OEHI Workshop

Edge AI with Sedna

In this talk, we share our experience with Sedna and KubeEdge to deploy AI applications on heterogenous edge-cloud infrastructure. We present our multi-edge computing model which was applied to different, real-life use-cases. For demonstration, we select the object re-identification problem (ReID) as our AI application to showcase our system design, workflow, and implementation.

HUAWEI

Edge AI with Sedna

About Us – Huawei MRC Edge AI Team

Our Research

We focus on solving the challenges of managing, deploying, and monitor AI applications on an heterogenous infrastructure. To do so, we build on top of well known components such as Kubernetes, Kubeedge, and Sedna.

Edge-Cloud Synergy

We optimize the use of edge and cloud resources to support flexible deployment of AI workloads based on user requirements.

AI Applications

Supports different classes of AI applications such as object detection, classification, re-identification (ReID).

Customizable

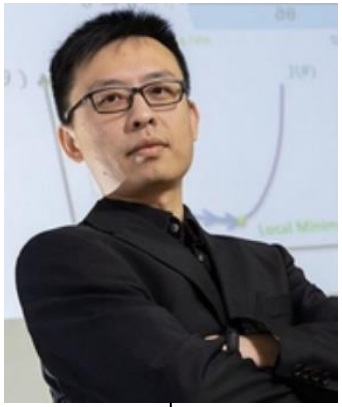
Fully customizable by offering ML practitioners the ability to add new models to the framework and creating new ML workloads on the fly, from scratch.

Open Source

Based on fully open-source projects, our code is aswell fully available on GitHub.

Vittorio Cozzolino
PhD, Senior Software Engineer
vittorio.cozzolino@huawei.com

Zi Yang
PhD Candidate
zi.yang1@huawei.com



Haojin Yang
PhD, Scientific Advisor
yang.haojin.ext@huawei.com

Soumajit Majumder
PhD, Senior AI Researcher
soumajit.majumder@huawei.com

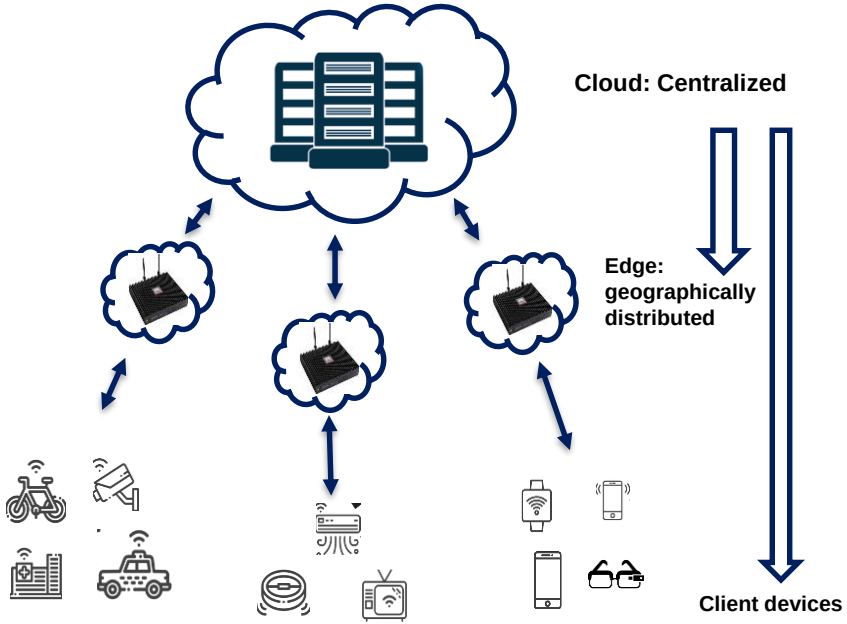
Jorge Cardoso
PhD, Chief Architect
jorge.cardoso@huawei.com

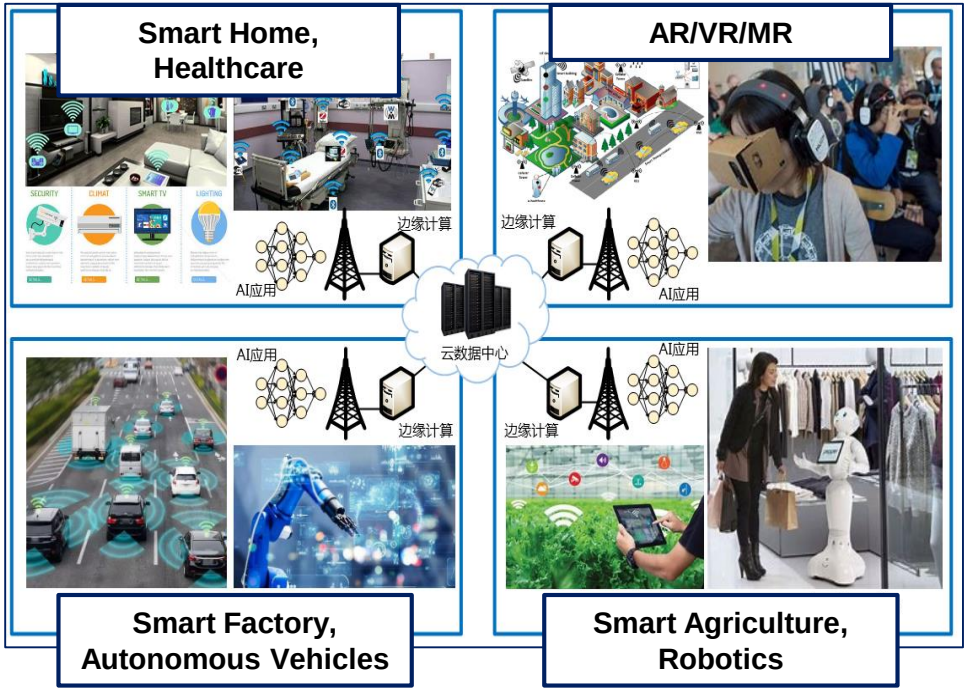
HUAWEI

Edge AI with Sedna

Why Edge AI?

The cloud, centralized AI computing paradigm cannot cope with the demand for real-time, accuracy, and strong interactivity of AI applications at the edge.





HUAWEI

Edge AI with Sedna

Edge AI Building Block

Kubernetes

- Open-source
- System for automating deployment, scaling, and management of containerized applications
- De-facto containerized application orchestration framework.

KubeEdge

- Open-Source
- Kubernetes Native API at Edge
- Seamless Cloud-Edge Coordination
- Edge Autonomy
- Low Resource Ready
- Simplified Device Communication
- Heterogenous

Sedna

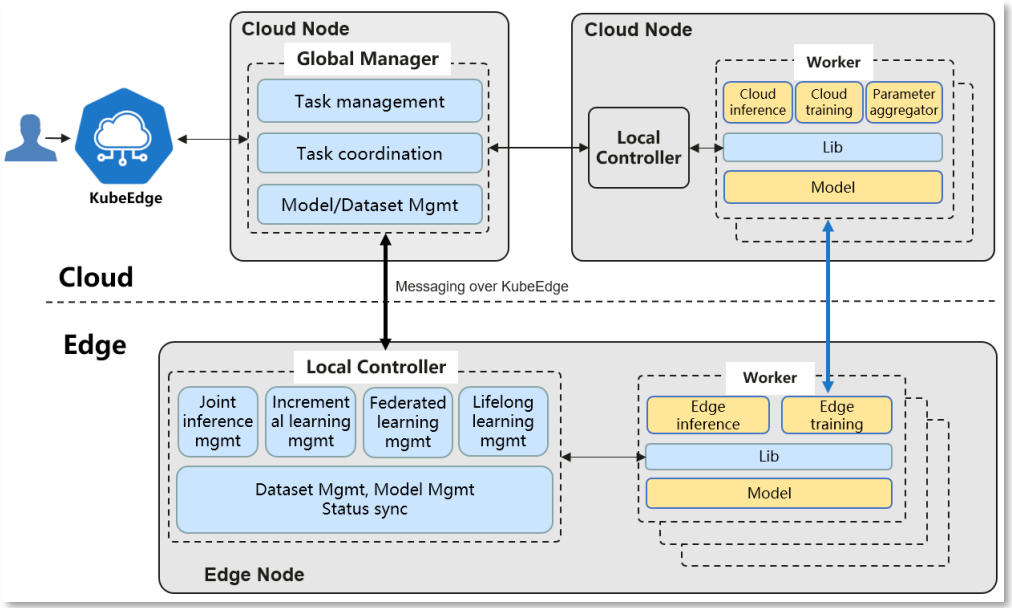
- Open-Source
- Unified orchestration of across edge-cloud applications
- Edge-cloud synergy AI framework
- Both training and inference frameworks
- Lifelong learning, multi-edge reasoning, joint inference etc..

HUAWEI

Edge AI with Sedna

Sedna – Distributed Collaborative AI

- Opensource project aimed at becoming a repository of AI models and algorithm to be easily reused and deployed in a distributed fashion.
- Based on the edge-cloud collaboration capability provided by KubeEdge (or Kubernetes).
- Supports seamless integration of existing AI applications at the edge.
- Compatible with mainstream AI frameworks such as TensorFlow, Pytorch, PaddlePaddle, and MindSpore.
- Reduces build and deployment costs by reuse of algorithms and models existing in the framework.
- Enables data privacy with local processing of sensitive data directly at the edge.
- https://github.com/kubeedge/sedna



AI Paradigms

- Joint Inference
- Incremental learning
- Federated Learning
- Lifelong learning
- Multi-edge AI

Compatibility

- Mainstream AI frameworks
- Module algorithms
- Extensible interfaces

Framework

- Dataset management
- Model management
- ……

HUAWEI

Edge AI with Sedna

Sedna – Components

GlobalManager

- Unified edge-cloud synergy AI task management
- Cross edge-cloud synergy management and collaboration
- Central Configuration Management

LocalController

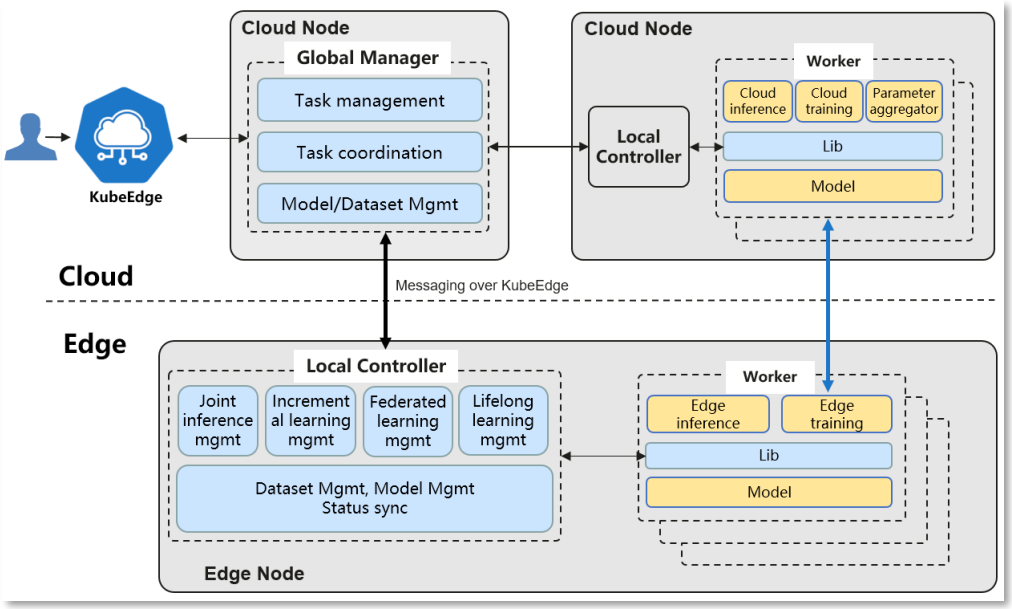
- Local process control of edge-cloud synergy AI tasks
- Local general management: model, dataset, and status synchronization

Worker

- Do inference or training, based on existing ML framework.
- Launch on demand, imagine they are docker containers.
- Different workers for different features
- Could run on edge or cloud

Lib

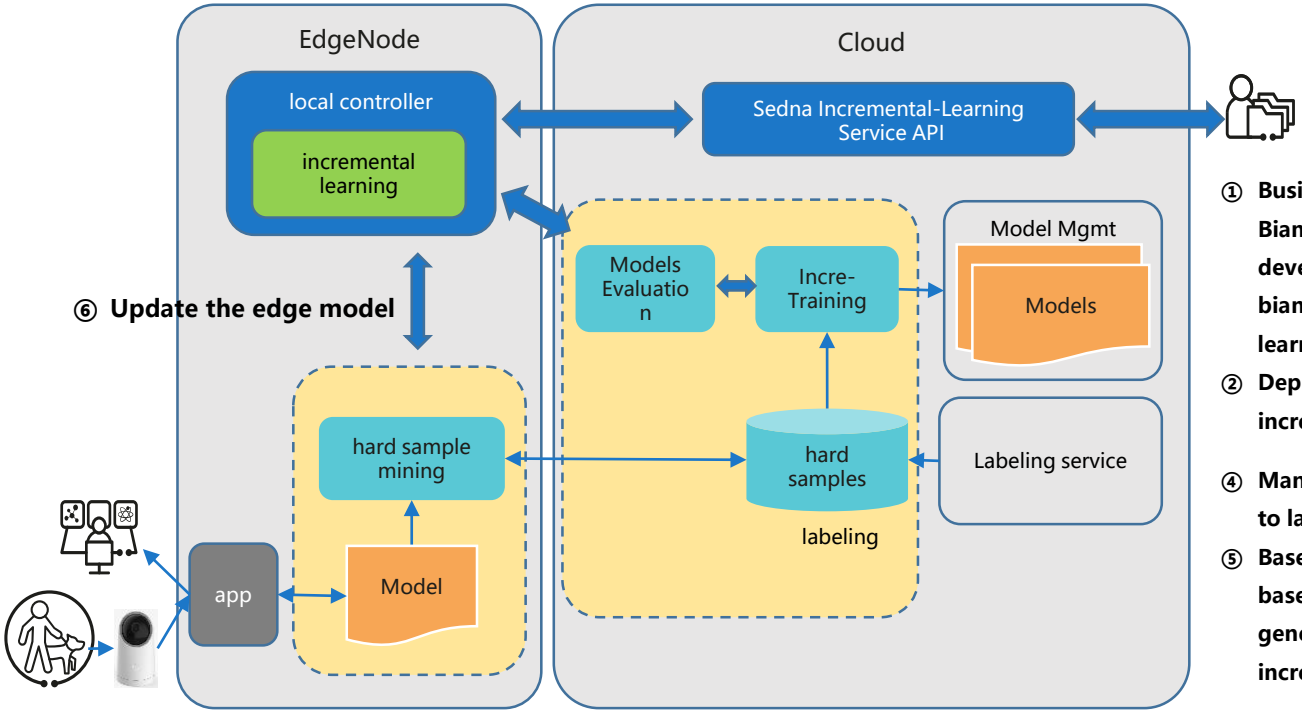
- Expose the Edge AI features to applications, i.e. training or inference programs



HUAWEI

Edge AI with Sedna

Sedna Applications: Edge-Cloud Collaborative Incremental Learning



① Business APP developers: Use the Bianyun AI Lib library during development, and integrate the bianyun collaborative incremental learning function.

② Deploy the service APP and start incremental learning.

④ Manual intervention periodically to label the sample.

⑤ Based on the task knowledge base, the system automatically generates new models for incremental training

③ The APP runs, identifies difficult cases, and uploads them to the cloud annotation service

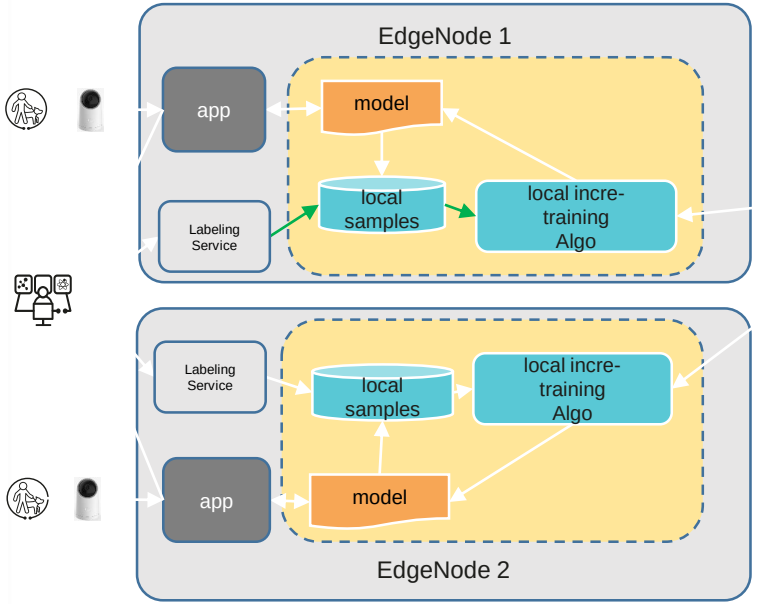
HUAWEI

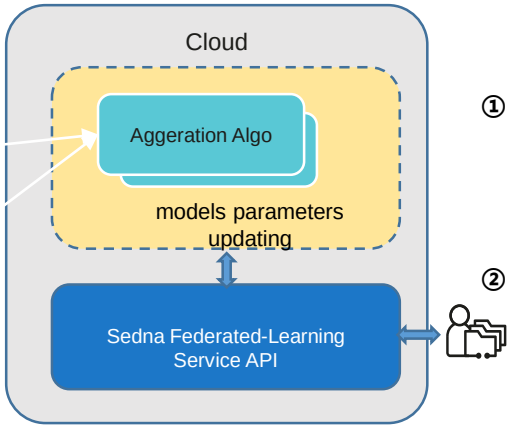
Edge AI with Sedna

Sedna Applications: Cloud-Side Collaborative Federated Learning

③ **Multi-task detection, divide non-IID sample sets, and cooperate with the cloud to identify similar tasks**

④ **For local training, model parameters are uploaded to the cloud, and cross-edge migration + model aggregation algorithms are run in the cloud.**





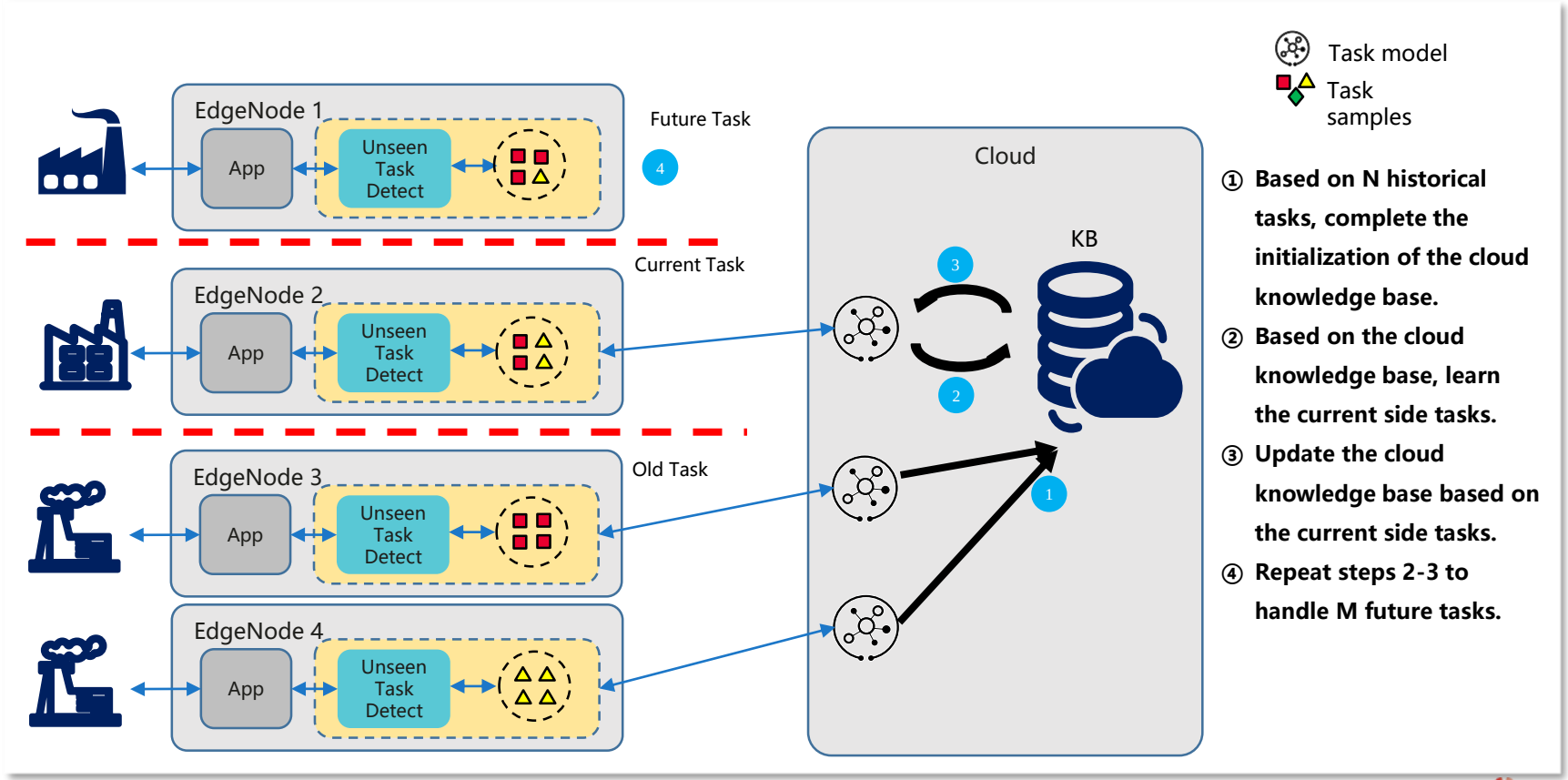
① **Developers: Import edge AI Lib libraries and develop edge cloud collaborative federated learning programs.**

② **Launch a federated learning task and deploy the training program to the edge**

HUAWEI

Edge AI with Sedna

Sedna Applications: Collaborative Lifelong Learning



① **Based on N historical tasks, complete the initialization of the cloud knowledge base.**

② **Based on the cloud knowledge base, learn the current side tasks.**

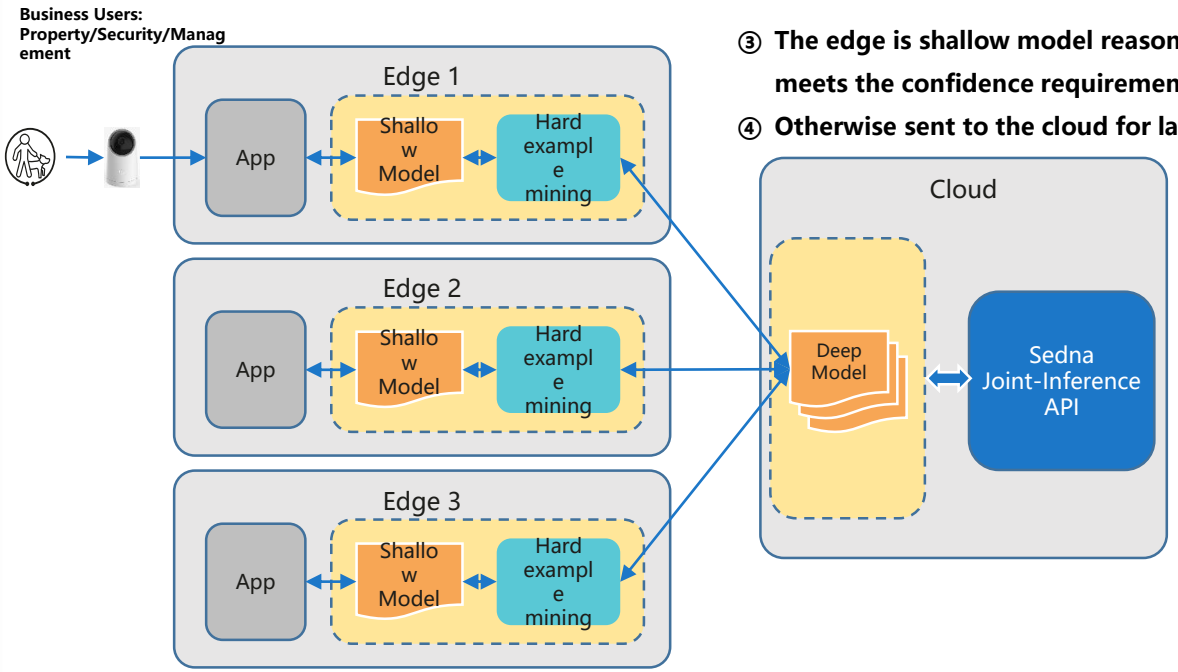
③ **Update the cloud knowledge base based on the current side tasks.**

④ **Repeat steps 2-3 to handle M future tasks.**

HUAWEI

Edge AI with Sedna

Sedna Applications: Edge-Cloud Joint Inference



③ The edge is shallow model reasoning, and the judgment meets the confidence requirements is returned directly

④ Otherwise sent to the cloud for large model inference

① AI developers: Provide data training to generate deep and shallow models

② Business developers: Invoke the synergy model through the Lib library and deploy it to the edge

HUAWEI

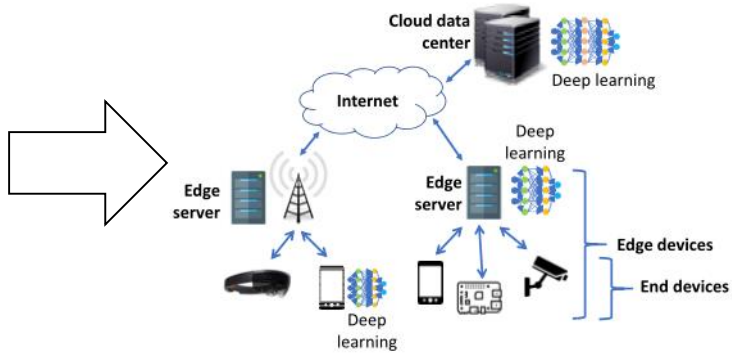
Edge AI with Sedna
Our Application

Applications

- **Object tracking, Computer vision**, Cloud Gaming, Augmented reality, and IoT use cases can be realized with edge computing architectures with machine learning (ML) techniques.
- Develop an Edge AI approach for automatic object tracking and identification in multi-camera systems.
- Comprehensively understand and analyze the use case and business problem.

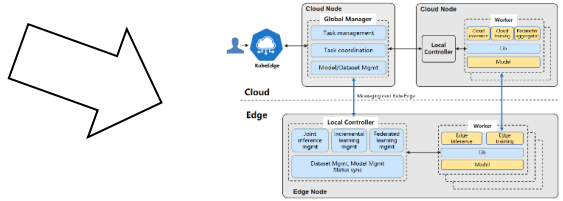
System

- To support the applications, data needs to be collected, stored, processed, analyzed, and visualized across the **distributed components of edge platforms using AI.**
- Approaches are needed to support the training, deployment, and inferencing ML models in edge computing environments.
- Furthermore, ML techniques need to be optimized to provide services with low latency for end users.



Sedna Integration

- Develop a framework to systematize and drive the implementation of technology selections and to optimize end-to-end system for the computing continuum
- Translate new computing models into new functionalities of the software framework **Sedna**



Edge AI with Sedna
Application Scenarios

Smart Port

Monitor the location of **shipment containers** in a smart port to boost logistic operations and allow quick retrieval of potentially lost goods.

When a shipment container enters a monitored area, it's **detected**, **recognized** and **localized.**

Smart Campus

Locate **pedestrians** in a smart campus for occupancy monitoring and to offer personalized services.

When a pedestrian enters a monitored area, it's **detected**, **recognized** and **localized.**

COVID-19

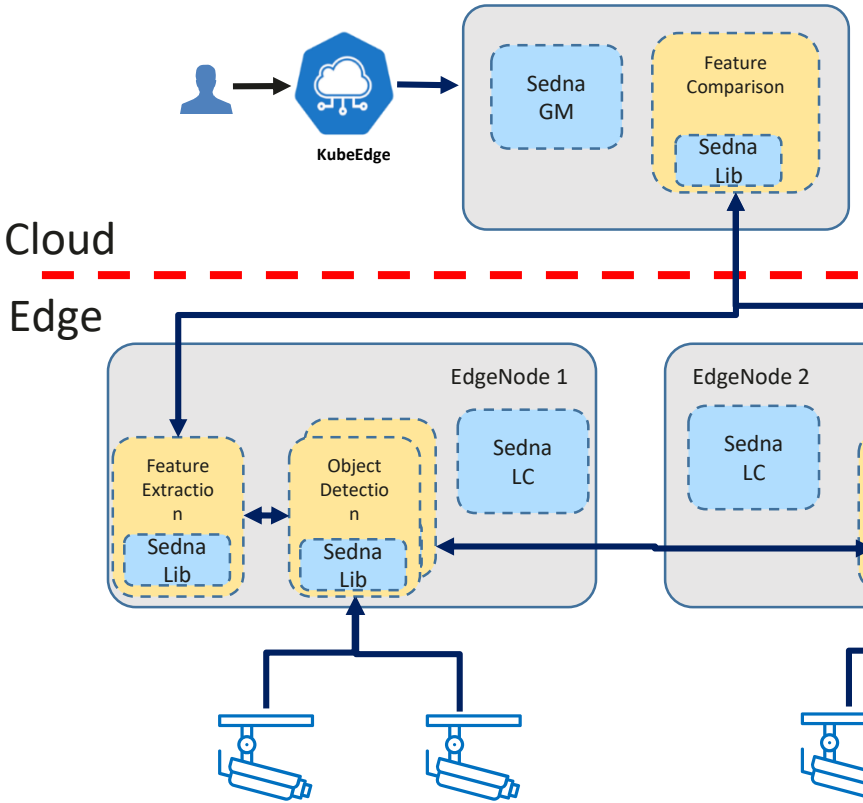
Find proximity contacts between a potential carrier of the virus and other people to tackle the current pandemic crisis.

When a carrier gets in contact with someone in a monitored area, the system will identify and notify the subject of the potential risk of infection**.**

HUAWEI

Edge AI with Sedna

Multi-Edge-Cloud Reasoning with Workload Repartition

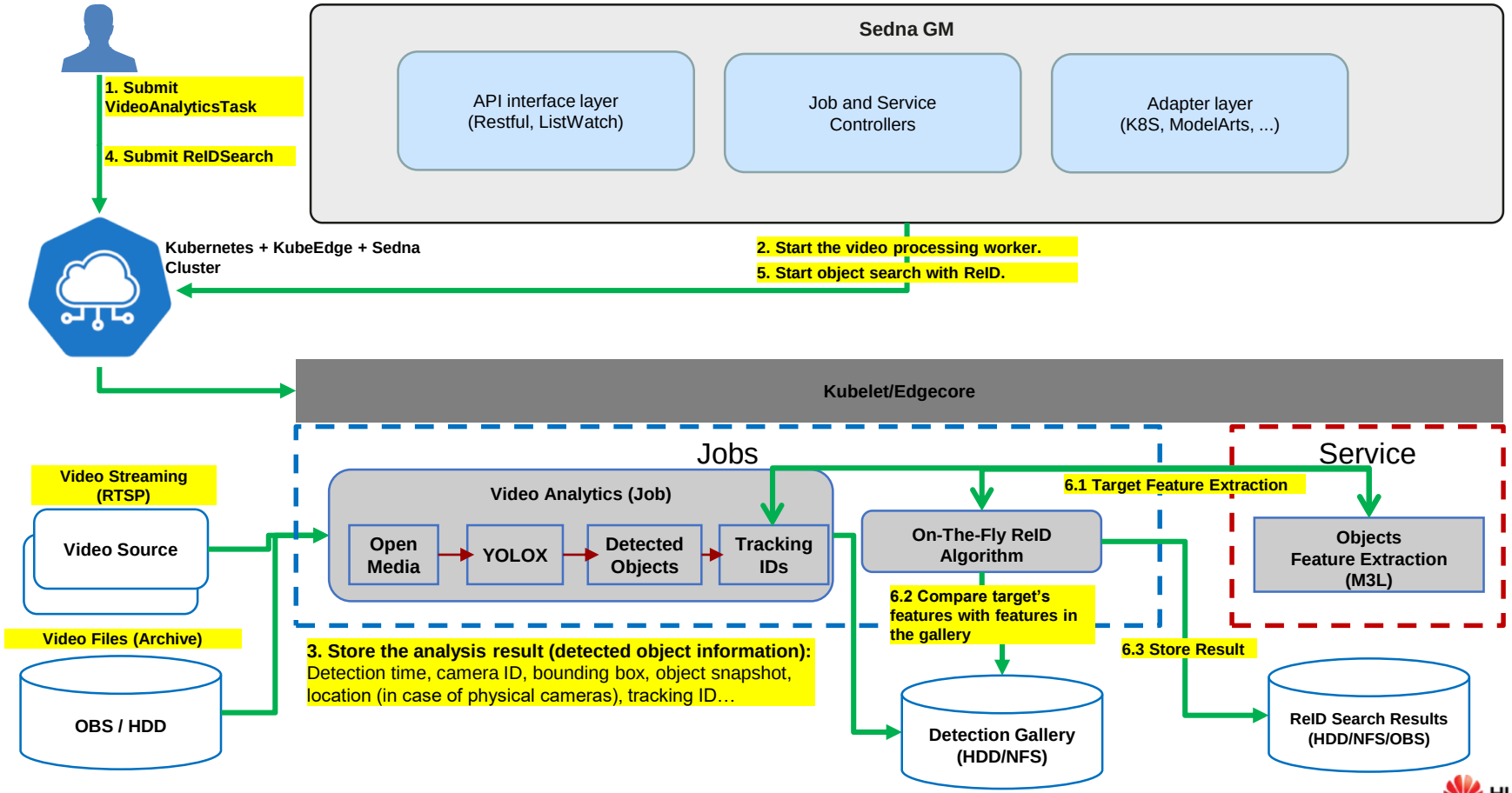


② AI algorithms running at different edge nodes can be flexibly unloaded, including offloading to other edges or the cloud, to improve overall inference efficiency.

① Multiple cameras deployed at different edge nodes can provide multi-dimensional information with different viewing angles, different times, and different lighting to improve the overall recognition accuracy.

HUAWEI

Edge AI with Sedna
System Design

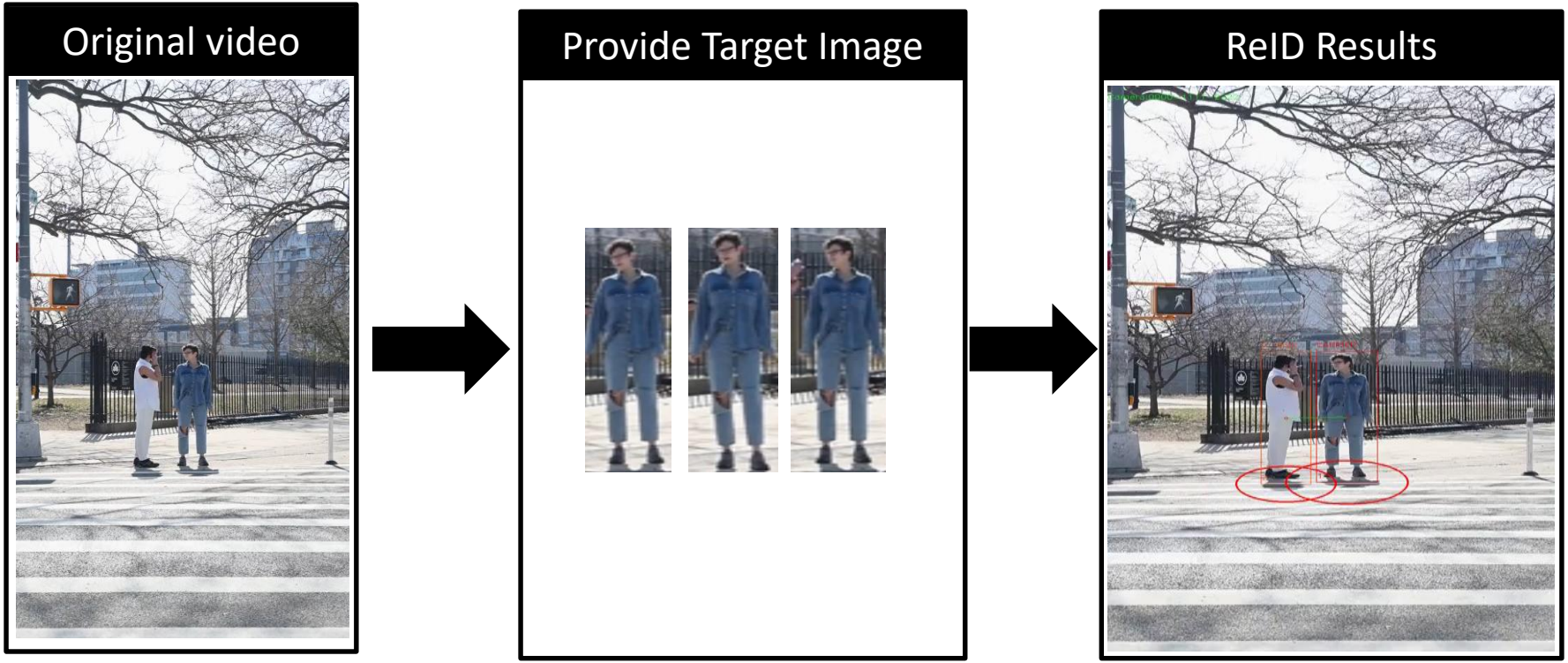


HUAWEI

Edge AI with Sedna

ReID in Practice

ReID can be applied the any domain, given that the feature extraction network is properly trained. In the example below, we show how it can be applied to humans due to the vast amount of data for this category.



Video from: https://www.pexels.com/video/two-people-standing-by-the-pedestrian-crossing-while-talking-7263169/

HUAWEI

Edge AI with Sedna
Hands-On

Getting started:

- Login into your VM
- Pull the training repo @ https://github.com/haicgu/training
- Open the README.md

HUAWEI