Robust Learning and Reasoning for Complex Event Forecasting

Project Acronym:	EVENFLOW
Grant Agreement number:	101070430 (HORIZON-CL4-2021-HUMAN-01-01 – Research and Innovation Action)
Project Full Title:	Robust Learning and Reasoning for Complex Event Forecasting

DELIVERABLE

D3.1 – Data handling, Requirements Analysis & Scenario Definition

Dissemination level:	PU - Public, fully open
Type of deliverable:	R - Document, report
Contractual date of delivery:	31 March 2023
Deliverable leader:	BSC
Status - version, date:	Final – v5.0, 2023-03-28
Keywords:	data governance, requirements, use cases

Funded by the European Union

This document is part of a project that is funded by the European Union under the Horizon Europe agreement No 101070430. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Commission. The document is the property of the EVENFLOW project and should not be distributed or reproduced without prior approval. Find us at www.evenflow-project.eu.

Executive Summary

This document discusses data handling, requirements analysis, and scenario definition in the context of the three use cases of EVENFLOW: Industry 4.0, Personalized Medicine, and Infrastructure Life Cycle Assessment. These aspects are crucial to ensuring that the objectives of the use cases are met, and the needs of the stakeholders are fulfilled.

In the Industry 4.0 use case, the main objective is to ensure the autonomous transportation of sensitive cargos in a factory setting by implementing efficient path planning and forecasting functionalities. The Personalized Medicine use case involves the utilization of latent representation learning models to scrutinize cancer progression data, which in turn facilitates molecular characterization. The Infrastructure Life Cycle Assessment use case is aimed at optimizing equipment efficiency, reducing waste, and enabling predictive maintenance in the manufacturing industry, with a specific focus on identifying leakage points in water distribution and critical infrastructure.

In all three use cases, the process of data handling encompasses the collection, processing, and storage of information derived from the use cases. Equally important is requirements analysis, which involves the vital task of gathering and documenting the use case requirements, ensuring that they are understood by all members of the consortium. Furthermore, scenario definition is an essential element that involves anticipating potential scenarios or situations in the domains of the use cases. Its purpose is to mitigate the impact of unexpected events and ensure that the required quality standards are met.

Deliverable leader:	Davide Cirillo (BSC)
Contributors:	Guillermo Prol Castelo (BSC) Benjamin Blumhofer (DFKI) Karim Ladjeri (EKSO) Ioannis Christou, Evangelos Kolias (INTRA)
Reviewers:	Nikos Katzouris (NCSR), Karim Ladjeri (EKSO)
Approved by:	Athanasios Poulakidas, Dimitrios Liparas (INTRA)

Document History			
Version	**Date**	**Contributor(s)**	**Description**
0.1	2023-02-04	BSC	ToC preparation
0.2	2023-03-01	BSC, INTRA, DFKI	Initial information reported
0.3	2023-03-07	BSC	First draft completed
0.4	2023-03-14	INTRA	Rearranged Chapter 4, added new content, modified content
0.5	2023-03-28	BSC	Final revision for QA after review
1.0	2023-03-31	INTRA	QA and final version for submission

Table of Contents

Executive Summary....2
Table of Contents....4
Table of Figures....5
List of Tables....5
Acronyms and Abbreviations....6
1 Introduction....7
1.1 Project Information....7
1.2 Document Scope....8
1.3 Document Structure....8
2 Scenario definition....9
2.1 EVENFLOW Use Case Scenarios Summary....9
2.2 Industry 4.0....9
2.3 Personalized Medicine....10
2.4 Infrastructure Life Cycle Assessment....11
3 Requirements analysis....12
3.1 EVENFLOW Use Cases Requirements Analysis Summary....12
3.2 Industry 4.0....12
3.3 Personalized Medicine....12
3.4 Infrastructure Life Cycle Assessment....13
4 Data handling....14
4.1 EVENFLOW Use Cases Data Handling Summary....14
4.2 Data Schemas per Use Case....14
4.2.1 Industry 4.0....14
4.2.2 Personalized Medicine....15
4.2.3 Infrastructure Life-Cycle Assessment....15
4.3 Streaming Data Handling....16
4.4 Database Data Handling....16

Table of Figures

Figure 1. Navigation Concept.....9

List of Tables

Table 1: EVENFLOW Consortium. 7
Table 2. Minimum required data from DFKI partner.....12

Acronyms and Abbreviations

Acronym/ Abbreviation	Title
AGV	Autonomous Guided Vehicles
RGBD-Camera	Red Green Blue Depth Camera
IMU	Inertial Measurement Unit
EKF	Extended Kalman Filter
TCGA	The Cancer Genome Atlas
AI	Artificial Intelligence
VAE	Variational AutoEncoder
CBM	Condition-Based Maintenance
PdM	Predictive Maintenance
LCA	Lifecycle Assessment
OEE	Overall Equipment Efficiency
DL	Deep Learning
ID	Identifier
HPC	High-Performance Computing
PFlops/s	PetaFlops/seconds
CSV	Comma Separated Value
ROS	Robot Operating System
Rot	Rotation
EUDAT	European Data Infrastructure
RDBMS	Relational Database Management System

1 Introduction

1.1 Project Information

The EVENFLOW project aims to develop hybrid learning techniques for complex event forecasting, which will combine deep learning with logic-based learning and reasoning into neuro-symbolic forecasting models. The envisioned methods combine neural representation learning techniques, powerful symbolic learning and reasoning tools, to synthesise high-level, interpretable patterns of critical situations to be forecast.

Crucial in the EVENFLOW approach is the online nature of the learning methods, which makes them applicable to evolving data flows and allows to utilise rich domain knowledge that is becoming available progressively. To deal with the brittleness of neural predictors and the high volume/velocity of temporal data flows, the EVENFLOW techniques rely on novel, formal verification techniques for machine learning, in addition to a suite of scalability algorithms for federated training and incremental model construction. The learnt forecasters will be interpretable and scalable, allowing for fully explainable insights, delivered in a timely fashion and enabling proactive decision making.

EVENFLOW is evaluated on three challenging use cases related to (1) oncological forecasting in precision medicine, (2) safe and efficient behaviour of autonomous transportation robots in smart factories and (3) reliable life cycle assessment of critical infrastructure.

Expected impact:

- New scientific horizons in integrating machine learning and machine reasoning, neural, statistical and symbolic AI.
- Breakthroughs in verification, interpretability and scalability of neuro-symbolic learning systems.
- Interpretable, verifiable and scalable ML-based proactive analytics and decision-making for humans-in-the-loop and autonomous systems alike.
- Robust, resilient solutions in critical sectors of science and industry.
- Accurate and timely forecasting in vertical sectors (healthcare, Industry 4.0, critical infrastructure monitoring).
- Novel FAIR datasets for scientific research.
- Novel resources and approaches for verifiable, interpretable, scalable and knowledge-aware machine learning.

Table 1: EVENFLOW Consortium.

Number[1]	Name	Country	Short name
1 (CO)	NETCOMPANY-INTRASOFT	Belgium	**INTRA**
1.1 (AE)	NETCOMPANY-INTRASOFT SA	Luxembourg	**INTRA-LU**
2	NATIONAL CENTER FOR SCIENTIFIC RESEARCH "DEMOKRITOS"	Greece	**NCSR**

[1] CO: Coordinator. AE: Affiliated Entity. AP: Associated Partner.

Number[1]	Name	Country	Short name
3	ATHINA-EREVNITIKO KENTRO KAINOTOMIAS STIS TECHNOLOGIES TIS PLIROFORIAS, TON EPIKOINONION KAI TIS GNOSIS	Greece	**ARC**
4	BARCELONA SUPERCOMPUTING CENTER-CENTRO NACIONAL DE SUPERCOMPUTACION	Spain	**BSC**
5	DEUTSCHES FORSCHUNGSZENTRUM FUR KUNSTLICHE INTELLIGENZ GMBH	Germany	**DFKI**
6	EKSO SRL	Italy	**EKSO**
7 (AP)	IMPERIAL COLLEGE OF SCIENCE TECHNOLOGY AND MEDICINE	United Kingdom	**ICL**

1.2 Document Scope

The scope of this document entails the description of data handling, requirements analysis, and scenario definition in the context of the EVEFLOW use cases: Industry 4.0, Personalized Medicine, and Infrastructure Life Cycle Assessment. These aspects are essential to ensure that the objectives of the project are completed successfully and meets the needs of the use case stakeholders.

Data handling involves the collection, processing, and storage of data. Requirements analysis involves gathering, documenting, and validating the use case requirements and ensuring that they are clearly understood by the whole consortium. Scenario definition involves the creation of potential scenarios or situations that may occur in the domains of applications of the use cases, helping minimize the impact of unexpected events and to ensure the required quality standards.

1.3 Document Structure

This document is comprised of the following chapters:

Chapter 1 presents an overview of the scenario and the challenges faced in use cases.

Chapter 2 presents the requirements for the implementation of forecasting functionalities in the use cases.

Chapter 3 presents the handling and pre-processing the data used in the use cases.

2 Scenario definition

2.1 EVENFLOW Use Case Scenarios Summary

This chapter is concerned with presenting the use case scenarios. The Industry 4.0 use case (Section 2.2) focuses on autonomous transportation of sensitive cargos in a factory environment in a timely manner. By integrating forecasting functionalities into the cost maps used for robot navigation, the robot can efficiently plan its path, anticipating the situations ahead and avoiding collisions through replanning. The Personalized Medicine use case (Section 2.3) involves latent representation learning models of cancer progression data. Tracing probabilistic trajectories of individual data points in the latent space enables forecasting undesirable outcomes and facilitates molecular characterization. The Infrastructure Life Cycle Assessment use case (Section 2.4) encompasses the implementation of AI in the manufacturing industry to optimize the overall equipment efficiency, reduce waste, and enable predictive maintenance, focusing on smart pipes in water distribution and critical infrastructure to identify leakage points and reduce management costs.

2.2 Industry 4.0

In Industry 4.0, autonomous guided Vehicles (AGV) take a key role in production processes by transporting sensitive cargo. The timeliness of those transports a crucial to keeping the entire factory process on track. This requires a robust and time-optimal delivery of the goods. The scheme shows the general structure of the implemented navigation pipeline, which is based on Nav2. The robot uses a 3D Lidar, a RGBD-Camera, an Inertial Measurement Unit (IMU), and wheel encoders as sensors. It is steered by three independent mecanum wheels, allowing simultaneous movement in x, y, and the steering angle theta.

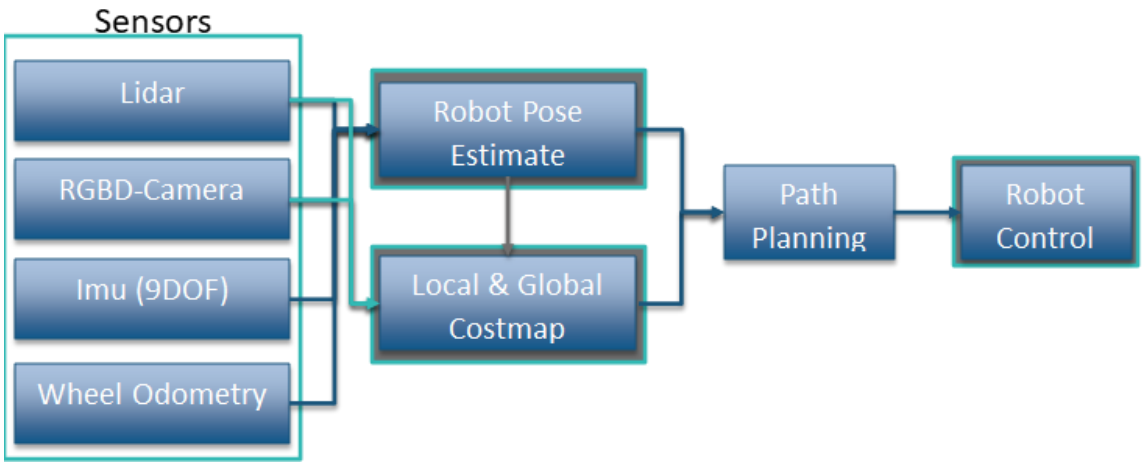

Figure 1. Navigation Concept.

To determine the pose of the robot, extended Kalman filter (EKF) is used. IMU measurements and wheel odometry are fed in the EKF. In addition, visual and lidar-based pose estimation is used as input to the EKF. For navigation, it is highly relevant to know the surrounding of the robot. The current state of the art simplifies the environment to 2D Gridmaps modelling the occupancy of a certain gridpoint based on costs. These maps are called costmaps. The

costmaps are implemented in two different layers, the global- and local costmap. The global costmap usually contains all the static objects and is used for planning from start to finish by avoiding static objects. Additionally, a local map that models only the area close to the robot is implemented to avoid dynamic objects. The local planner plans trajectories based on the global plan to avoid dynamic obstacles. The planned path is then fed into a controller to follow the path with the robot.

To navigate complex factory environments, robots use costmaps containing information about space occupancy. However, these costmaps do not incorporate semantic knowledge about objects or forecasting functionalities. The goal of this scenario is to use the forecasting availability of EVENFLOW to create a costmap including information on possible future trajectories of objects based on the past movement patterns. Early detection of such events will allow for replanning and avoiding collisions far ahead in the path planning of the robot. Avoiding scenarios with humans in the vicinity of robots leads to less replanning efforts and a faster delivery of goods.

To create the costmap with forecasting functionalities, EVENFLOW will be used. The task of the forecasting system is to forecast trajectories of objects based on known patterns. The trained model will be integrated with the robot's navigation system to provide forecasting information. The information will be incorporated into the costmap, allowing the robot to replan its path far ahead of the situation, leading to a more optimal path.

Processed information about detected objects in the environment will be used as input information for the EVENFLOW approach. The information will be obtained through sensors such as cameras and lidars. The information will include additional information about the detected object, e.g., the class and directional velocity of the object. The processed information will be incorporated into the costmap, allowing the robot to avoid potentially risky situations.

The use of forecasting availability and processed information about detected objects in the environment can significantly improve the efficiency and safety of the delivery process. Using the costmap with forecasting functionalities will allow the robot to navigate while avoiding potentially risky situations and predicting the future position of detected objects. Integrating EVENFLOW and processed information will lead to a more efficient and safer delivery process, making autonomous robots even more important in various industries.

2.3 Personalized Medicine

Although understanding the molecular changes that occur during cancer progression is critical for developing effective treatment strategies, the current scarcity of molecular data of intermediate stages of tumour progression is a significant limitation in cancer research. Indeed, obtaining samples that comprehensively cover the entire long-term process of cancer progression can be challenging due to difficulties and limitations of imaging and histopathological techniques, the invasiveness of certain procedures, and patient-related issues like delays in seeking medical attention. As a result, the current landscape of available molecular data of cancer progression is extremely fragmented and it often consists of

measurements that are taken at different time points during the progression of cancer in different groups of people.

An example is the molecular data of cancer staging that can be found in The Cancer Genome Atlas (TCGA). TCGA collects large amounts of molecular data from cancer patients, including DNA sequencing data, RNA expression data, and epigenetic profiling data, among others. Given the pressing need for new technologies that can help characterize or infer molecular changes that occur during breast cancer progression, BSC aims to develop artificial intelligence (AI) approaches with the goal of effectively forecasting cancer progression and facilitating the identification of effective treatment strategies.

The approach that is employed consists in training and evaluating a Variational Autoencoder (VAE) using TCGA breast cancer RNA expression data. The resulting lower-dimensional representation of such data, or latent space, is then explored to trace probabilistic trajectories of individual data points as they move from one cancer stage to another. These trajectories enable the exploration of possible outcomes and future directions facilitating the development of forecasting models.

2.4 Infrastructure Life Cycle Assessment

The use of AI allows the development of added-value use cases in the manufacturing shopfloor and the manufacturing chain, including intelligent asset management, condition-based maintenance (CBM), predictive maintenance (PdM) and lifecycle assessment (LCA) for key assets. The latter use cases enable manufacturers and other users of assets to avoid the catastrophic consequences of unplanned downtimes, boost the optimization of the Overall Equipment Efficiency (OEE), and contribute essentially to waste reduction in-line with the twin transition agenda of most industrial organizations.

In parallel, based on experience, water losses in the underground potable and irrigation networks exceed in many cases 50%. Beyond the significant ecological footprint, the possibility of accurately identifying the point of leakage determines a substantial reduction of these management/maintenance costs, improving the efficiency of networks and infrastructures. Moreover, the availability of information in (near) real time on the proper operation of factory plants and critical infrastructures becomes of great interest in the case of industrial applications where safety aspects are of primary importance. Applying machine learning techniques is a challenging task.

The use case scenario will provide the means for gathering and analysing digital data about the conditions of the pipes towards optimizing their lifecycle management including their maintenance, services, repair, and other lifecycle management processes. In this direction, the project will develop an innovative digitally enabled lifecycle assessment tool for pipes, which will provide the means for optimizing both economic and environmental parameters, while providing recommendations for creating new pipes and resolving relevant trade-offs. The use case aims at designing and developing an innovative predictive maintenance application based on Deep Learning (DL) techniques and manufacturing data related to smart pipes for water distribution/irrigation and critical infrastructure.

3 Requirements analysis

3.1 EVENFLOW Use Cases Requirements Analysis Summary

This chapter is focused on presenting the main requirements of the use cases. The Industry 4.0 use case (Section 3.2) requires real-time navigation with efficient forecasting capabilities to enable better replanning. The Personalized Medicine use case (Section 0) necessitates access to public molecular data of cancer progression and appropriate computational resources for model development. The Infrastructure Life Cycle Assessment use case (Section 3.4) requires regular data gathering of pipe conditions and the availability of suitable computational resources for implementing models that can analyse any anomalies referred to a base case scenario.

3.2 Industry 4.0

The Industry 4.0 Scenarios have multiple requirements in different areas. The first requirement comes from the nature of the problem. As navigation needs to run in real-time for the forecasting to be useful there is a need for the forecasting algorithm to run with a desired update rate faster than 5Hz on the robot's computation device. To be able to include EVENFLOW forecasting in costmaps in it is necessary to receive information on the future behaviour of the object from the forecasting algorithm. This can be information on the involved object identifier (ID) and additional information like the location and time occurrence of the event. The more information is available on the forecasted event the better replanning based on costmap is possible.

Table 2. Minimum required data from DFKI partner.

Information per detected object	**Provider**
Object ID	DFKI
Position	DFKI
Velocity	DFKI
Array of possible trajectories (x, y, time)	NCSR
Probability per trajectory	NCSR

3.3 Personalized Medicine

The requirements of the Personalized Medicine use case consist in the availability of public molecular data of cancer progression and the availability of suitable computational resources for the implementation of models for generating probabilistic trajectories and forecast relevant events.

The data that is currently used is RNA expression of breast cancer from TCGA, which contains information on 1084 female subjects belonging to four distinct cancer stages in varying proportions. Breast cancer stages are classified based on the size of the tumour and the extend of spread:

- Stage I involves a small tumour, typically less than 2 cm in diameter, that is confined to the breast tissue.
- Stage II, the tumour may be larger, up to 5 cm in diameter, and may have spread to nearby lymph nodes.
- Stage III, the tumour is larger and may involve nearby tissues, such as the chest wall or skin, and may have spread to multiple lymph nodes.
- Stage IV represents the most advanced stage of breast cancer, where the cancer has spread to distant parts of the body, such as the bones, liver, or lungs.

This data is publicly available, and it can be accessed at the Genomic Data Commons Data Portal: https://portal.gdc.cancer.gov/. The study of TCGA liver cancer staging is under consideration for further applications in the context of the project.

The use case has access to world-class computational resources. The BSC is a national research and supercomputing centre in Spain, specialised in high-performance computing (HPC), which manages MareNostrum, one of the most powerful supercomputers in Europe. The centre has several supercomputing clusters, among which are two outstanding ones. MareNostrum4: It is the most powerful supercomputer in Spain, with a disk storage capacity of 14 Petabytes and is connected to the Big Data infrastructures of BSC, which have a total capacity of 24.6 Petabytes, connected to the European research centres and universities through the RedIris and Geant networks; MareNostrum5: a pre-exascale heterogeneous supercomputer that is 18 times more than the current MareNostrum4. It will have a target performance above 200 PFlops/s in two major working partitions, one based on general-purpose nodes and the other based on accelerated nodes.

3.4 Infrastructure Life Cycle Assessment

The requirements of the use case consist in the periodic and regular data gathering of the pipe conditions and the availability of suitable computational resources for the implementation of models for analysing any anomality referred to a base case scenario.

The pilot will build and validate policy models that will reduce physical and commercial water losses and maintenance cost, while helping the management entity (private or public entity) establish effective maintenance and repair schedules/policies. The project's tools will recommend, simulate, and identify policies associated with pipe/water data measurement. Based on the results and experience, the AI pilot can suggest upscaled wider monitored network sections.

4 Data handling

4.1 EVENFLOW Use Cases Data Handling Summary

Currently, all use cases in the project seem to provide data in the form of Comma Separated Value (CSV) files that are ASCII text files representing records that contain attribute values separated by a standard separator symbol (usually, the comma “,” character). In the case of time-series data, usually there exists one column in the CSV file that represents a timestamp, or at least the order (position) of the record in the sequence within which it belongs. Such data can be represented in their raw format (flat text files) or they can be processed and entered either in the chosen database for handling time-series (InfluxDB) or in the Kafka distributed message bus as data in one or more topics. In the latter case, the data can be stored as a link to the external file system where they live, or they can be stored themselves as (potentially huge) strings in a topic, which however is against the design assumptions of Kafka, and any other message bus in general, or finally, the data in each CSV file can be “decomposed” and stored as chunks, or rows themselves in possibly different topics, depending on the need of the consumers of this data. In many cases, the data in CSV form never enter either of the data management solutions, and instead remain in certain folders that can be accessed by various protocols (e.g. FTP protocol).

4.2 Data Schemas per Use Case

4.2.1 Industry 4.0

Generated data in the DFKI scenario come in the form of time series, commonly as measurements from robot sensors. The process of data generation and data handling do not make use of Kafka, nor InfluxDB. We record our data and store them as ROS bags (a file format containing ROS messages from topics of interest e.g., position, velocity…), then convert all the recordings into .csv format. Moreover, the data are offered for further usage by involved EVENFLOW partners via ownCloud.

The data collection is done in the following sensors: Lidar, camera, IMU, wheel odometry, and magnetometer.

In general, the dataset contains of .csv files and a detailed description of the recorded scenario and signals. Each .csv file contains:

Headers (first row) of all .csv files contain descriptive information on the quantities being measured.

Possible content of the datasets can be:

- In the “imu.csv” file, first column corresponds to timestamp (Unix epoch). Columns 2,3, and 4 correspond to angular velocities along x, y, and z axis. Columns 5,6, and 7 correspond to linear accelerations along x, y, and z axis, and the remainder of the columns correspond to the orientation of IMU. Content of all columns is of numerical values.
- In the “odmetry.csv” file, first column corresponds to timestamp (Unix epoch). The rest of the columns correspond to robot pose along x, y and z axis, as well as

orientation of the robot using quaternion representation (last 4 columns). Position and orientation are calculated from wheel encoder. Content of all columns is of numerical values.

- In the "detected_obj.csv" file, first column corresponds to timestamp (Unix epoch). The second column contains the number of detected objects. From the third column on there are x columns per detected object with the following information: ID, Class, accuracy of the detection, Pose (x, y, z), Rot(x, y, z), linear velocity.

4.2.2 Personalized Medicine

The RNA expression data of TCGA breast cancer was downloaded from the corresponding source (see Section 3.3) and stored locally at BSC. Prior to model implementation, the data has been pre-processed and thoroughly analysed using statistical techniques to summarize and visualize it.

The VAE model has been implemented in PyTorch and trained and tested locally at BSC. The obtained patient trajectories are stored in comma separated values (.csv) files and shared with NCSR. In the original .csv files we can find one column per gene and one row per patient, and the corresponding gene expression, a numerical value. Meanwhile, the data processed through the VAE keeps the same patients in the rows, but the columns are now a linear combination of genes, greatly reducing this dimension. The data sharing platform used is B2DROP (https://eudat.eu/services/userdoc/b2drop) that is the official BSC service for data sharing outside the institution. The focus of B2DROP is to facilitate collaboration and file sharing among scholars and researchers who handle significant volumes of data. As a result, it has emerged as the preferred solution for EUDAT (European Data Infrastructure), which is a pan-European data infrastructure that offers data services to the research community.

4.2.3 Infrastructure Life-Cycle Assessment

The Infrastructure Life-Cycle Assessment data (Pozzalo data) are time-series data that represent experimental measurements from vibration sensors (number and location to be determined) attached to certain water pipes having a number of taps that can be open or closed, with the objective to be able to detect leakages in these pipes, and eventually even be able to localize leakage spots and other phenomena: flow, pressure, etc.

The datasets come in a set of file-system folders, appropriately named using the convention

"<sensor-name>_Scenario<X>_<timestamp>" and

"<sensor-name>_Leakage_<timestamp>".

The scenarios correspond to a series of events that correspond to various simulations of situations where some tap may be open (emulating a leakage scenario), or not; they may also correspond to a transient state where an initial empty set of pipes start filling in with flowing water etc. Within each top-level folder therefore, there exist further sub-folders with names that follow the convention "<lag-in-seconds>_<event-name>" where <event-name> can be for example "Tap1Open" or "AllTapsClosed" etc.

Within each sub-folder, there is a single CSV file whose records follow the format:

<timestamp>, <sensor-value>

There is always a top (header) row that reads:

"Time,Modulus"

In fact, for experiments where read-outs from more than 1 sensor are taken, it is possible that the exact time of the measurement between the two sensors does not coincide; this creates the need for an algorithm that matches in an optimal sense the readings of the two (or more) sensors, for a classification/regression model to later use together.

Sampling frequency, strongly influencing DB dimensions, shall be calibrated carefully to reach max efficiency. Only anomalies referred to a base case scenario shall be evidenced, analysed and examined.

4.3 Streaming Data Handling

So far, all use cases we have seen above concern time-series (sequential data). However, this does not mean that the data are streaming data, though it is possible to convert them into streams when this is beneficial for the performance or, more basically, for the functionalities of the system. Both Apache Kafka, and InfluxDB are designed to work mainly with streaming data, but where the focus of Apache Kafka is to act as a broker that allows many consumers to concurrently read data that are sent non-stop from various producers, the focus of InfluxDB is to act as a database management system, allowing for the long-term storage and indexing of streaming data, of continuous queries using a NoSQL language, and of course, monitoring the status of the incoming data sources and raising alerts if some data source seems to be disrupted for some reason. Given the current needs of the project, it appears that sending appropriate subsets of some of the CSV files in Kafka is currently more useful to the consortium.

4.4 Database Data Handling

The InfluxDB time-series database comes with its own query language (InfluxQL) that allows complex querying with syntax and functionalities that sometimes resemble PL/SQL more than the standard declarative SQL language. We are currently in the process of deploying InfluxDB with the help of the helm package management tool for Kubernetes clusters.

However, for non-time-series, non-sequential data the best tool to use is a relational database management system. If the need arises, we shall be using MySQL (or MariaDB) as it represents the most popular free (for non-commercial purposes) RDBMS.