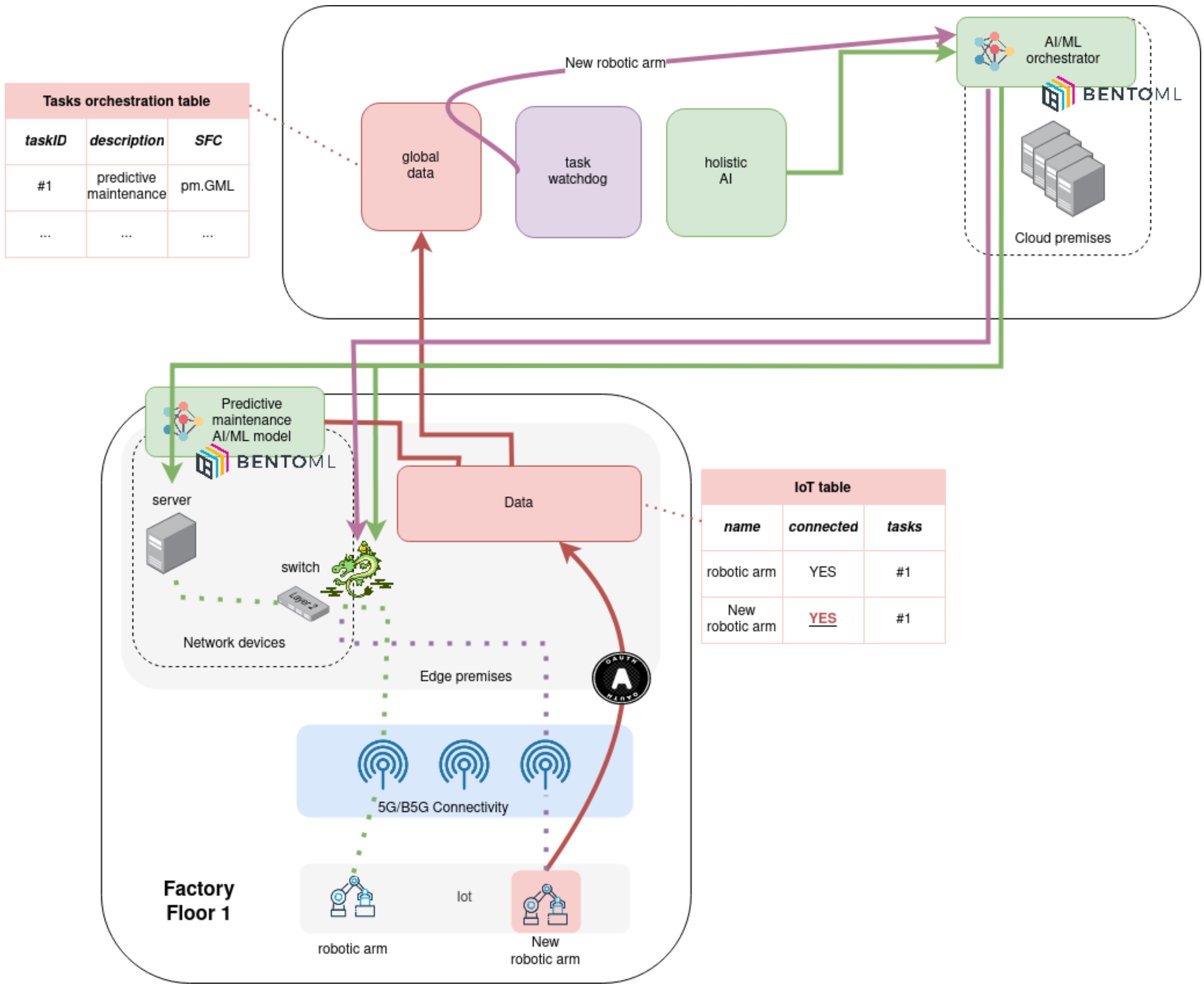

FIGURE 7: AUTOMATION INTERACTIONS

5.1. AI-based orchestration

For the initial system architecture we consider that the AI-based orchestration resides in the central cloud entity, i.e., the orchestration of industrial services runs in the cloud and has a global view of the factory floors within the pool.

The global Data entity keeps track of a task orchestration table that specifies the Service Function Chain (SFC) of the deployed industrial service. The SFC is a graph with each node representing a Virtual Network Function (VNF) – as the predictive maintenance AI/ML model –, that interconnects with other VNFs or Physical Network Functions (PNFs) as a robotic arm. The SFC graph also holds the bandwidth and computing requirements of the service to deploy. For instance, the predictive maintenance SFC could be a tree graph of depth one with each leaf being the robotic arm, and the root being the AI/ML predictive maintenance model.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

Each orchestrator task is passed to the holistic AI to decide the deployment of the associated task SFC, e.g., the predictive maintenance entry is fed to a pre-trained AI/ML orchestrator that resides in the holistic AI. The AI/ML orchestrator was trained following the procedure of Section 3, but rather than using data from the metrics tables, it uses data coming from the network devices table, so the AI/ML is aware of the available network resources. A possible approach is to define a reinforcement learning AI/ML model that receives all the tasks and learns how to orchestrate them in the training stage.

Once the AI/ML orchestrator is trained and deployed at the cloud as a BentoML API, each task orchestration entry is fed to the model to deploy the related task SFC, e.g., the predictive maintenance. Figure 7 illustrates in green arrows the interactions between the BentoML, the Edge server, and the RYU SDN controller to deploy the Predictive maintenance AI/ML model (see the server within the factory floor), and allocate the network resources to interconnect it with the robotic arm (see the dotted green lines).

5.2. Zero-touch deployment

Once an AI/ML orchestrator has been trained (see Figure 7), the BentoML model receives all the entries of the orchestration tasks and deploys all the tasks, e.g., the predictive maintenance task in Figure 7 example. All the orchestration procedure is performed autonomously without human intervention, for the orchestration tasks table contains all the SFCs to deploy.

Therefore, the deployment of the AI/ML tasks are autonomously deployed by the system architecture. However, it is yet to define what happens when a new device authenticates in the factory floor; Figure 7 illustrates such procedure. To that aim we define the task watchdog, an entity that periodically monitors the devices authenticated and recorded at the global data entity, to request orchestration updates for each task. In the following we detail such procedure.

First of all, the New robotic arm challenges and Oauth authentication to register at the factory floor. The data of the new device is exchanged with the global data entity of the central cloud entity, and the latter updates its IoT table. Second, the task watchdog notices that a new robotic arm has been authenticated and checks which tasks it has associated. In the case of Figure 7, the New robotic arm has associated the predictive maintenance task. Third, the task watchdog informs the AI/ML orchestrator that a New robotic arm should be included in the predictive maintenance task. Namely, the AI/ML orchestrator receives the predictive maintenance SFC and realizes that it has to allocate traffic resources to interconnect the New robotic arm with the AI/ML predictive maintenance running at the factory floor Edge premises. Specifically, the AI/ML orchestrator instructs the RYU controller at the factory floor to create new routes and allocate bandwidth to interconnect the AI/ML predictive maintenance to the 5G/B5G RUs and the New robotic arm (purple dotted lines in Figure 7).

The aforementioned interactions allow, with the help of the task watchdog, to trigger orchestration updates upon the authentication of new devices, hence, leading to a zero-touch deployment.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

6. Summary and Conclusions

This deliverable specifies the overall system architecture proposed to assess AI/ML automation in connected industry. First, the deliverable specifies the overall architecture considering Edge premises within each factory floor, and a central Cloud premise to gather and process the information of those factory floors belonging to a factory pool. Each factory floor has its own Data and AI entities to gather and process metrics of the devices that belong to the production pipeline.

The proposed architecture ensures the data privacy and collaborative training of AI/ML models to boost up the performance, and automates the registry and automatic integration of new devices to tasks within the production pipeline. Additionally, the architecture supports maintenance operations to ease the tracking and surveillance of the devices' status at the factory floor.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D

7. Referencias

Applications, O. N. (2022). *Khasa Gillani, Jorge Martín-Pérez, Milan Groshev, Antonio de la Oliva, Carlos J. Bernardos, Robert Gazda.* arXiv.

Milan Groshev, J. M.-P. (2022). FoReCo: a forecast-based recovery mechanism for real-time remote control of robotic manipulators. *IEEE Transactions on Network and Service Management*, 12.

Milan Groshev, J. S.-P. (2022). Demo: FoReCo – a forecast-based recovery mechanism for real-time remote control of robotic manipulators. *SIGCOMM '22: Proceedings of the SIGCOMM '22 Poster and Demo Sessions* (pág. 2). Amsterdam: ACM.