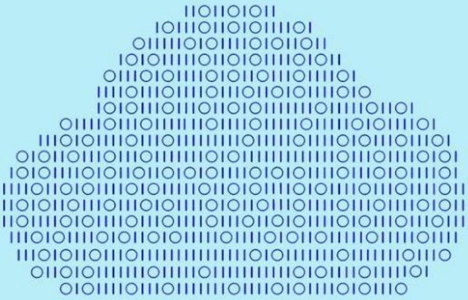
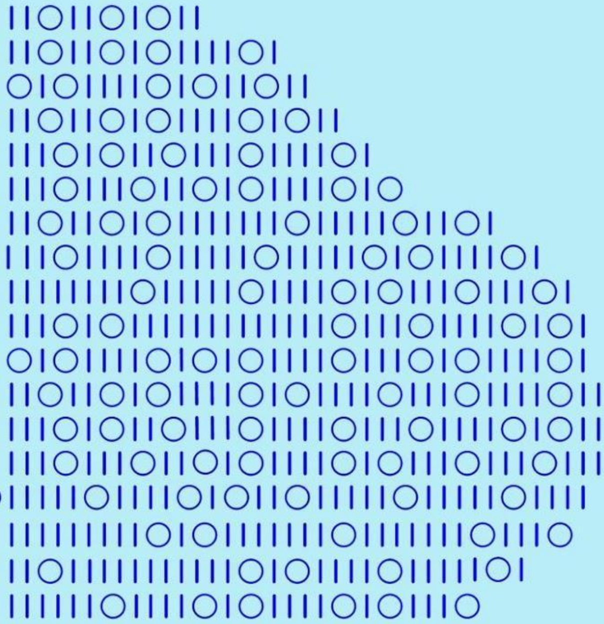
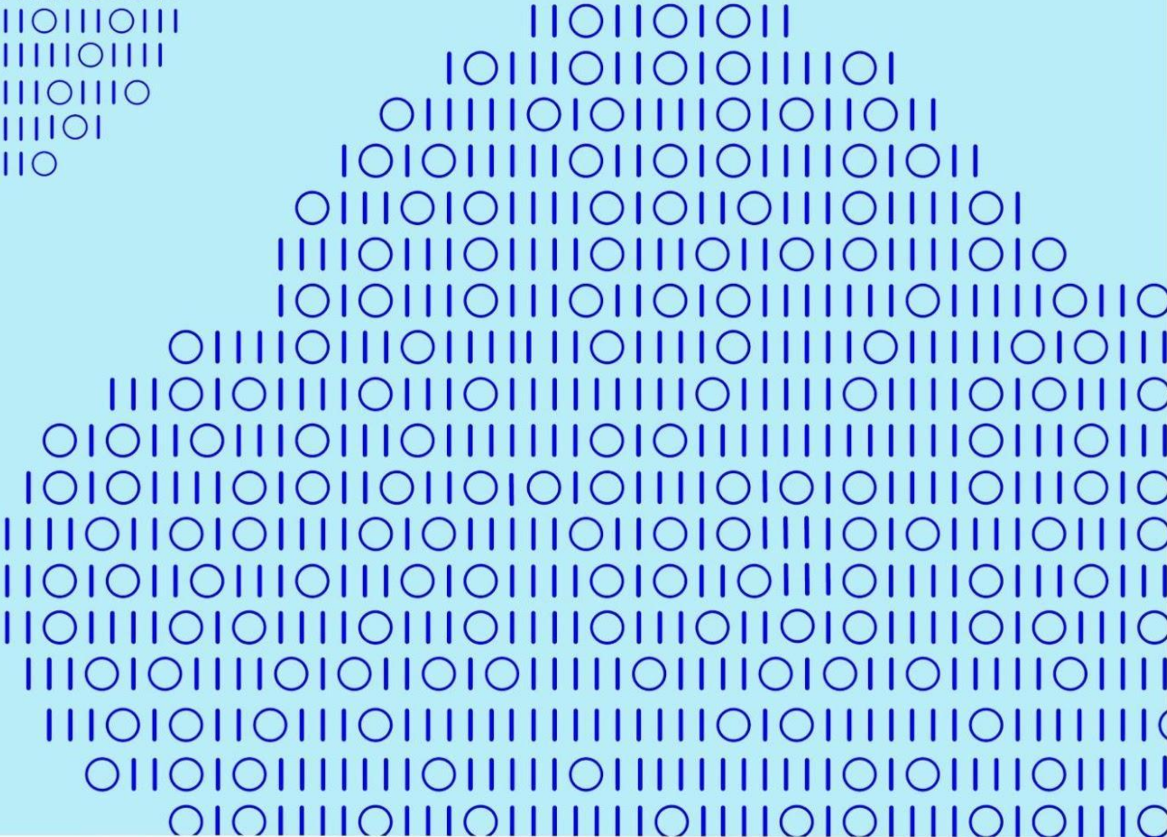
AIDA

Adaptive, Intelligent and Distributed Assurance Platform







Deliverable 1.2

Platform Architecture and APIs (Report)

Public

Date 15/05/2022

Activity 1 T1.2, T1.3

Version 1.0

Authors Bruno Sousa, Carlos Martins, João Vinagre, José Pereira, Nuno Antunes, Pedro Fidalgo, Ana Nunes, Ricardo Vilaça

COMPETE 2020 · PORTUGAL 2020 · UNIÃO EUROPEIA Fundo Europeu de Desenvolvimento Regional · FCT Fundação para a Ciência e a Tecnologia · Carnegie Mellon Portugal

Table of Contents

Executive Summary

Report describing overall design for the proposed risk monitoring platform: the high level architecture for the risk monitoring platform and its components: distribution/orchestration, distributed machine learning and security/privacy issues.

1. Introduction

RAID is Mobileum's platform that handles the entire risk management lifecycle of enterprises. It comprises a very fluid pipeline, which covers several steps as data collection, monitoring, notification, discovery and actuation, to provide several services. Companies over the world are served by RAID to capture revenue for all services rendered (revenue assurance), for business assurance and fraud management, in an end-to-end real-time manner.

The Adaptive, Intelligent and Distributed Assurance Platform (AIDA) project has the main goal of conceiving a new version of the current RAID platform where some of the pipeline phases can be dynamically moved to the edges of the system. Currently, the platform is fully deployed in physically co-located servers, either on premises or in the cloud. AIDA aims to provide highly configurable data collection and monitoring while preserving real-time, security and dependability guarantees, with ability to run in diverse hardware. The market pressure requires RAID to become effective in scaling to unprecedented levels while assuring data privacy and confidentiality of the data processed across the different administrative domains and owners. This poses significant and interesting research and development challenges in the area of security of data and infrastructure which need to be addressed.

AIDA faces the identified challenges through several perspectives that compose the activities within this project. Namely, focusing on distribution of the platform to encompass the cloud and the edge, which allows the platform to take advantage of 5G networks to have more effective scaling; utilization of distributed and federated machine learning to enable model training on distributed data and to learn with both behavior patterns and context evolution, leveraging this process for risk management purposes; and also assure security of the platform, through intrusion detection and tolerance to assure the continuity of service, and data privacy and confidentiality, which are fundamental concerns in the design of the platform, as it should be able to work with anonymized data, as well as provide secure data exchange between the different processing elements in distinct environments.

This document specifically focuses on the high level architecture and APIs for the risk monitoring platform and its components: distribution/orchestration, distributed machine learning and security/privacy issues.

1.1. Target Audience

This document is intended for public use. This document serves the purpose of defining the AIDA platform's design, architecture and general APIs.

COMPETE 2020 PORTUGAL 2020 UNIÃO EUROPEIA Fundo Europeu de Desenvolvimento Regional FCT Fundação para a Ciência e a Tecnologia Carnegie Mellon Portugal

1.2. Document Structure

The rest of this document is structured in the following four main sections.

Section 2 presents the AIDA plastform goals and requirements. Section 3 describes the AIDA and RAID high level architecture. Section 4 describes the APIs of the components exposed to the exterior, as well the internal service APIs. Finally, Section 5 concludes the deliverable.

2. Platform overview and design

5G presents an opportunity for telecom operators to capture new revenue streams from industrial digitization. In cases such as network-as-a-service (NaaS), network exposure is becoming a reality through the transformation of core telecom network assets into digital assets. With 5G, the dynamic provisioning and scaling of network capacity and resources are available for the first time.

The Adaptive, Intelligent, and Distributed Assurance Platform, AIDA, project aims to deliver this vision, an end-to-end 5G-ready fraud management platform that is able to protect the 5G ecosystem in its multiple layers, and deploy an strategy that manages high data volumes and real-time visibility through edges close to the monitoring points, contributing with scalability and local learning to global models. Additionally, 5G introduces challenges that previous generations did not have. The multitude of deployment scenarios between isolated or shared and private and public networks, and the multiple business entities and partners involved in the new business models, introduces intrusion, tampering, confidentiality, and data privacy requirements that need to be monitored and analyzed, ensuring system-wide protection of the ecosystem and value chain.

AIDA addresses 5G's scalability and privacy challenges by enhancing RAID's engine and expanding its automation capabilities. In particular, the project will be focusing on the following goals:

- Leverage edge computing and 5G - to distribute RAID platform components to delegate processing to the edge or use central servers, according to the nature of the computation and the type and localization of monitoring and reference data.
- Explore emergent federated machine learning techniques - to learn from local data and push incremental model updates to coordinator nodes that maintain global models based on the contribution of edge nodes and other relevant data sources.
- Test resilience to intrusion or tampering - by requiring the research and application of intrusion detection techniques at multiple levels of the architecture, with the goal of enabling system-wide intrusion tolerance.
- Protect data privacy and confidentiality – by maintaining the confidentiality of the operational data being monitored, analysed, and protecting the privacy of the entities to whom the data refers.

2.1. Requirements

In Deliverable D1.1 ,Preliminary Studies and Requirements, 5G Fraud the use cases and its requirements have been identified as follows:

1. **Service Disruption Detection** - For this use case there is the need for developing a service disruption detection that can work as a micro service and is able to monitor the overall health of both Platform and OTT Service ability to provide the service according with the contracted terms.

2. **Platform Service Abuse -** Development of a service abuse detection at the platform level, that can work as a micro service and is able to monitor the monitor subscriber service connection and usage patterns, together with the service contract data, in order to detect situations of abnormal usage that don't comply with the contracted service terms or have the risk to have the CSP incurring in financial or reputational losses. If such a case is detected, relevant data for the analysis should be collected and an alarm generated. Automated actions may be executed for certain scenarios.
3. **OTT Service Abuse -** Development of an OTT service abuse detection, at the platform level, that can make use of an wealthier set of data not at the reach of the OTT Service provider, and that can work as a micro service in order to monitor the connections to the OTT service and usage patterns, together with the service contract data, and spot situations of abnormal service usage that don't comply with the contracted service terms or have the risk to generate losses to either the CSP or the OTT Service provider. If such a case is detected, relevant data for the analysis should be collected and an alarm generated. Automated actions may be executed for certain scenarios.
4. **Platform Capacity Planning -** Development of a capacity planning micro-service, able to collect and process the current capacity indicators for the different technological platform together with the usage pattern, for each site or regional area, and with that information, to forecast the capacity levels and the ability to cope with the predicted usage levels for those sites and regions. Based on this, it should be able to pinpoint problematic areas / technologies / services (if there is a latency problem, only some services are affected). It should also recommend subscribers and services to be migrated, or investment measures to be made to cope with the growing needs.
5. **Customer Experience Segmentation -** A ML based Customer Experience Segmentation approach, that is able to run centrally based on the heavy historical data but with the ability to be adjusted at the Edge with the local usage and infrastructure data stored at the EDGE.
6. **OTT and platform Services Retention -** Development of OTT Services Churn prediction models able to use relevant information stored in the central servers or stored at regional servers or even at the edge, where sensible data that can't be moved to central location may reside, either in near real time or in batch. Models should be adaptive as much as possible to changing patterns, without or with minimal human intervention.

3. Architecture

The AIDA project has the main goal of conceiving a new version of the current RAID platform where some of the pipeline phases can be dynamically moved to the edges of the system. The AIDA general architecture, Figure 3.1, is a distributed architecture with cloud and edge environments that respectively allow global and local processing. It is composed of three vertical components that motivate environments from centralized cloud towards the edge nodes: comprising the orchestration of the distributed edge nodes for better scalability; monitoring/adaptation of the distributed platform to assure that data is collected for problem identification and for the adaptation of the platform; and intrusion detection and tolerance so that AIDA is resilient to intrusion by detecting and tolerating its occurrence. To complement these components, there are two components that ease the interaction between cloud and the edge to assure that communication channels among services are secured as well as the privacy and security of the data is also maintained.

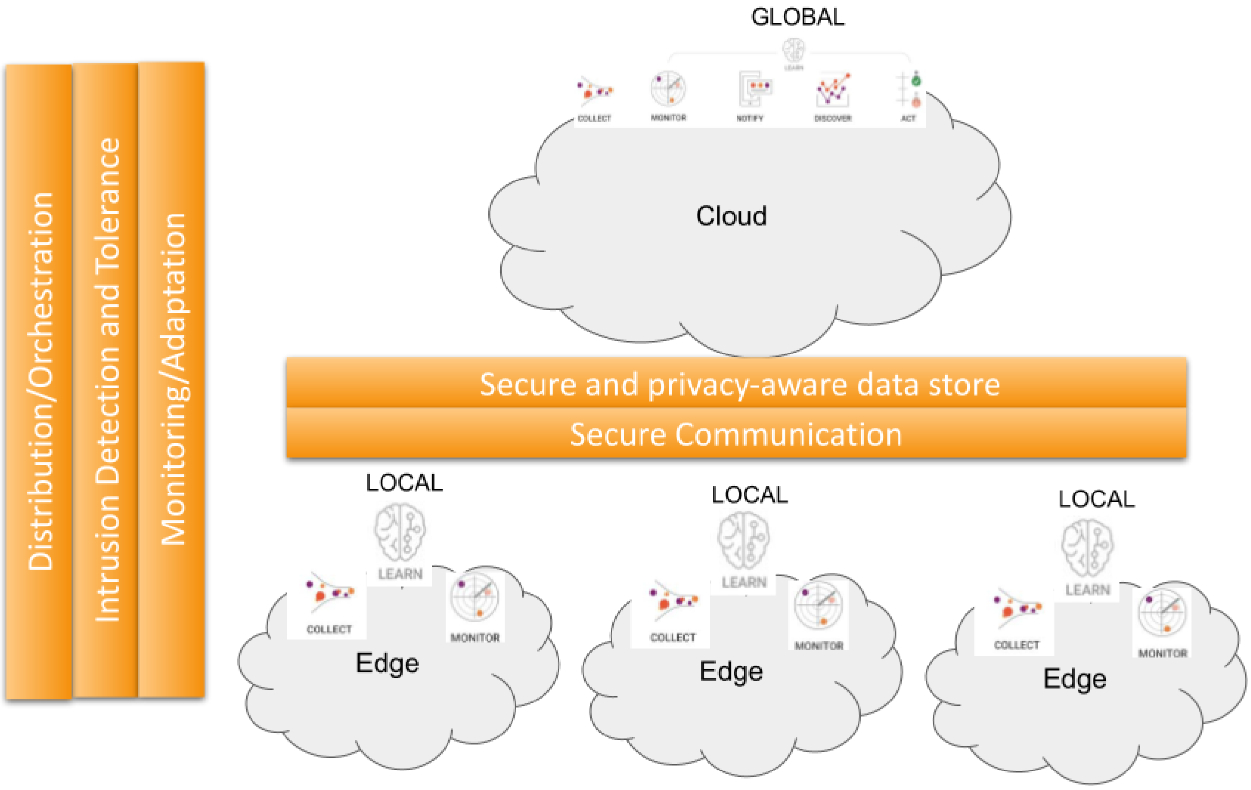


Figure 3.1: AIDA architecture overview.

RAID, Figure 3.2, at the technical level is mainly composed of four main areas:

- a presentation area
- a processing area with one of more RAID Risk Management Solution (RAS/FMS) modules

- a common area
- a build area

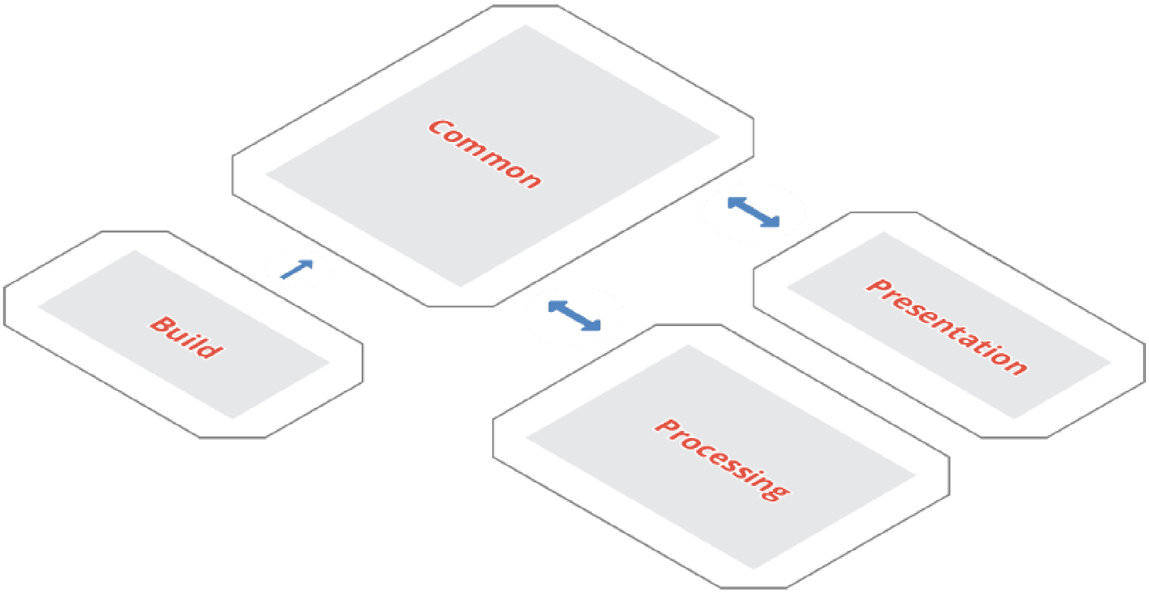


Figure 3.2 - Main WeDo Risk Management Solution areas

The presentation area provides the output interfaces of the solution, namely the solution's UI. The processing area is where input and historical data is processed to generate the necessary outputs. The common area provides all the common services used between the presentation area and the processing area and serves also as the operational platform for the solution. All service communication between the presentation and processing areas is handled in this common area. The build area provides the generation and management of the container images and setup data needed for each presentation and processing blocks or containers.

Taking into account the AIDA general architecture the new version of the current RAID platform is depicted in Figure 3.3, with the four components previously mentioned and the cloud and edge parts detailed. The cloud side of the platform is composed of the main components of the RAID platform divided into the Common, Processing and Presentation areas. In general, this architecture provides clear distinction between the layers across the cloud deployment and actuation area of each service. The Edge side of the platform is composed of lightweight services delegated to the edge of the network, based on containerized applications deployed to provide real-time, low latency service to the customers while assuring higher and more effective scalability capacity.

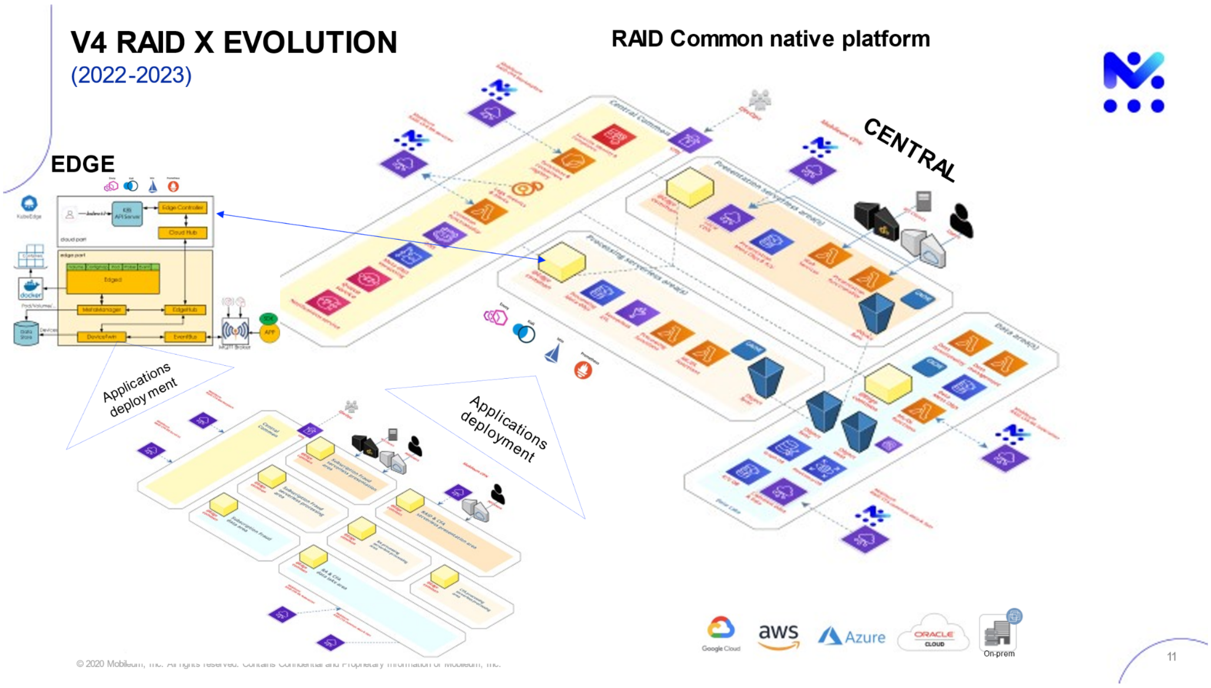

V4 RAID X EVOLUTION
(2022-2023)
RAID Common native platform
EDGE
CENTRAL
Applications deploy ment
Applications deployment
aws
Azure
ORACLE CLOUD
Google Cloud
On-prem
11
© 2020 Mobileum, Inc. All rights reserved. Contains Confidential and Proprietary information of Mobileum, Inc.

Figure 3.3: RAID edge and cloud components distribution overview.

In the following sections we detail the AIDA architecture across different views: data, learning, deployment and security.

3.1. Data

The Edge Computing paradigm aims at leveraging the computational and storage capabilities of edge devices, while resorting to Cloud Computing services for more demanding processing tasks that cannot be done at edge devices. Edge devices generate large volumes of data that may need to be transferred to the cloud and that come from several types of data sources. Therefore, as depicted in Figure 3.4, AIDA includes a polyglot middleware and a synchronization middleware.

The polyglot middleware is used to process data in several different formats, harbouring data hailing from different sources, and thus, often encoded according to different data models. This component should be capable of efficiently querying over an integrated view of the data, regardless of the backing data sources, with transaction isolation.

The synchronization middleware component is in charge of the transfer of data from the edge to the cloud, in an efficient manner, by trying to reduce the volume of data transferred. The synchronization middleware detects differences between the data from edge devices and the data stored in the cloud and synchronizes data while maintaining a balance between extra storage used, processing time, and accuracy.

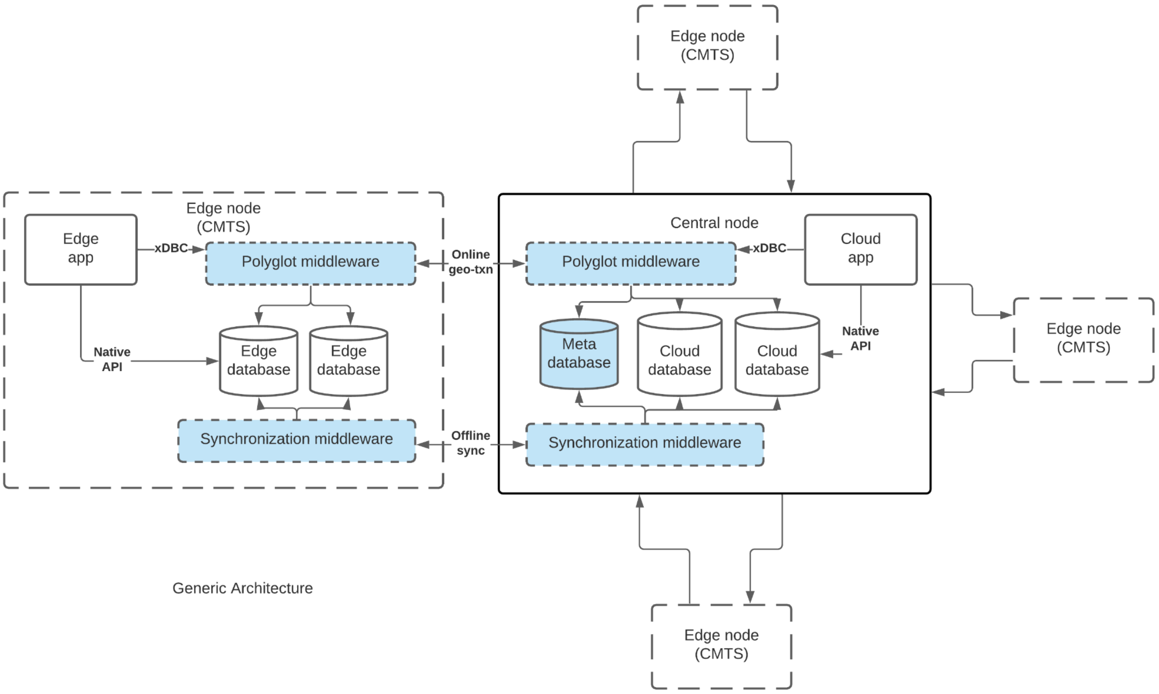


Figure 3.4: Data view of the AIDA platform

3.2. Deployment

The AIDA orchestration and management is responsible for the cluster deployment, with the Cloud cluster and several edge clusters. This implies the deployment of the physical infrastructure to on-premises cloud platforms or public clouds and to deploy the edge nodes and services. Moreover, efficient and secure solutions to orchestrate and monitor the RAID platform in 5G deployments are required, not only to leverage the benefits of edge computing, but also to enhance resilience support in RAID components. AIDA rely on KubeEdge as its interconnection with core kubernetes platforms enables the edge-cloud continuum paradigm. Components providing simple functionalities, for instance, caching services, can be deployed at the edge in KubeEdge pods to provide contents in a more efficient fashion, faster and reducing resource usage at the core network. The resilience aspect is accomplished with the efficient scheduling mechanisms that are able to distribute RAID components according to resource usage or other policies. The monitoring mechanisms provide valuable data for the scheduling algorithms and for the overall purpose of ensuring correctness and security of the diverse micro-services of the RAID platform. The deployment view of the AIDA platform is depicted in Figure 3.5.

KubeEdge is an open source system for extending native containerized application orchestration capabilities to hosts at Edge. It is built upon Kubernetes and provides fundamental infrastructure support for network, application deployment and metadata synchronization between cloud and edge. Kubernetes has become the gold standard for orchestrating containerized workloads running in the data center and public cloud. The control plane of Kubernetes is designed to handle tens of thousands of containers running across hundreds of nodes. This architecture is well-suited to manage scalable, distributed edge deployments. Each edge computing device can be treated as a node while one or more connected devices can be mapped to pods.

In KubeEdge, unlike the nodes of a Kubernetes cluster, edge nodes will have to work in a completely disconnected mode. KubeEdge elegantly tackles this problem through the combination of a message bus and data store that makes edge nodes autonomous and independent. The desired configuration stored in the control plane is synchronized with the local datastore of an edge device which gets cached till the next handshake. Same is the case of the current state of devices persisted in the datastore of the edge device.

For machine-to-machine communication and duplex communication between the edge and the control plane, KubeEdge relies on Mosquitto, a popular open source MQTT broker from the Eclipse Foundation. SQLite is used as the datastore to persist the device twin state and the messages flowing back and forth from the edge to the control plane. WebSockets are used to enable the communication between the edge and the master nodes.

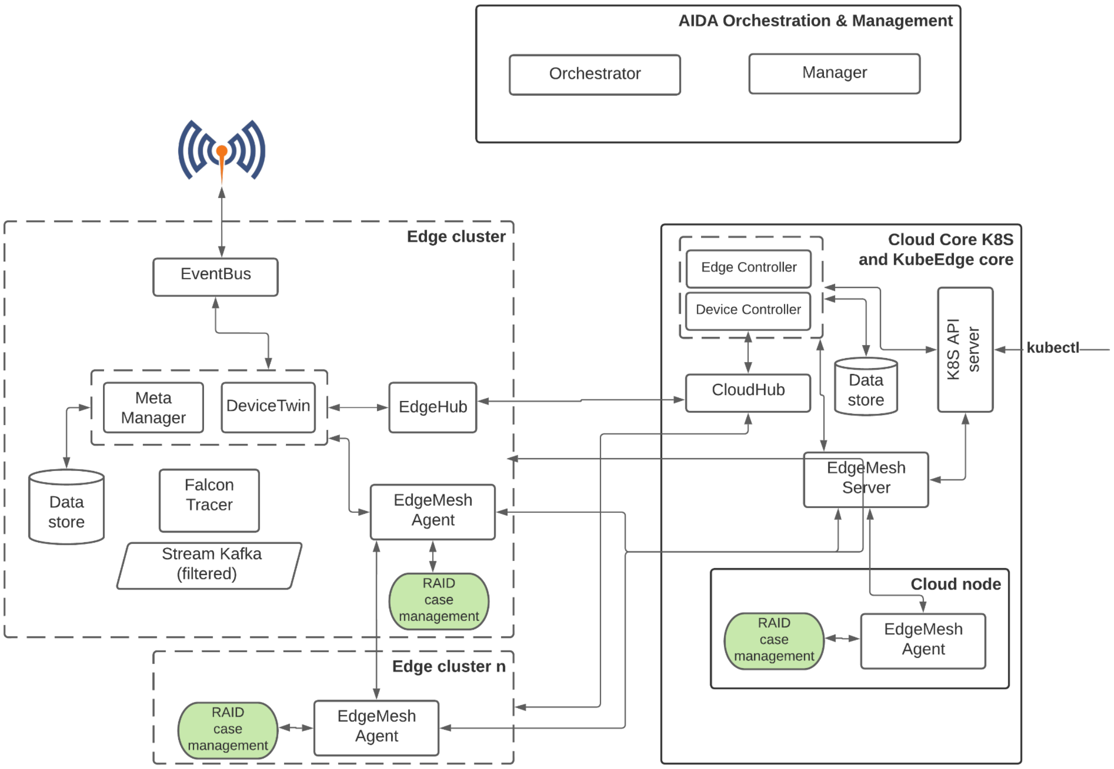


Figure 3.5: Deployment view of the AIDA platform

3.3. Learning

Another key aspect of 5G is the number of new stakeholders in the fraud landscape, which brings new types of fraud that are difficult to anticipate now. This is where the use of AI, especially unsupervised learning algorithms for abnormal behavior detection can help to address the unknown patterns or smart fraud, designed to dissimulate any abnormal patterns and create blind spots, evading detection. In an Integrated Risk Management approach, data feeds that traditionally are not considered in fraud management systems will strengthen the linkage between the different sources enhancing the relations between different domains, like fraud, security, or network fault and performance.

Federated learning is a new approach for distributed machine learning which enables training on a large corpus of decentralized data residing on multiple devices, which often requires computing on the edge. Edge devices should be able to learn from local data and push incremental model updates to coordinator nodes that maintain global models based on the contribution of edge nodes and other relevant data sources.

In AIDA, Figure 3.5, federated learning will be used for the OTT Piracy and DDoS use cases, and the learning and prediction data source is a Kafka stream.

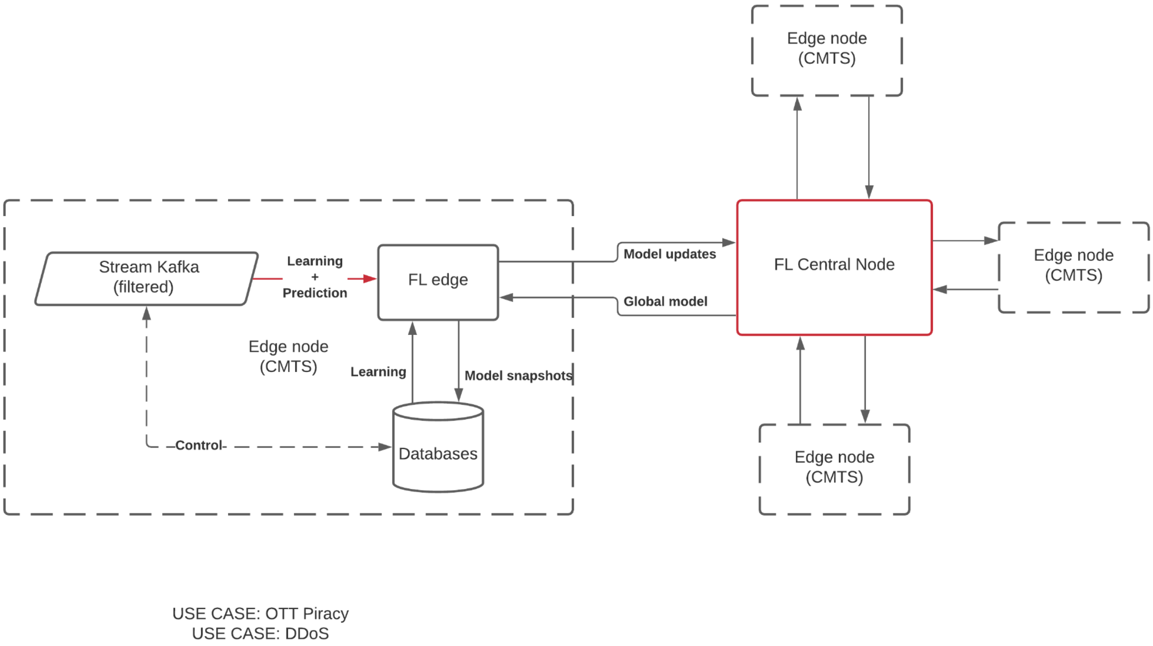


Figure 3.5: Federated learning view of the AIDA platform

3.4. Security

The evolution of the RAID platform during the AIDA project enables RAID to become more effective in scaling to unprecedented levels, but also many security and privacy concerns to be considered. Thus, discovering and implementing measures that address these new risks, while not degrading performance, is of utmost importance. The main challenges are related to the transition to edge, pushing the computational power to the edges of the network, to the integration of 5G supporting multiple tenants and network slicing, and finally to the privacy of the data gathered and analyzed. The security and privacy strategy for the platform, communications and data is detailed in Deliverable 4.1.

Figure 3.6 shows the alignment of the security and privacy strategy with the AIDA architecture. The AIDA orchestration and management component includes a monitoring and self-adaptation solution to secure microservice-based systems in cloud and edge environments. AIDA includes an intrusion detection mechanism for microservice-based systems, both in the edge cluster and cloud nodes, and in AIDA, intrusion tolerance component provides operation capabilities in the presence of security intrusions. The AIDA platform will include multiple microservices running in different environments (cloud, edge), requiring secure exchange of data. Secure communications are required to assure privacy and data integrity between two communicating microservices. This also includes scalable authentication and authorization mechanisms. Regarding data privacy the platform will provide a distributed and privacy-preserving framework for machine learning workloads and a privacy framework for heterogeneous data types.

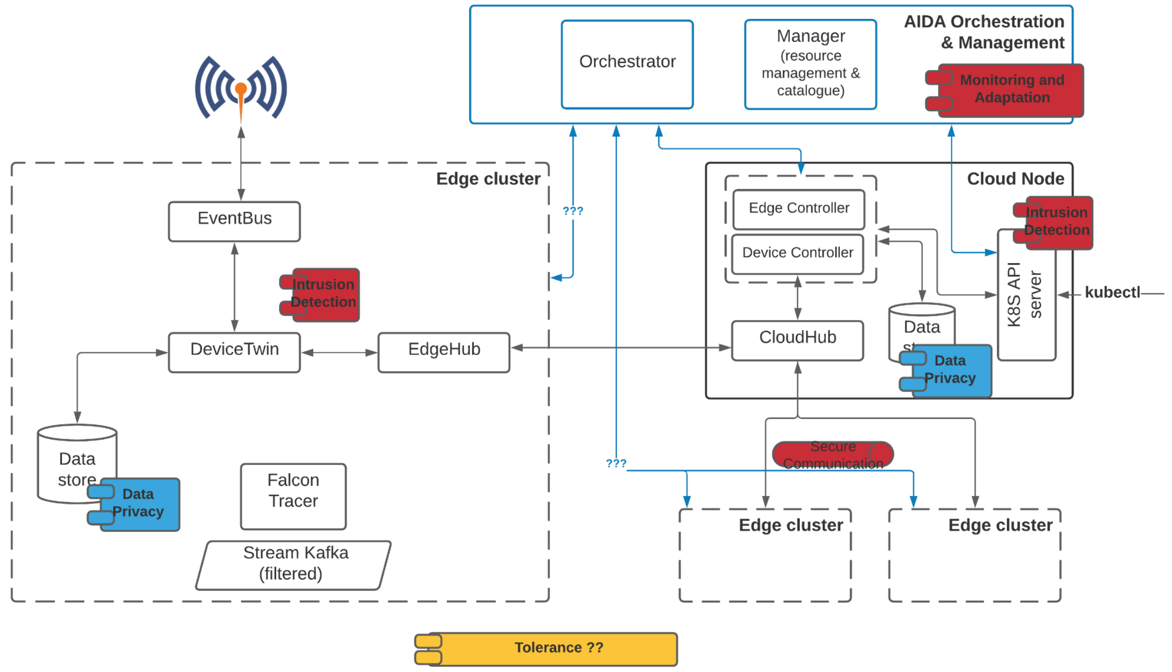


Figure 3.6: Security view of the AIDA platform

4. APIs

RAID platform exposes three main components to the exterior, namely the presentation layer, the web services and the Extract, Transform and Load (ETL) layer. The presentation layer serves the users of the platform with user-friendly interfaces, following the Rich Internet Applications (RIA) architecture with state-of-the-art technologies, for the creation, management, and deletion of the objects mentioned above. The users can navigate through the pages available to perform the intended operations. The web services are available to the clients that feed the system with data, communicating through REST and using the Tomcat server as support. With regard to the ETL layer is where the processing of the information collected is conducted which can also partially occur at the edge of the network, as a consequence this layer is also exposed to external agents that can feed the information to get it processed.

Multi-access Edge Computing enables the implementation of MEC applications as software-only entities that run on top of a virtualisation infrastructure, which is located in or close to the network edge. The multi-access edge system consists of the MEC hosts and the MEC management necessary to run MEC applications within an operator network or a subset of an operator network. ETSI ISG MEC specifies Multi-access Edge Computing technologies, and in particular, a set of APIs that allows applications that are virtualized in the edge, to access network and users information from the local node. AIDA architecture and APIs were designed in order to be compatible and interoperable with ETSI ISG MEC APIs that may exist at the telco operators.

As listed in Section 2.1 and detailed in Deliverable D1.1, AIDA focus on 5G fraud use cases. In order to automate how to prevent and stop the deliberate act to obtain an unauthorized benefit by using unethical means, we defined 6 groups of service APIs for AIDA: case, alarm, alert, edge monitoring, edge performance, and model. The APIs are depicted in Figure 4.1.

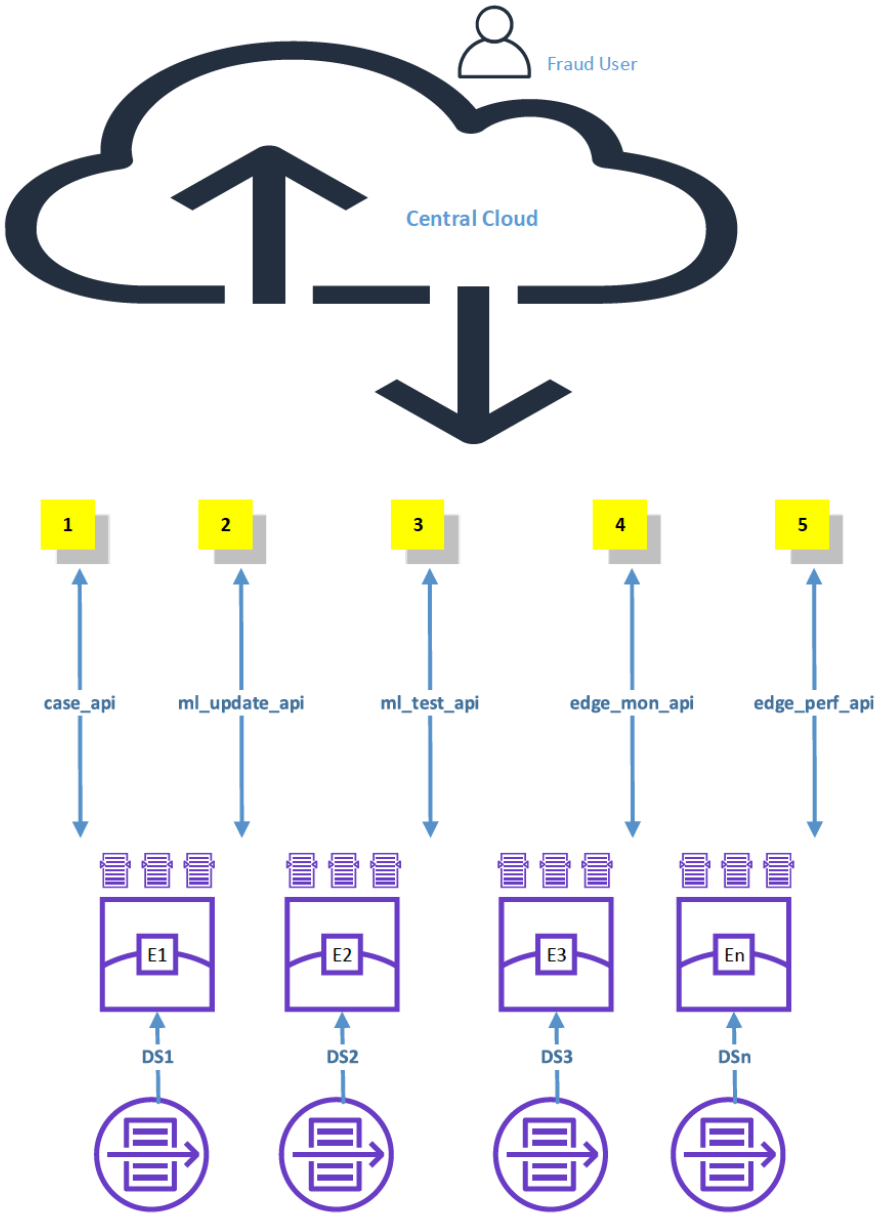


Figure 4.1: AIDA service APIs

- case: A core component of RAID is case management, which helps fraud examiners manage their workloads and digital information. Cases are opened in the cloud by each edge location, implying the transmission of details which will allow the analyst to review and classify it. This API includes the following operations:
  - c_open_case to open a new case from the cloud;
  - c_reinforce_case to reinforce an existing case from the cloud;

  - e_review_case to review an existing case from the edge;
  - e_check_status to check the status from edge.
- ml update: Each edge location will have its own model trained with local data, this federated model will be shared with the cloud whenever it is updated . The cloud will produce a new model with the combined model and share with the multiple edge locations. This API includes the following operations:
  - e_check_model_vs;
  - c_check_model_vs;
  - c_update_model_cs;
  - e_update_model_vs.
- ml test: Monitor the performance of the machine learning model and benchmark with previous versions identifying performance issues , degradation or possible poisoning attacks. This API includes the following operations:
  - e_test_model;
  - e_benchmark_model;
  - e_revert_model;
  - c_promote_model.
- edge monitoring: Monitor the health of the edge location resources ensures that its live and performing under expected thresholds, allowing automatic or remote configurations to be deployed, including reference data. This API includes the following operations:
  - e_health_status;
  - e_check_backlog;
  - e_remote_adm;
  - e_update_refdata.
- edge performance: Each edge location needs to be monitored in terms of business performance, eg : there are no data feeds to be processed , or they have wrong formats , how many cases per minutes are opened , closed , high variations on trends or silence systems even though no errors being reported. This API includes the following operations:
  - a

The Figure 4.2 details the data each edge cluster needs to maintain in order to provide the 6 service APIs.

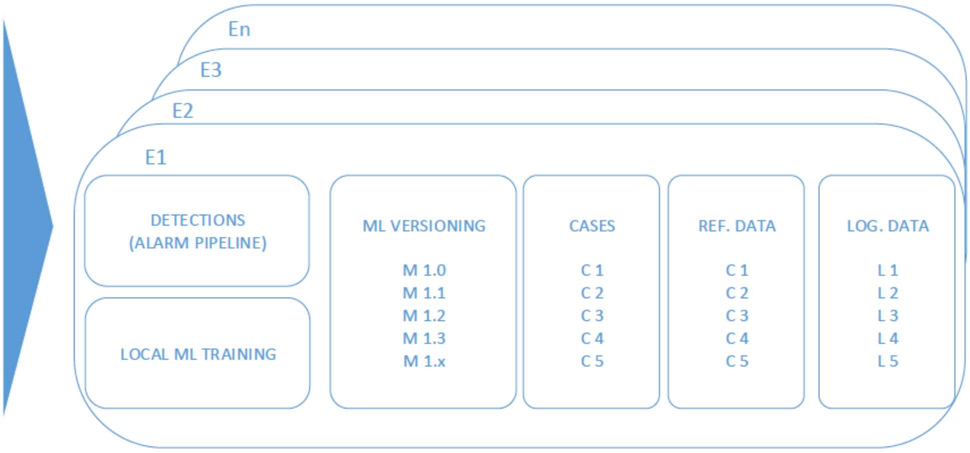


Figure 4.2: AIDA service APIs edge data

5. Conclusions

This deliverable details the architecture and interfaces of the AIDA platform. As discussed throughout the document the proposed design will a new version of Mobileum's RAID platform enabling: the distribution of the platform to encompass the cloud and the edge platforms; utilization of distributed and federated machine learning to enable model training on distributed data and to learn with both behavior patterns and context evolution; and also assure security of the platform, through intrusion detection and tolerance.

As another contribution, the deliverable discusses the internal architectures and interfaces for AIDA components: Orchestration, Intrusion Detection and Tolerance, Monitoring/Adaptation, Secure and privacy-aware data store, and secure communication. Also, it shows how these components interact and integrate with each other.

The information contained in this document will drive the next major step of the project, namely, the implementation of the AIDA platform.

COMPETE 2020 PORTUGAL 2020 UNIÃO EUROPEIA Fundo Europeu de Desenvolvimento Regional FCT Fundação para a Ciência e a Tecnologia Carnegie Mellon Portugal