D1.1

Application Cases and Architecture Requirements

IT4I@VSB

Public

Funded by
the European Union

Deliverable Properties

Version	v1.0
Dissemination level	PU
Deliverable type	R
Work package	WP1
Task	T1.1
Due date	30 April, 2023
Submission date	30 April, 2023
Deliverable lead	IT4I@VSB
Authors Name	Mohamad Hayek (BADW-LRZ), Martin Golasowski (IT4I@VSB), Pinar Karagosz (METU), David Číž (IT4I@VSB), Jan Zahradník (VALEO), David Hurych (VALEO), Piyush Harsh (TERRAVIEW), Jérôme Freani (ALTRNATIV)
Reviewers	Antonín Vobecký (CVUT), Ismail Hakki Toroslu (METU)

Abstract

This document is the first deliverable of the EXA4MIND project. It contains requirements provided by the project's application-case work packages WP4-WP6 and their mapping to the EXA4MIND Platform features. The document is roughly divided into two parts. The first part is containing a unified description of each application case and its requirements. The second half of the document contains the mapping of the requirements to the technical features of the EXA4MIND Platform and the project objectives provided by the technical work packages WP1-WP3.

Keywords

requirements; application cases; technical questionnaire

Document Revision History

Version	Date	Description of change	Contributor(s)
v0.1	22 March, 2023	Table of Content	Mohamad Hayek (BADW-LRZ), Martin Golasowski (IT4I@VSB)
v0.2	7 April, 2023	Integrated contribution from application cases	David Číž (IT4I@VSB), Jan Zahradník (VALEO), David Hurych (VALEO), Piyush Harsh (TERRAVIEW), Jérôme Freani (ALTRNATIV)
v0.3	14 April, 2023	Outcomes from technical questionnaire integrated	Mohamad Hayek (BADW-LRZ), Martin Golasowski (IT4I@VSB), Pinar Karagosz (METU)
v0.4	20 April, 2023	Pre-final version for internal review	Martin Golasowski (IT4I@VSB)
v1.0	30 April, 2023	Final check done by coordinator and science/co-design coordinator	Kateřina Slaninová (IT4I@VSB), Jan Martinovič (IT4I@VSB), Stephan Hachinger (BADW-LRZ)

Disclaimer

Copyright Notice

Dissemination levels

PU	Public	Public, fully open, e.g. website
SEN	Sensitive	Confidential to EXA4MIND project and Commission Services

Deliverable types

R	document, report (excluding periodic and final reports)
DEM	document, report (excluding periodic and final reports)
DEC	websites, patent filings, press and media actions, videos, etc.
OTHER	software, technical diagrams, etc.

Table of Contents

List of Figures

List of Tables

Glossary

ADAS	Autonomous driver-assistance system
ADC	Analog to Digital
AI	Artificial Intelligence
AL	Active Learning
AoI	Area of Interest
AQIS	Advanced Query and Indexing System
CAN	Controller Area Network
DoA	Description of Action
ECU	Electronic Control Unit
EDD	Extreme Data Database
EO	Earth Observation
FF	Force Field
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
HW	Hardware
HIL	Hardware-in-the-Loop
HPC	High Performance Computing
IAM	Identity and Access Management
IMU	Innertial Measurement Unit
KPI	Key Performance Indicator
LAN	Local Area Network
LiDAR	Light Detection and Ranging
MD	Molecular Dynamics
ML	Machine Learning
ODD	Operational Design Domains
RTK	Real-Time Kinematics
SDG	Strategy Development Goal
SIL	Software-in-the-Loop
SW	Software
USGS	United States Geological Survey

WAN	Wide Area Network
WP	Work Package

Table of Partners

Short Name	Partner
IT4I@VSB	IT4Innovations at VSB – Technical University of Ostrava (IT4Innovations) https://www.it4i.cz/en
BADW-LRZ	Leibniz Supercomputing Centre (LRZ) of the Bavarian Academy of Sciences and Humanities https://www.lrz.de/english/
METU	Middle East Technical University (METU) https://www.metu.edu.tr/
AUSTRALO	AUSTRALO Interinnov Marketing Lab https://www.australo.org/
EURAXENT	Euraxent https://www.linkedin.com/company/euraxent/
CVUT	Czech Institute of Informatics, Robotics and Cybernetics at the Czech Technical University (CIIRC CTU) https://www.ciirc.cvut.cz/
VALEO	Valeo https://www.valeo.com/en/valeo-ai/
TERRAVIEW	Terraview https://www.terraview.co/
ALTRNATIV	Altrnativ https://altrnativ.com/?lang=en

1 Executive Summary

This document is the first deliverable of the EXA4MIND project. It contains a description of each application case provided by work packages WP4-WP6, results of the analysis performed by the technical work packages WP2-WP3 and a list of initial requirements on the EXA4MIND Platform.

The document is roughly divided into two parts. The first part contains a unified description of each application case. The second half of the document contains the mapping of the requirements to technical features of the EXA4MIND Platform and project objectives provided by the technical work packages WP1-WP3. The content of the deliverable is the initial input for the co-design process. The first draft of the architecture due in M9 will be based on these inputs.

2 Application Cases Description

In this section, we provide a structured description of each application case. The cases are described by their respective work packages (WPs). Each case provides its short description, visual representation of the planned workflow of the application, concrete usage scenarios, and stakeholders with their roles. The industrial application case is split into two parts in comparison with the original DoA. The first describes the data handling part of the case, which is then followed by an AI case, which directly uses the data handled in the previous step.

2.1 Scientific Application Case - ADAMS4SIMS (WP4)

2.1.1 Description

Modelling has become essential in many fields ranging from engineering through weather research to chemistry. With advances in HPC and theoretical chemistry, molecular dynamics (MD) simulations witnessed a significant progress (Smith et al. 2017). Having important applications ranging from drug discovery over structure and function relationships to protein/nucleic acid design, MD simulations support tens of thousands of scientific publications. This has been supported by the development of empirical force fields (FFs), which capture many structural features of biologically relevant systems. On the other hand, researchers often face FF artifacts, related to over-parameterization and overfitting of the FF parameters.

Based on the long-term experience of the EXA4MIND team in FF development for nucleic acids (Zgarbová et al. 2011; Kuhrova, Best, et al. 2016; Kuhrova, Mlynsky, et al. 2019; Frohlking et al. 2022), where the popular AMBER framework (Case et al. n.d.) uses our FFs for DNA/RNA simulations, we can state that the era of manual or semi-automatic tuning/development of FF parameters (Šponer et al. 2018) (and concepts, e.g. additional FF terms) is coming to its end, as humans cannot capture the whole complexity of this process. We are certain that the future development must be automated and assisted by Machine Learning (ML) approaches on top of Extreme Data acquired from HPC MD simulations. FF development requires an extensive testing

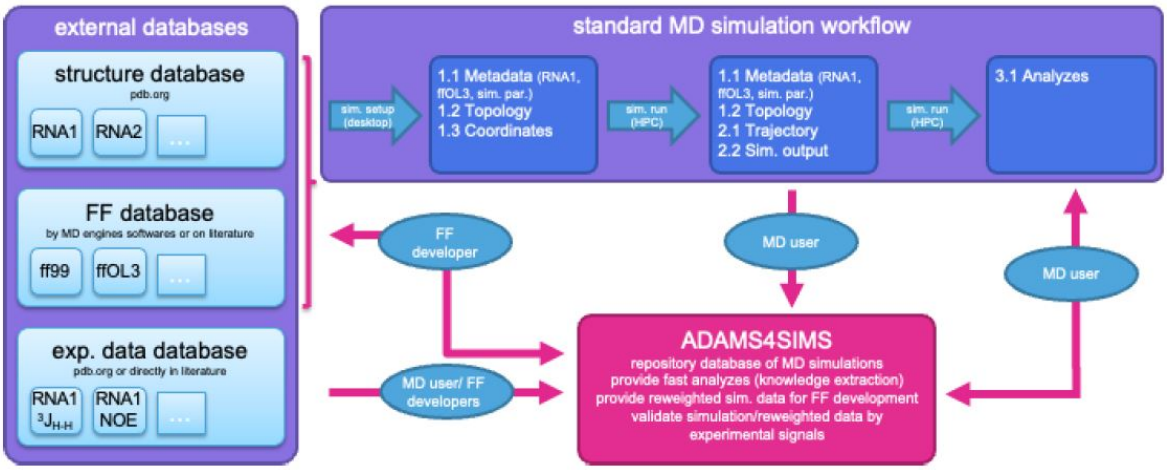

Figure 1: Overview of ADAMS4SIMS application

carried out via massive MD simulations. Each state-of-the-art MD simulation takes weeks to months of HPC computational time. The proposed solution will enable reusing the simulation data and application of reweighing schemes that significantly shorten the requirements from weeks/months to minutes during FF testing. It speeds up the process of FF development and enables an efficient exploitation/mining of the MD simulation data. To facilitate the analysis and visualisation of the simulation results, we propose to use a datacube structure to represent and organise multiple dimensions of the data, such as parameters, systems, and snapshots. Moreover, we envision a query method that enables an efficient and flexible access to any subset of the datacube, based on the user's needs and preferences.

2.1.2 Usage Scenarios

Current state-of-the-art FF development relies on experience-based trial and error approaches, combined with FF parameterizations based on high-level quantum chemical calculations and massive testing. One of the strongest tools of FF development are reweighting methods (Bottaro et al. 2018; Köfinger, Różycki, and Hummer 2019; Shen and Hamelberg 2008), which can re-evaluate the results of MD simulations under the assumption of a modified FF parameterization without repeating the entire simulation. Using these methods, FF parameters are fine-tuned so that the observables are predicted to meet the corresponding experimental values. However, this approach is typically applied to one or at most a few molecular systems Mlynsky et al. 2022, and over-parameterization/overfitting then tends to prevent a general applicability. With massive testing on a wide range of molecular systems, we expect to overcome this problem and generate robust FF reparameterizations. This requires the automation of the tuning process, based on a huge database of simulations and observed parameters for benchmark molecular systems.

The main phases of the application case and the associated workflows are presented in Figure 1.

2.1.3 Elements/Actors/System

There are different types of stakeholders depending on the type of usage scenarios:

- **MD simulation community.** Members of the MD simulation community will be able to upload their MD simulations and process the knowledge extraction (analysing the observables in their simulations) using the tool of ADAMS4SIMS. Members of MD simulation community are also expected to upload corresponding experimental data for benchmarking the simulation outcomes. In such a way, they will feed the MD simulation database including the data coming from the knowledge extraction. These data have to be curated, and the access level of users has to be managed/approved.
- **FF developers.** FF developers will be able to upload new FFs and launch the reweighting process via ADAMS4SIMS GUI. FF developers are assumed to act also as members of MD simulation community including uploading MD simulations, extracting the knowledge (setting analyses) and uploading experimental data.
- **General user.** The general user is expected to view the MD simulations, simulation outcomes and FF performances. They will be able to set new tasks for additional analysis (knowledge extraction data).

2.2 Industry Application Case - AI (WP5)

2.2.1 Description

The AI part of the VALEO application case is about the development of a model toolbox based on ML/AI for recorded ADAS (Autonomous driver-assistance system) data content interpretation (metadata extraction). To this purpose, both publicly available data sets (e.g. Waymo (Sun et al. 2020), SemanticKITTI (Behley et al. 2019), nuScenes Caesar et al. 2020, ONCE (Mao et al. 2021), Cityscapes (Cordts et al. 2016), ACDC (Sakaridis, Dai, and Van Gool 2021), Dark Zurich (Sakaridis, Dai, and Van Gool 2020),) in tens of terabytes in total size, as well as VALEO application case data scaling up to 3PB of raw data, will be used. The meta-data extracted at the end of the project may scale the total amounts of data 2 to 10 times, depending on the storage and computational power limitations.

2.2.2 Usage Scenarios

AI-based usage scenarios include a variety of AI-based use cases as follows:

1. Video Object Detection and Tracking for Data Pre-annotation: This use case aims to integrate large-scale data storage systems and powerful computing infrastructures with a novel automated an data management and effective data staging. Due to the massive amount of data that needs to be processed, this use case provides an automatic annotation system based on machine learning and computer vision to pre-process and pre-annotate the raw data before a manual annotation. The main focus is a video object detection and tracking,

which involves locating objects and tracking them on the raw data without manual annotations. This provides preliminary annotations of object locations and identities to the data before annotators get involved, significantly reducing the time and cost of the data annotation process.

2. Virtual Validation: This use case focuses on the use of generative artificial intelligence for system validation in the automotive industry. The use case will explore the possibility of using various Generative Neural Networks for the controlled creation of image data suitable for assessing the error of tested ADAS, i.e., virtual testing data generation. The goal is to get the ability to generate an arbitrary composition of road users in an image and use it as a challenging testing scenario for ADAS. This will help to assess the performance and overall safety of such a system before its production in millions of cars. The size of the testing set for ADAS is scaling to units of Petabytes and being able to partly or fully replace it with virtual data is a must. Therefore, huge computational resources are essential to be able to train such models that generate image data of great quality in a high resolution.

3. Open-vocabulary Object Detection/Discovery: We aim at developing an open-vocabulary detection framework that (i) can localise and recognise a large number of object classes in imagery inputs, and (ii) can be extended to handle novel or more fine-grained classes with a minimal cost and complexity. To this end, we plan to train a universal detector on existing public data sets for detection and classification. The detector is equipped with the zero-shot power of large vision-language models like CLIP, as such it can be extended to recognise new concepts. The universal detector designed in this use case will serve as the starting backbone for any pre-annotation campaigns coming with different requirements in the future. This case also requires access to large-scale data management and powerful computing infrastructures to satisfy its requirements.

4. Unsupervised Domain Adaptation for LiDAR Point Clouds: In order to achieve high performance, today's neural networks for point cloud semantic segmentation or 3D object detection require large annotated data sets. Several public data sets are available to train these networks: SemanticKITTI (Behley et al. 2019), nuScenes (Caesar et al. 2020), Waymo (Sun et al. 2020), ONCE (Mao et al. 2021), etc. Yet any network trained on one data set performs poorly on the others. One of the reasons is that these data sets are acquired with different LiDARs and deep neural networks do not generalise well across LiDARs. Therefore, we need techniques to avoid labelling new data sets or, at least, reduce the number of new annotations. In this use case, we are specifically interested in unsupervised domain adaptation strategies. In particular, we will consider two types of approaches. First, we will evaluate methods that leverage pre-trained image models and use distillation techniques to transfer their knowledge to point cloud processing networks. These approaches require data sets with synchronised and calibrated LiDAR and cameras, such as in the public data sets KITTI, nuScenes, Waymo and ONCE, but also in the data sets captured internally at VALEO. Second, we will evaluate domain adaptation techniques that work solely on LiDAR data. These methods typically leverage a source data set, that is fully annotated,

and a target data set, that is not annotated at all, and mitigate the distribution shifts between the source and target data sets to boost the performance of a network on the target data set. Finally, depending on the performance of the first and second type of approaches, we will investigate whether such methods bring any benefit, and if they do, which approach is better and if they can be combined together. The techniques considered above leverage self-supervised learning strategies, which do not require any manual annotations but for which the gain appears when trained at scale: on large data sets with large batch sizes, requiring multiple GPUs processing data in parallel.

5. Reducing cost annotation with Active Learning: Many perception tasks are currently tackled using deep learning methods. However, such methods are usually trained in a fully-supervised fashion and therefore require data to be entirely annotated. Annotation costs (and corresponding slow processes) remain a major bottleneck when considering the definition of a new perception problem. One way to mitigate these costs is by the application of Active Learning (AL) methods which allow to select the best data to be annotated. In the context of the EXA4MIND project, we will investigate how such methods can be best applied to our industrial data, which are more redundant and noisy than academic data sets.

6. Self-supervised image representation learning for autonomous driving: This case aims to develop self-supervised learning methods for pre-training neural networks using large volumes of raw/unlabelled autonomous driving images. This pre-training can help to learn useful image representations that can be applied to downstream vision tasks, such as object detection and semantic segmentation, in autonomous driving. Using self-supervised learning, which utilises only the information within an image and does not require human-generated labels, the use case aims to reduce the requirement for large labelled training sets for high-capacity image model architectures, like vision transformers. This can help improve the performance of autonomous driving systems while reducing annotation costs. Ultimately, the goal is to develop self-supervised methods that are able to exploit raw/unlabelled data to enable efficient learning of perception networks that are robust and require limited human-generated labels.

7. Multi-Modal Large-Scale Representation Learning for Driving Scene Understanding: The research aims to train machine learning models to produce useful representations for driving scene understanding via using large-scale multimodal data. This is an important problem, as the annotated data for driving scenes is expensive to obtain, i.e., a single image can take up to three hours of manual labelling. The key open problem in these low-data regimes is low model accuracy in rare or missing situations in the annotated training data or overfitting to the training data distribution, which inherently differs from testing and validation. On the other hand, large amounts of unannotated data with multiple sensor modalities (multiple sensory captures for the same scene and time) are available in automotive setups. In this use case, we will develop a new generation of large-scale multi-modal self-supervised learning techniques to overcome the

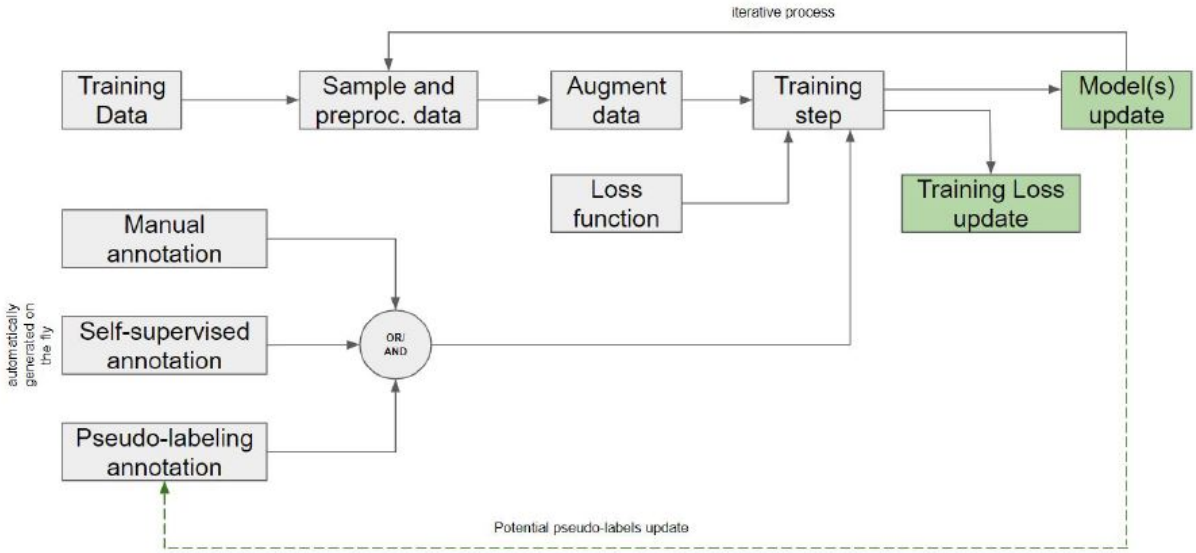

Figure 2: Overview of AI Model Construction in VALEO Application

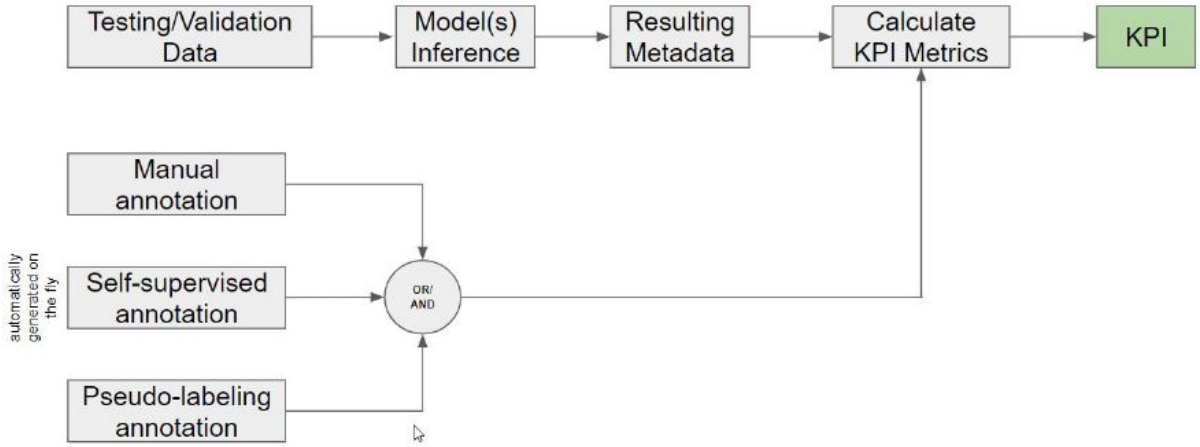

Figure 3: Overview of AI Model Use in VALEO Application

need for costly and hard-to-obtain annotations. Models trained using the developed approach need fewer annotated samples to perform the downstream task to reach the same performance as fully-supervised models and therefore lower the requirements for manual annotation. Moreover, the resulting models trained with our approach are more robust due to the amount of data they use during the training. The potential impact of this work is an increased safety on roads with better, more accurate, and robust car perception systems.

The AI-based model construction and model use pipelines are presented in Figure 2 and Figure 3, respectively.

2.2.3 Elements/Actors/System

In the cases described above, multiple stakeholders/roles are involved:

- **Tool developer**. Develops data processing tools (like convertors, metadata extraction KPI scripts) and/or automated workflow for their big-scale execution.
- **HIL (Hardware-in-the-Loop)**. An engineer, which takes care of preparation, maintenance and smooth HILs operation (as automated workflow).
- **Test Engineer (or Trace Analyst)**. Analyses results and creates test reports.
- **System Engineer**. Analyses issues and fixes/improves the system algorithms.
- **AI/ML Engineer**. Develops algorithms and models to extract metadata and/or to automatically pre-annotate (Ground Truth extraction algorithms).

2.3 Industry Application Case - Data Management (WP5)

2.3.1 Description

The VALEO application case includes the following data management-focused use cases:

- Advanced Driver Assistance System (ADAS) Data Management, Workflow, Processing, and Analytics platform,
- Toolbox of Machine Learning (ML) models for recorded ADAS data content interpretation (metadata extraction),
- Toolbox for automatic Ground Truth pre-annotation.

To develop and validate ADAS, high-quality appropriate data sets have to be recorded. They must represent the statistical distribution of real conditions of target product use (as original safety/comfort equipment of vehicles). Data recorded in ADAS test vehicle fleet are ingested in the data platform, and processed through multiple stages up to the ADAS System KPIs, which are necessary to release ADAS System for production. In Table 1, we present data rates produced by different sensors in a test vehicle.

2.3.2 Usage Scenarios

A simplified validation data lifecycle consists of the following steps:

- Data are transported from test vehicles (driving worldwide on target markets) to data centers, typically via high-volume external data media.
- Data are checked for quality. Some of the commonly used checks are as follows:
  - Is file integrity problem-free?
  - Are all required inputs recorded?

Data Rate	Content
50MiB/s - 6GiB/s	Raw data from ADAS sensor under test, e.g. full-scale images from ADAS Front/Surround-view Cameras, point cloud from LiDAR or raw ADC samples from RADAR front end
50MiB/s - 1GiB/s	Reference-system data (additional LiDARs or cameras sets or other vehicle data)
50KiB/s - 5MiB/s	Vehicle bus data (CAN, FlexRay, BroadR-Ethernet)
10KiB/s	GNSS data (Global Navigation Satellite System - e.g. GPS with RTK extension (Real-Time Kinematics, allowing cm-level accuracy location))
50KiB/s	IMU data (Innertial Measurement Unit data including 3D gyro, 3D accelerometer and 3D compass)

Table 1: Typical Recorded Data Content in ADAS Test Vehicles

- Are all frames from all sensors available and complete?
- Are all vehicle configurations and calibration data problem-free?
- Delete scrap data (fix test vehicle equipment, etc.).

- Basic metadata is extracted, including the following commonly used ones:
  - Weather conditions (using weather API based on vehicle Global Navigation Satellite System (GNSS) position and time),
  - Ego vehicle speeds (min, max, avg, etc.),
  - Other road users,
  - Country.
- Statistical model matching of newly recorded data is checked with the required distribution (based on the metadata).
- Decisions are taken as to which parts of the recorded data is to store or to delete.
- Decisions are taken on whether to use the recorded data and annotate as Ground Truth.

Simplified development and validation cycle stages (automated, high-scale) include the following steps:

- Recorded raw ADAS sensor data are prepared for reprocessing by a new SW (subject of validation). It can be executed on real ECU HW (HIL - Hardware-in-the-Loop) or as a software package only (SIL - Software-in-the-Loop – typically as a Docker image).
- Processed data is stored.
- System KPIs are calculated (and processed data is compared with Ground Truth)

- System KPIs are analysed.
- Deviations and issues are analysed (If occurrences of similar situations are found based on metadata, software issues are fixed/tuned, and tested on similar situations.).

The steps are repeated for each new software.

2.3.3 Elements/Actors/System

In above described cases, multiple stakeholders/roles are involved:

- **Campaign planner**: Plans locations to drive for each vehicle to record a statistically distributed data set.
- **Fleet/Operations manager**: Monitoring of test vehicle fleet and data collection.
- **Logistics engineer**: Managing delivery of data media between ingest site/data centre and test vehicle and between data centre and external annotation sites.
- **Ingest engineer**: takes care of data ingestion process from data media up to target data storage and ingest processing.
- **Data platform product owner**: Collects requirements from stakeholders for data platform, maintains and prioritises product backlog and product increments, organises definition of user stories.
- **Data platform architect**: Designs core components, SW technologies and services for big data platform - data/content management, encryption standards, IAM, databases, monitoring, cloud services integration.
- **IT Infrastructure Architect**: Designs and configures IT technologies as enablers for big data processing (data storage, networking (LAN, WAN), data compute nodes, cloud).
- **DevOps Engineer**: Takes care of Data Platform automated deployment and production environment monitoring.
- **Automation Engineer**: Using Data Platform to develop, automate and execute workflows.
- **Dataset engineer**: Takes care of data set preparation for specific use cases/ODDs (Operational Design Domains), well distributed, annotated, etc.

2.4 SME Application Case - Health (WP6)

The global objective of this application case is to transpose EXA4MIND capabilities into a concrete application. The proposed application case aims to demonstrate the platform ease of use in the domain of secured health data management. It will emphasise how the platform can gather heterogeneous information from different sources (i.e.: numerous hospitals), transform them, and output an exploitable database. It will give the capability to various users, as defined below, to integrate data, analyse and exploit them.

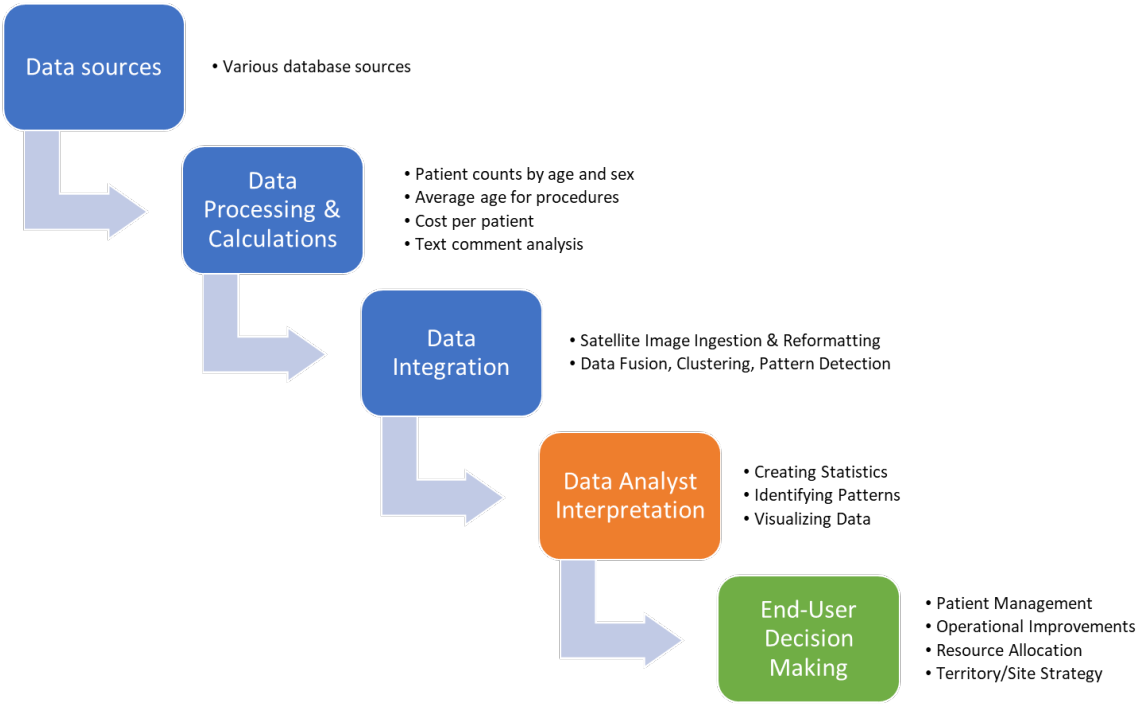

Figure 4: Overview of ALTRNATIV Application

2.4.1 Usage Scenarios

In this application case, health data will be securely mined from different sources, containing patient demographics and clinical histories, respectively. Our analysis will involve processing 4.5 million patients and 100 million procedures, calculating key statistics such as patient counts by age and sex, the average age for specific procedures, and costs per patient. We will also analyse text comments for insights, create tables in the Extreme Data Database (EDD) for efficient data storage, and ultimately combine our results with satellite images for geolocation and pattern detection purposes. This comprehensive approach offers valuable insights for decision-makers, allowing for better resource allocation and improved patient care. The application case is presented in Figure 4.

2.4.2 Elements/Actors/Systems

For the above-described case, the following users' roles are involved:

1. **Data engineer**: Takes care of pipeline creation (data collection, preparation, ingestion, transformation), for example:
   - Import a CSV file containing millions of rows and fields with information about the patients and performed procedures,
   - Define concrete SQL queries performed on the imported data (aggregation, grouping, counting, etc.)

2. **Data analyst**: Takes care of the interpretation and exploitation of the data, for example:

   - Creating Statistics: Data analysts can summarise the raw data into meaningful statistics, such as average age for procedures, procedure frequencies, and costs associated with treatments. These statistics can help identify trends and provide actionable insights to decision-makers.
   - Identifying Patterns: By examining patterns in patient demographics, procedures, and costs, data analysts can highlight areas that may require attention or reveal opportunities for improvement. For example, they can detect if certain procedures are more common in specific age groups or if costs are disproportionately high for particular treatments.
   - Visualising Data: Data analysts can create visual representations of the data, such as graphs, charts, and heatmaps, to better communicate findings to end-users. This can make it easier for stakeholders to understand the patterns and trends in the data.
   - Monitoring Trends: By continuously analysing the data over time, data analysts can monitor the effectiveness of implemented changes and track the progress of specific initiatives, ensuring that improvements are sustained and resources are used efficiently.
   - Providing Recommendations: Based on the insights derived from the data, data analysts can offer actionable recommendations to decision-makers. These suggestions can help optimise resource allocation, streamline processes, and improve patient care.

3. **End-users** (doctors, management, service director, administrations): Can leverage this health data analysis to make informed decisions regarding patient care, operational improvements, and resource allocation. By examining the demographic distribution, average age for specific procedures, and cost statistics, users can identify trends and pinpoint areas needing attention or improvement. For instance:

   - Global Patient Management: By understanding patient demographics and procedure frequency, end-users can allocate resources, staff, and equipment more efficiently, ensuring that patients receive timely care.
   - Reinforcing Operational Services: The analysis of procedure costs and the frequency of medical services per patient can highlight areas where operational efficiency could be improved, leading to better budgeting and strategic investments.
   - Identifying Procedure Deficiencies: Analysing the text comments and procedure patterns can reveal gaps or deficiencies in current procedures, allowing end-users to take corrective action or implement new protocols.
   - Territory/Site Strategy: Combining health data with geolocation information from satellite images enables end-users to identify geographic patterns and optimise resource allocation across multiple sites. This can help establish

new facilities in underserved areas or reallocate resources to better serve the needs of the population.

By leveraging this data analysis, end-users can make more informed decisions, leading to improved patient care, operational efficiency, and better overall health outcomes for the community.

2.5 SME Application Case - Agriculture (WP6)

Soil moisture profiler service is an enabling component within TerraviewOS architecture that empowers variable rate irrigation prescription maps for customers who currently have weather stations deployed that can report evapo-transpiration rates. The problem is that more than 98% of customers do not have such sensors deployed. The proposed application case will use the data management and transformation capabilities of EXA4MIND to empower the sensor-less soil moisture profiler service which will use the capabilities of a machine learning pipeline with EO data sets and forecasted weather trends over an area of interest (AoI) to generate predicted moisture profiles to empower irrigation prescription maps even for those land parcels which have no evapo-transpiration sensors available.

2.5.1 Usage Scenarios

Sensor-less soil moisture mapping service will allow agriculture practitioners to optimise the use of water for agriculture. Optimisation of water usage is a high-priority task within the united nations' sustainable development goals (SDGs) (UN Resolution 2015). 70% of fresh water globally is used for agriculture, of which further just 30% is used efficiently, rest is washed away as runoff. The application case is presented in Figure 5.

2.5.2 Stakeholders

The stakeholders are key decision makers in the agriculture farmland management company, or even the farmers directly. The usage sequence will look as follows:

1. The user will select AoI using a drawing tool on the map shown in a web-UI.
2. The requested AoI will be sent to the TERRAVIEW Sensorless Soil Moisture Mapping service module.
3. If a pre-trained model for the AoI exists, the user will get a prediction of the soil moisture profile for the selected AoI for the next 2 weeks.
4. The user can take irrigation decisions based on the predicted soil moisture profile.
5. This service will be made available as a standalone service for the benefit of global agriculture communities.

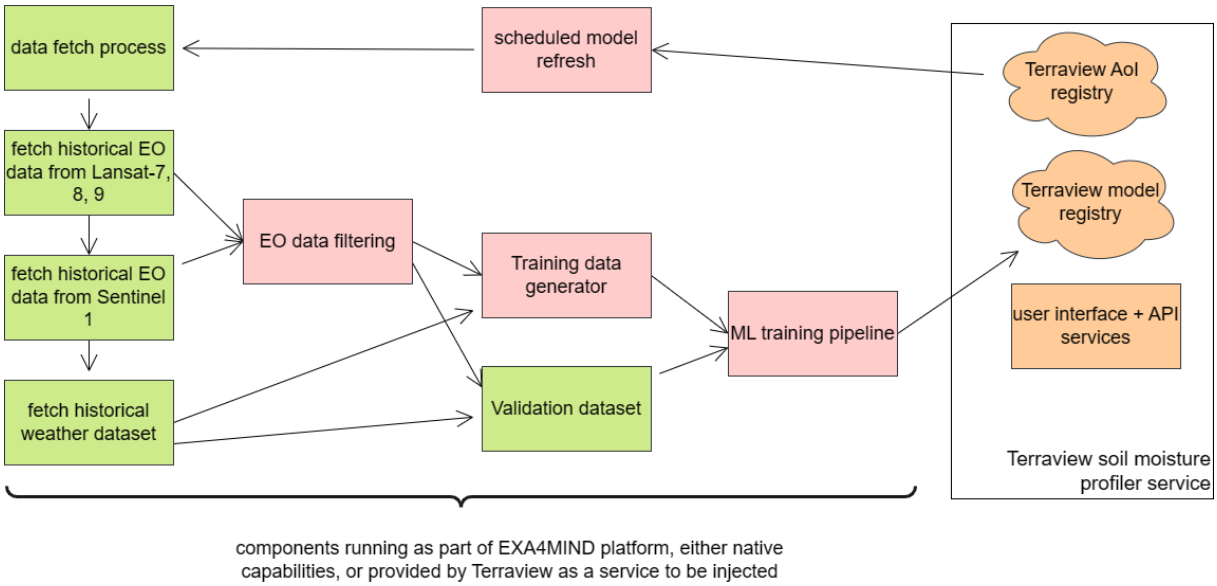

Figure 5: Overview of TERRAVIEW Application

6. If the user is an existing paid user of the TerraviewOS platform, they will receive detailed variable irrigation prescription maps for their vineyard plots powered by the soil moisture profiler independent service.

If the pre-trained ML model does not exist for the AoI, the model generation request goes into the model generator queue, which will first finish the model generation, and then send the forecast profiles to the user once completed.

2.5.3 Elements/Actors/Systems

1. **ESA Copernicus program** (European Commission / Copernicus 2023a): The source of optical and thermal, as well as radar observation data sets over the area of interests.
2. **ESA Copernicus historical weather data sets** (European Commission / Copernicus 2023b) as well as **short-range / mid-range forecasts** powered by ECMWF.
3. **USGS / NASA Landsat program** (NASA / USGS 2023): The source of optical/thermal EO datasets over an area of interests.
4. **Data prefetching service**: An archival data facility offered by EXA4MIND where we can specify AoI or a set of AoIs and the service gets all available historical EO data for the AoI from Copernicus, USGS, and also historical weather data set.
5. **Training data generator**: This service is written and provided by TERAVIEW which uses the prefetched data of an AoI to generate the training and validation data sets for a specific AoI – staged for temporarily use on the HPC parallel file system by EXA4MIND.
6. **Trained model registry**: This service is provided by TERRAVIEW which will curate the models for different regions.

7. **Model Training and validation workflow**: Ability in EXA4MIND to schedule ML workflows.
8. **Data refresh capability**: Ability in EXA4MIND to refresh the archived data sets bringing in the latest EO and weather data values from Copernicus / USGS service endpoints since the last collection time.
9. **Stream data ingestion**: The ability to ingest IoT sensor data values on a continuous basis.

2.5.4 Data Model Management

TerraviewOS soil moisture profiler service will operate in multi-model mode. Due to variation in land topography, as well as soil types, it is foreseen that no one model will be sufficient enough to provide reasonable accuracy at the global scale. The ML training workflows will be designed to generate soil type specific models as well as regional and topography specific models where ever possible. The models will be periodically refreshed as new satellite and ground truth data gets ingested.

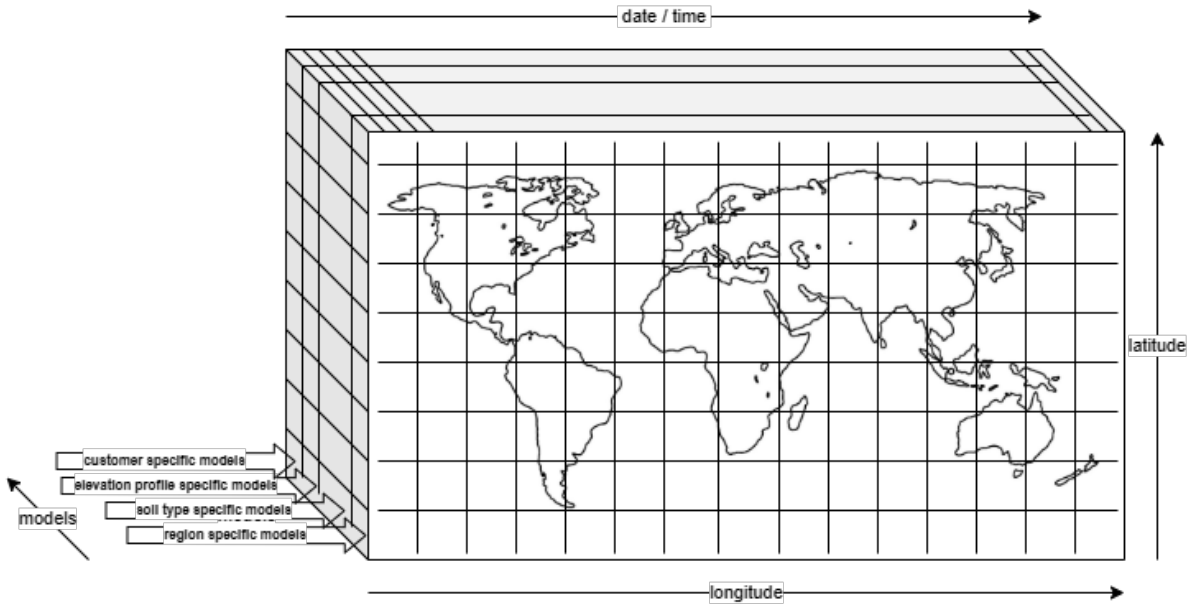

Figure 6: Overview of TERRAVIEW Data Cube

The inference workflows using various models will be done periodically where different models will be used depending on the location of the agricultural plots. In this aspect, the data management appears similar to that of a large data cube. Visualisation of TERRAVIEW's data cube requirement can be best represented via Figure 6. Each cubelet in the figure would represent the data value chain resulting in multiple soil moisture models for the geographical-region represented by the cubelet.

3 Architecture Requirements

This section contains a result of a discussion between the technical work packages (WP2 and WP3) and the individual application case work packages (WP4, WP5, and WP6), and a set of initial requirements on the EXA4MIND Platform, which will be transformed to the platform features in the first version of the platform architecture in M9. The list of questions and answers provided during the discussion is located in Section 3.1. The resulting set of requirements is listed in Table 6 which is then mapped back to the technical question tables with requirements summary for each WP (see Table 2, Table 3, Table 4, and Table 5).

3.1 Technical Questionnaires

The requirement collection was done in two steps:

1. Initial application case description,
2. Application case detailed questionnaire.

The goal of the initial description was to get an overview of the application case. Each application case owner was asked to provide a short description that included:

1. One paragraph description of the application case,
2. Plain words description of the steps in the current application case,
3. Expectations from the EXA4MIND Platform.

After getting an overview of each of the application cases in phase one, a detailed questionnaire was created. The aim of the questionnaire is to get a deep insight into the application cases and their needs. The questionnaire has the following main sections:

1. Detailed description of the application case,
2. Data Management,
3. Data Analytics,
4. Workflow.

3.1.1 Distributed Data Space Management (WP2)

Table 2 and Table 3 contain questions related to the EXA4MIND Platform requirements from the WP2 point of view and answers provided by the application cases. It focuses on various aspects of data provided by the application cases which the platform should handle.

Question	WP4 ADAMS4SIMS	WP5 AI	WP5 Data	WP6 Health	WP6 Agriculture	REQ
Will you need to import data into the platform?	Yes	Yes	Yes	Yes	Yes	R1
What kind of data will you import and how?	MD simulations (∼1 TB), Metadata text files, FF file (∼MBs)	Public data sets and VALEO Data	Sensor raw data (∼1-6GB/s), raw unprocessed data, config files	Various data and satellite images	Earth observation data as GeoTIFFs and associated metadata files	R2
How frequently will you need to import new data sets?	Weekly/monthly	Initial data + every 3 months	Full-recordings - 3 shifts per day (8 WH each), 1-5 cars	Initial data set once, then depending on availability of the streaming sources	Weekly for all satellite missions used	R1
What size are these data sets?	TBs + web app can generate hundreds of 1-100 MB files in a single run	Provided by WP5 Data	2.5 PB	100 TB	Varies, each area of interest is ∼4TB	R1
Does the application case use different types of storage systems?	To be clarified in the project	Fast access storage - parallel filesystem on HPC clusters	AWS S3 on the commercial cloud, local storage arrays	AWS S3	Any POSIX compatible storage is acceptable	R3
Do data sets need to be encrypted or compressed?	Benefits of compression of binary format files to be checked in the project	No	Unencrypted, optional compression can be beneficial for staging	No compression, encryption, based on the data analysis	No	R4, R5
What type(s) of database would you envisage using and are you currently using a specific database software?	Subject to further investigation leveraging EXA4MIND solution	N/A	ElasticSearch, MS SQL, PostgreSQL, InfluxDB, Prometheus, Neo4j	SQLite, ElasticSearch, Parquet	PostGIS, PostgreSQL	R2, R3
How is your data structured? And what types of format do you use?	Four layers - MD simulation data set, knowledge extraction files, force field parameter files, experimental data	jpeg, png, avi, mpeg, pcd, lvx, json, xml	Data containers (Elektrobit/ADTF DAT file, PCAP, MDF4 - Bosch, npz - numpy format), Sensor specific formats: (PCD, *.lvx, CSV, XML, json, MPEG4, JPEG, PNG)	CSV and possibly document pictures and satellite imagery	GeoTIFFs, GeoJSON, plaintext files with metadata	R2
What is the expected volume of the data to be processed/produced/transferred in each step of your application case?	up to ∼1 TB	∼2.5 PB	Target application specific	100 TB	Depends on the area of interest (expected ∼4 TB)	R8

Table 2: WP2 Requirements Summary - Part 1

Funded by the European Union

Question	WP4 ADAMS4SIMS	WP5 AI	WP5 Data	WP6 Health	WP6 Agriculture	REQ
Do you provide special metadata with your data i.e are you using a field specific metadata scheme? If yes, please mention the scheme or provide an example.	Standards-based metadata management for molecular simulations (Grunzke et al. 2014)	Metadata is usually stored in XML/JSON format	JSON, no standard metadata structure	No	Proprietary GeoJSON	R9
Do you want to share your data with high bandwidth to internal/external collaborators/stakeholders?	Yes, PIDs required, no requirement on bandwidth	Only in case VALEO decides to make a data set with metadata public	Only in case VALEO decides to make a data set with metadata public	Depends on the data source	Limited only for the consortium	R9
Can your software environment or container images be stored at the public repositories as DockerHub or do you have/need any private repository?	Yes, OpenSource software is used	No, only in private repositories shared with the consortium	No, only in private repositories shared with the consortium	No, only in private repositories shared with the consortium	No, only in private repositories shared with the consortium	R1, R9
Do you keep/do you need to keep your data versioned?	Possibility for simulation extension (prolongation) - the question is if it is new version or linking to original data.	Yes	Yes	Yes	No for input data, versioning of the models already being used	R9, R10

Table 3: WP2 Requirements Summary - Part 2

Funded by the European Union

3.1.2 Extreme Data Analytics and Processing (WP3)

This section contains another set of questions in Tables 4 and 5 from the point of view of WP3 focusing on development of the advanced query and indexing system (AQIS). It is focused more on handling different types of databases within the platform, query interfaces used by the cases, and the way the cases are using them.

3.2 Identified Requirements

Table 6 contains the list of initial identified requirements on the EXA4MIND Platform, which will be reflected in the platform architecture as its features. The list is the result of an extended discussion between the application cases and technical WPs and provides an initial insight on the required features of the EXA4MIND Platform.

In Figure 7, we provide a mapping of the identified requirements to the original concept of the EXA4MIND Platform described in the project proposal. This will be used as a basis for identification of the data flows and to derive the first version of the EXA4MIND Platform architecture.

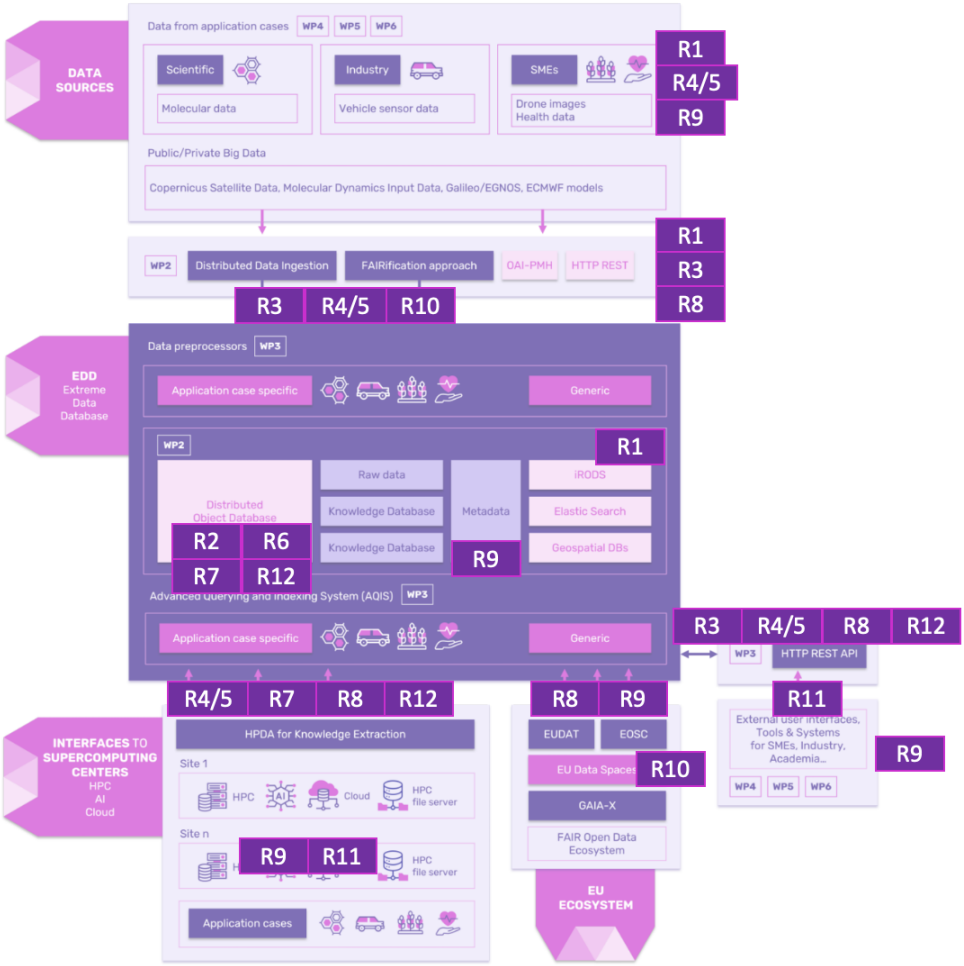

Figure 7: Mapping of Requirements on Original EXA4MIND Concept

Question	WP4	WP5 AI	WP5 Data Management	WP6 Health	WP6 Agriculture	REQ
What type(s) of database query languages are you currently using and would you envisage using?	None, no preference	N/A	Elastic queries, SQL, GraphQL, Neo4j	Will be adapted	PostgreSQL/PostGIS for postprocessed results	R7
What kind of database queries are you currently using and would you envisage using?	None, hypercube approach will be evaluated	N/A	ElasticSearch (LUCENE), SQL	SQL, examples provided in the application description	SQL with GIS extensions	R7
Which software tools and libraries do you use for format conversion on the data?	Not applied currently	Not defined yet	Raw data are converted to toolchain specific formats. Tools such as Python (NumPy, Pandas), ADTF libraries, National Instrument Labview, Matlab are used.	Data is transformed from CSV (and potentially images & satellites pictures) to a SQLite or parquet database. Alternative solutions, such as DBMS, will be explored.	PyroSAR, raster IO, GDAL	R11
What is the current average time cost of format conversion operations?	N/A	N/A	Time costs vary with the project	In the order of hours/days	Not evaluated yet	R11
Which software tools and libraries do you use for data cleaning?	Not applied currently	Not defined yet	Quality checks statistical model matching. Similar tools as in the above question.	ALTRNATIV's proprietary data transformation module	PyroSAR, raster IO, GDAL	R11
What is the current average time cost of data cleaning?	N/A	N/A	Time costs vary with the project	In the order of hours/days	Not provided	R11
Which statistical software tools and libraries do you use?	Data-driven scripting language is used	Statistical results of AI model performance as KPI.	Statistical functionality of Python modules like NumPy, Pandas, SciPy	ALTRNATIV's proprietary data analytics module	Simple statistical functions to generate useful metrics for our users which are captured within GeoJSON object.	R11
What is the current average time cost of statistical function?	Typically performed within minutes (considering trajectory with hundred thousands of snapshots)	Average time cost is proportional to the amount of data processed (GBs to PBs).	Time costs vary with the project	In the order of hours/days	Not provided	R11

Table 4: WP3 Requirements Summary - Part 1

Funded by the European Union

Question	WP4	WP5 AI	WP5 Data Management	WP6 Health	WP6 Agriculture	REQ
Which machine learning software tools and libraries do you use for knowledge discovery/extraction?	Not yet currently	Mainly Pytorch library, some state of the art method public codes e.g. in Tensorflow	Semantic segmentation and static object detection, clustering of LiDAR point clouds Tools built upon Keras, Tensorflow, PyTorch, SciPy are used.	For detection of abnormal patterns of object classification/recognition	Time-series extrapolation will be used to generate predictions of relative soil moisture content for the particular AoI.	R12
Are there any pre-processing/data analytics tasks that you particularly experience difficult (in terms of time cost, lack of practical solution or available tool, etc.)?	Advanced querying of a hypercube with extensive amounts of data	Sensor intrinsic and extrinsic calibration especially for the captures, where the sensors calibration data are not known in advance. Sensor to sensor calibration (e.g. LiDAR to camera).	Pre-selection of good quality data. Time and space domain sensor calibration.	No	Full historical image scene processing due to the storage space requirements.	R12
In a data analytics workflow, do you keep/need to keep intermediate results?	No	Yes, but mainly to save the computational efforts and for debugging reason	Intermediate results are kept for different versions of the pipeline so that the pipeline can resume from the intermediate state in an event of infrastructure failure	Intermediate results need to be kept to ensure reversibility of data and errors management.	Not at the moment, but if delta model training or transfer learning is possible, intermediate results could be be stored for incremental training .	R8, R9
Do you use any visualisation tools or libraries? Which ones do you use?	xy – plots, distribution plots/bar charts	Mainly standard Python modules	Elastic Stack (Kibana), Python (matplotlib, etc.), Proprietary tools (DASTE, FrontCam studio) CloudCompare, Grafana (mainly for monitoring of systems), Annotation tools (Philosys Label Editor / Ground Truth Annotator, Pixano, Vicomtech SAA)	Various representations such as bar charts, pie charts, radar charts, etc. ALTRNATIV's proprietary solution or via other solutions such as Tableau Software or Kibana.	Satellite data is rendered on user devices using Leaflet and linked data is packed as GeoJSONs. In browser, web-charting libraries are used to show sensor streaming data as charts.	R6, R11

Table 5: WP3 Requirements Summary - Part 2

Funded by the European Union

Number	Requirement description
R1	Data import capability
R2	Support for various types of data
R3	Support for different protocols including preprocessing
R4	Data encryption
R5	Data compression
R6	Support for different DB technologies
R7	Efficient query execution
R8	Data staging
R9	Metadata ingestion support, support for FAIR data, reproducibility and trust
R10	Support for versioning and handling large datasets
R11	Querying by natural language processing prompts, support for Large Language Models
R12	Support for advanced data querying on large data collections

Table 6: Initial EXA4MIND Platform Requirements

4 Conclusion

In this deliverable, we conducted an extensive analysis of the application cases in the project and derived an initial set of requirements on the EXA4MIND Platform. These requirements will be used to define the data flows to be further described in detail in deliverable D2.1 - *Extreme Data Flow Patterns*. Both D1.1 and D2.1 will also serve to define the initial architecture of the platform which is due in M9 as milestone M1.1 - *Initial Architecture Defined*.

The first part of the deliverable, Section 2, contains a structured description of each application case, including the graphical representation of the application workflow and the identification of usage scenarios and involved stakeholders.

The next part (Architecture Requirements, Section 3), first shows the results of a dialogue on this topic between the technical WPs and the application case WPs. Based on targeted questions, we obtained a structured response, presented in the form of several tables. Together with the application case descriptions, these have then been used to derive a first set of requirements on the EXA4MIND architecture (Table 6). We also provide a mapping of these requirements back to the original questions asked and to the EXA4MIND concept diagram presented in the project proposal. This is a first input to the co-design process which will iteratively address requirements coming up at later stages as well.

5 References

Behley, J. et al. (2019). "SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences." In: **Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)**.

Bottaro, Sandro et al. (2018). "Conformational ensembles of RNA oligonucleotides from integrating NMR and molecular simulations." In: **Science Advances** 4.5, eaar8521. DOI: 10.1126/sciadv.aar8521. eprint: https://www.science.org/doi/pdf/10.1126/sciadv.aar8521. URL: https://www.science.org/doi/abs/10.1126/sciadv.aar8521.

Caesar, Holger et al. (2020). "nuScenes: A multimodal dataset for autonomous driving." In: **CVPR**.

Case, D.A. et al. (n.d.). **Amber 2022, University of California: San Francisco, CA, USA. 2022**.

Cordts, Marius et al. (2016). "The Cityscapes Dataset for Semantic Urban Scene Understanding." In: **Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**.

European Commission / Copernicus (2023a). **Homepage | Copernicus**. https://www.copernicus.eu, Cited 16 Apr 2023.

— (2023b). **Homepage | Copernicus | Climate | Change**. https://cds.climate.copernicus.eu/, Cited 16 Apr 2023.

Frohlking, Thorben et al. (2022). "Automatic learning of hydrogen-bond fixes in the AMBER RNA force field." In: **Journal of Chemical Theory and Computation** 18.7, pp. 4490–4502.

Grunzke, Richard et al. (July 2014). "Standards-based Metadata Management for Molecular Simulations." In: **Concurrency and Computation: Practice and Experience** 26, pp. 1744–1759. DOI: 10.1002/cpe.3116.

Köfinger, Jürgen, Bartosz Różycki, and Gerhard Hummer (2019). "Inferring Structural Ensembles of Flexible and Dynamic Macromolecules Using Bayesian, Maximum Entropy, and Minimal-Ensemble Refinement Methods." In: **Biomolecular Simulations: Methods and Protocols**. Ed. by Massimiliano Bonomi and Carlo Camilloni. New York, NY: Springer New York, pp. 341–352. ISBN: 978-1-4939-9608-7. DOI: 10.1007/978-1-4939-9608-7_14. URL: https://doi.org/10.1007/978-1-4939-9608-7_14.

Kuhrova, Petra, Robert B Best, et al. (2016). "Computer folding of RNA tetraloops: identification of key force field deficiencies." In: **Journal of chemical theory and computation** 12.9, pp. 4534–4548.

Kuhrova, Petra, Vojtech Mlynsky, et al. (2019). "Improving the performance of the amber RNA force field by tuning the hydrogen-bonding interactions." In: **Journal of chemical theory and computation** 15.5, pp. 3288–3305.

Mao, Jiageng et al. (2021). "One Million Scenes for Autonomous Driving: ONCE Dataset." In.

Mlynsky, Vojtech et al. (2022). "Toward convergence in folding simulations of RNA tetraloops: Comparison of enhanced sampling techniques and effects of force field modifications." In: **Journal of Chemical Theory and Computation** 18.4, pp. 2642–2656.

NASA / USGS (2023). **Homepage | Copernicus | Data | Access**. https://www.usgs.gov/landsat-missions/landsat-data-access, Cited 16 Apr 2023.

Sakaridis, Christos, Dengxin Dai, and Luc Van Gool (2020). "Map-Guided Curriculum Domain Adaptation and Uncertainty-Aware Evaluation for Semantic Nighttime Image Segmentation." In: **IEEE Transactions on Pattern Analysis and Machine Intelligence**. DOI: 10.1109/TPAMI.2020.3045882.

— (Oct. 2021). "ACDC: The Adverse Conditions Dataset with Correspondences for Semantic Driving Scene Understanding." In: **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)**.

Shen, Tongye and Donald Hamelberg (2008). "A statistical analysis of the precision of reweighting-based simulations." In: **The Journal of chemical physics** 129.3, p. 034103.

Smith, Louis G. et al. (2017). "Physics-based all-atom modeling of RNA energetics and structure." In: **Wiley Interdisciplinary Reviews: RNA** 8.

Šponer, Jiří et al. (2018). "RNA Structural Dynamics As Captured by Molecular Simulations: A Comprehensive Overview." In: **Chemical Reviews** 118.8. PMID: 29297679, pp. 4177–4338. DOI: 10.1021/acs.chemrev.7b00427. eprint: https://doi.org/10.1021/acs.chemrev.7b00427. URL: https://doi.org/10.1021/acs.chemrev.7b00427.

Sun, Pei et al. (July 2020). "Scalability in Perception for Autonomous Driving: Waymo Open Dataset." In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**.

UN Resolution (Oct. 2015). "Transforming our world : the 2030 Agenda for Sustainable Development." In: **Issued in GAOR, 70th sess., Suppl. no. 49.** URL: https://digitallibrary.un.org/record/3923923.

Zgarbová, Marie et al. (2011). "Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles." In: **Journal of Chemical Theory and Computation** 7.9. PMID: 21921995, pp. 2886–2902. DOI: 10.1021/ct200162x. eprint: https://doi.org/10.1021/ct200162x. URL: https://doi.org/10.1021/ct200162x.

This is the version of the deliverable before the review by European Commission.