Red Hat Research Days
presents

AIDA
A holistic AI-driven networking and processing framework for industrial IoT applications

September 21st | 3:00 - 4:30 PM CEST

Speaker
Anna Brunstrom
Karlstad University

Speaker
Muhammad Usman
Karlstad University

Speaker
Bestoun Ahmed
Karlstad University

Conversation Leader
Toke Høiland-Jørgensen
Red Hat

Conversation Leader
Simone Ferlin-Reiter
Red Hat

Contents

- Brief introduction to Karlstad University
- Background and overview of AIDA
- Distributed observability framework
- ML pipeline

Quick Facts about Karlstad University

- 19,000 students
- 260 doctoral students
- 1,500 staff
- Established 1999
- Teacher education since 1843
- Excellent research groups
  - **Computer Science (CS)**
  - Service Research Centre (CTF)

Computer Science

- 800 students
- ~60 staff
  - 20+ doctoral students
- Research profiles
  - Distributed systems and communications (DISCO)
  - Privacy and security (PriSec)
  - Software quality and digital modernization (SQuaD)

Our employees come from eighteen countries around the world and represent four continents.

Speaker
Anna Brunstrom
Karlstad University

AIDA

A HOLISTIC AI-DRIVEN NETWORKING AND PROCESSING FRAMEWORK FOR INDUSTRIAL IOT APPLICATIONS

RED HAT RESEARCH DAYS 2023-09-21

KARLSTADS UNIVERSITET

ERICSSON

Red Hat

tietoevry

EL & STADSNÄT

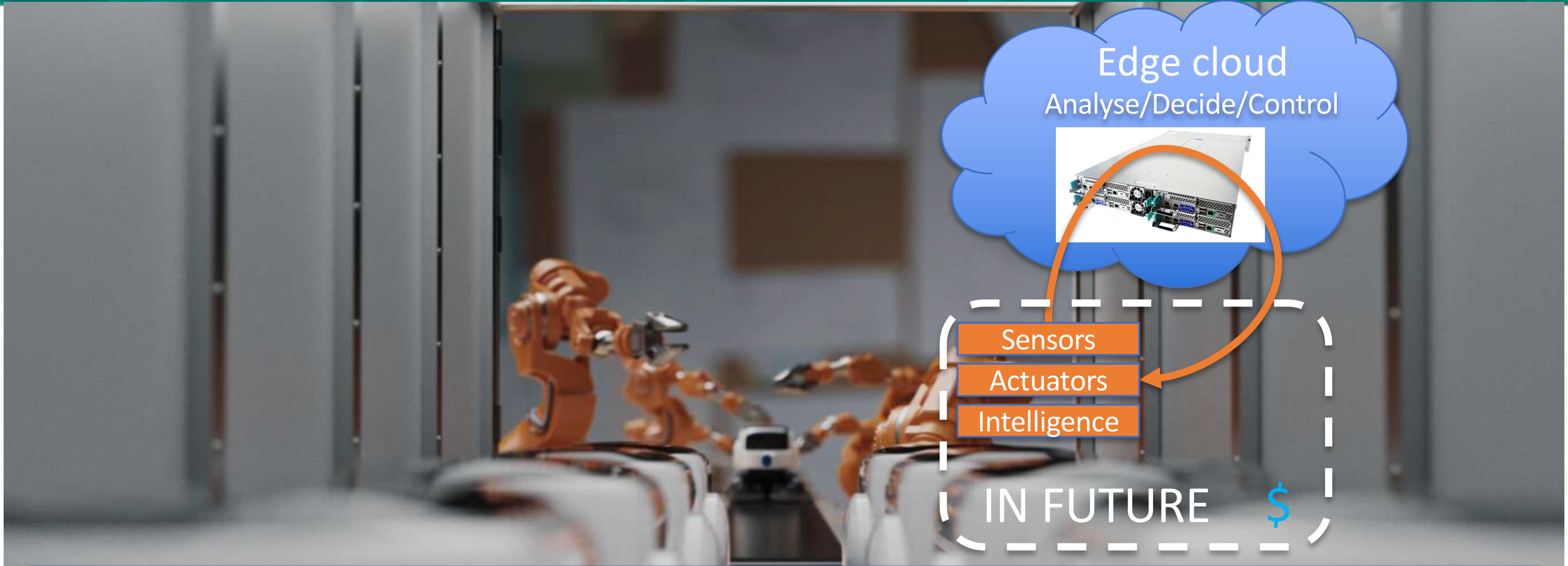
UDDEHOLM

TIME CRITICAL NETWORKS

CONNECTED CYBER-PHYSICAL SYSTEMS → NOW

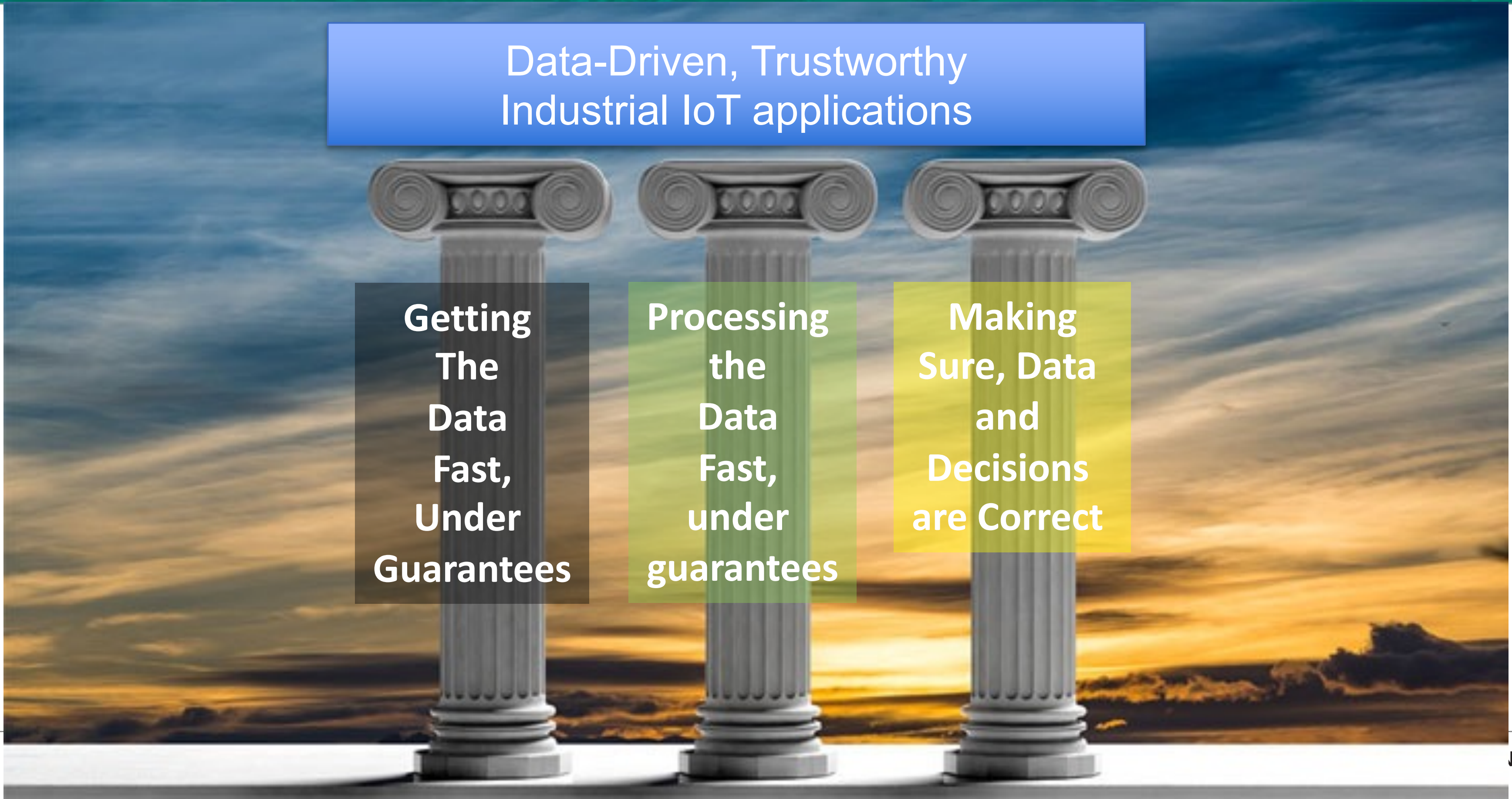
- Needs/Trends:
  - Collecting and Making use of billions of sensor data → IoT
  - Analyzing data and acting upon it in Real-time → Analytics
  - Autonomous Decisions guided by algorithms → ML

CONNECTED CYBER-PHYSICAL SYSTEMS → IN FUTURE



- Characteristics and Benefits
  - In software, virtualized, programmable, upgradable, commodity infrastructure, open, interoperable, customizable
  - Increase flexibility, reduce deployment time and cost

3 MAIN PILLARS FOR TRUSTWORTHY I-IOT APPLICATIONS



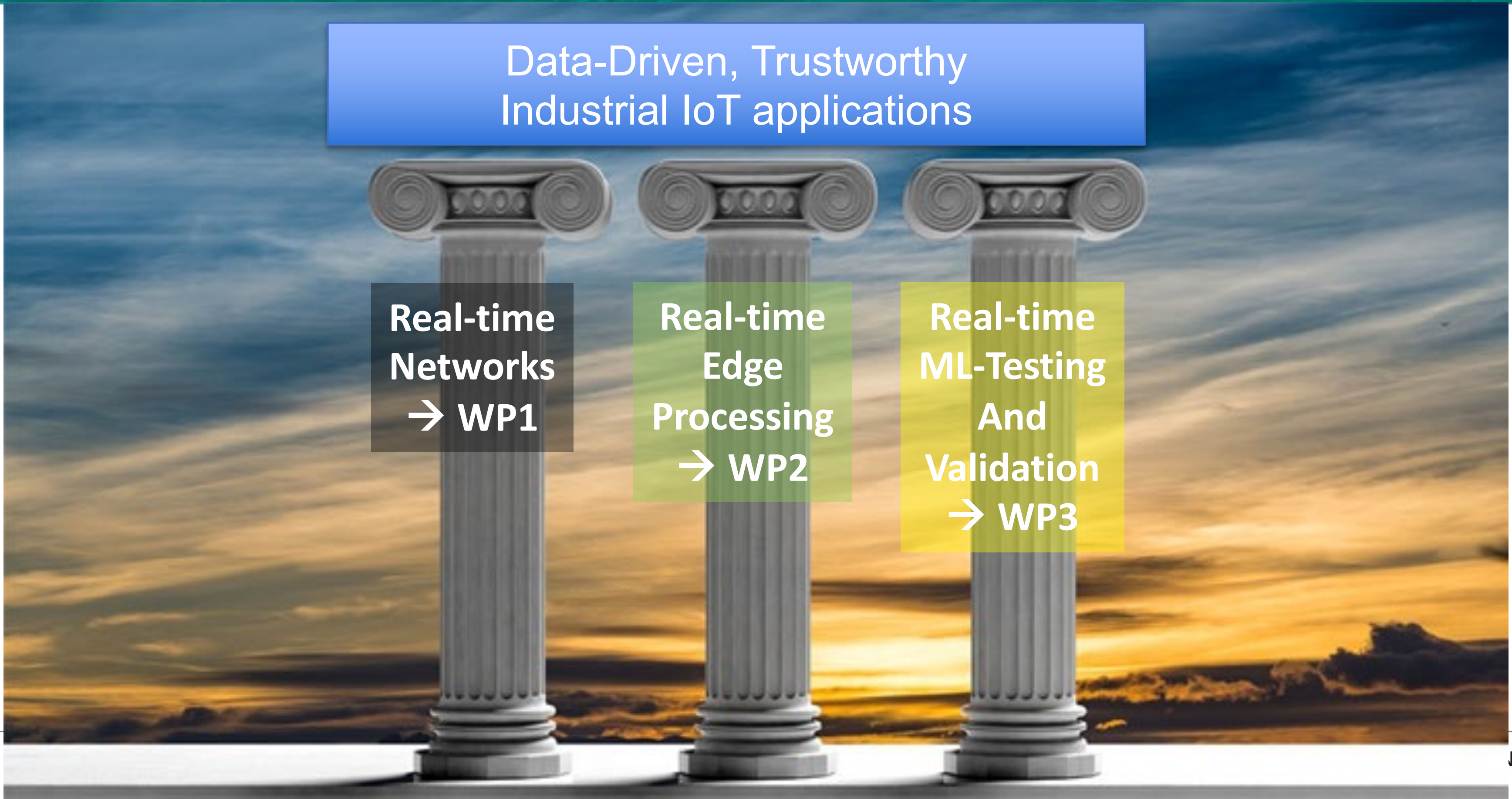
Data-Driven, Trustworthy Industrial IoT applications

Getting The Data Fast, Under Guarantees

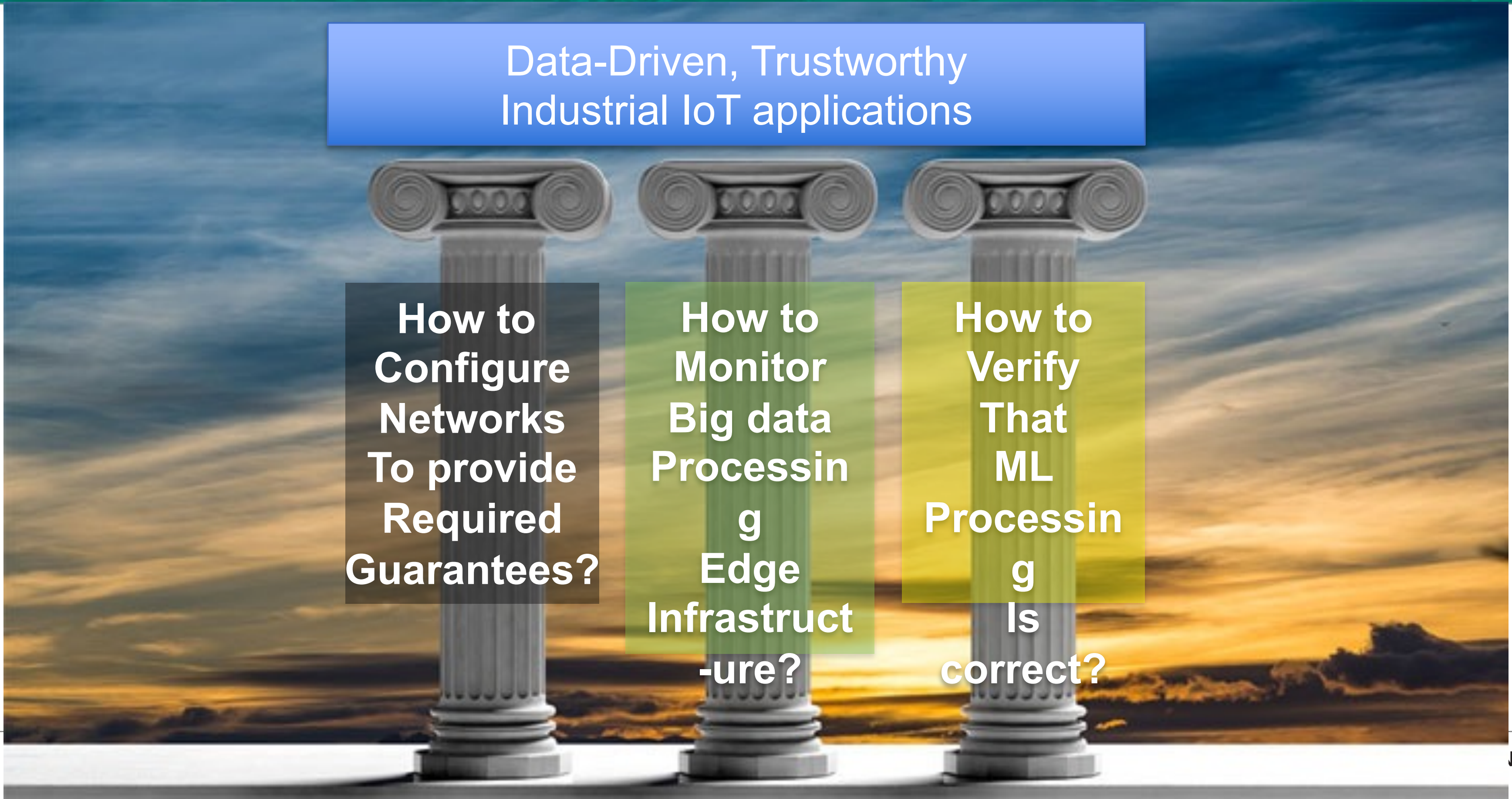
Processing the Data Fast, under guarantees

Making Sure, Data and Decisions are Correct

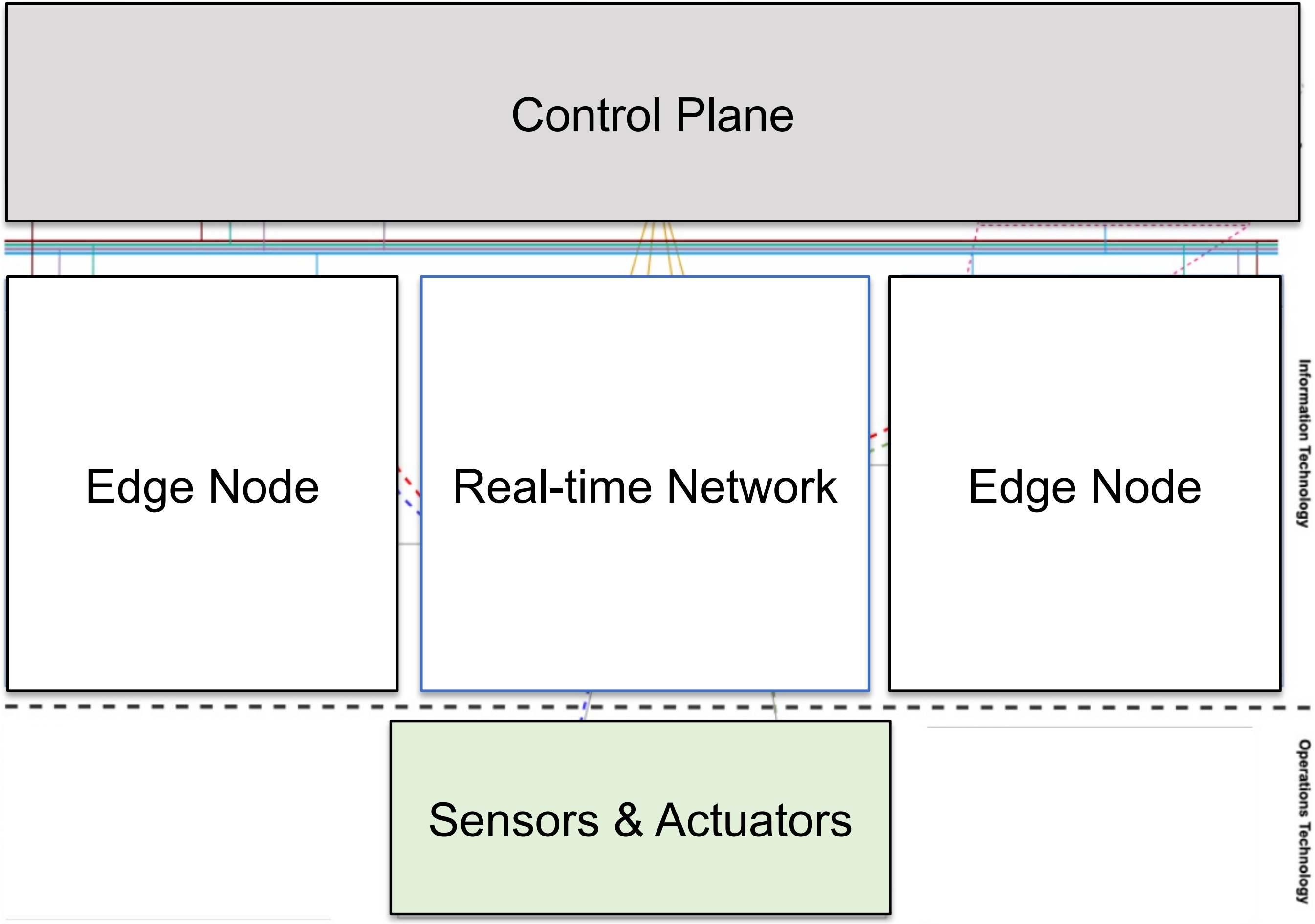
3 MAIN PILLARS FOR TRUSTWORTHY I-IOT APPLICATIONS



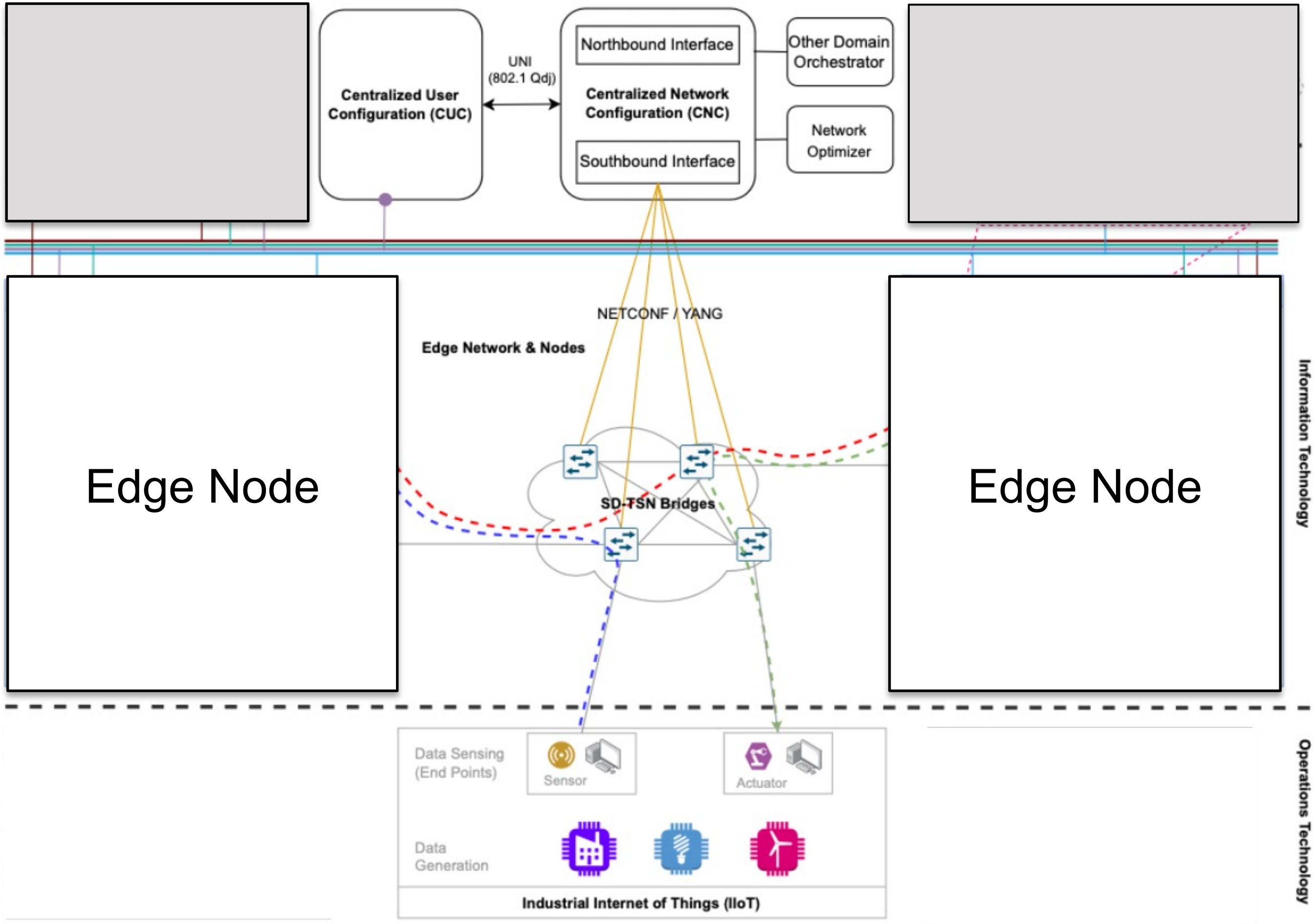
3 MAIN PILLARS FOR TRUSTWORTHY I-IOT APPLICATIONS



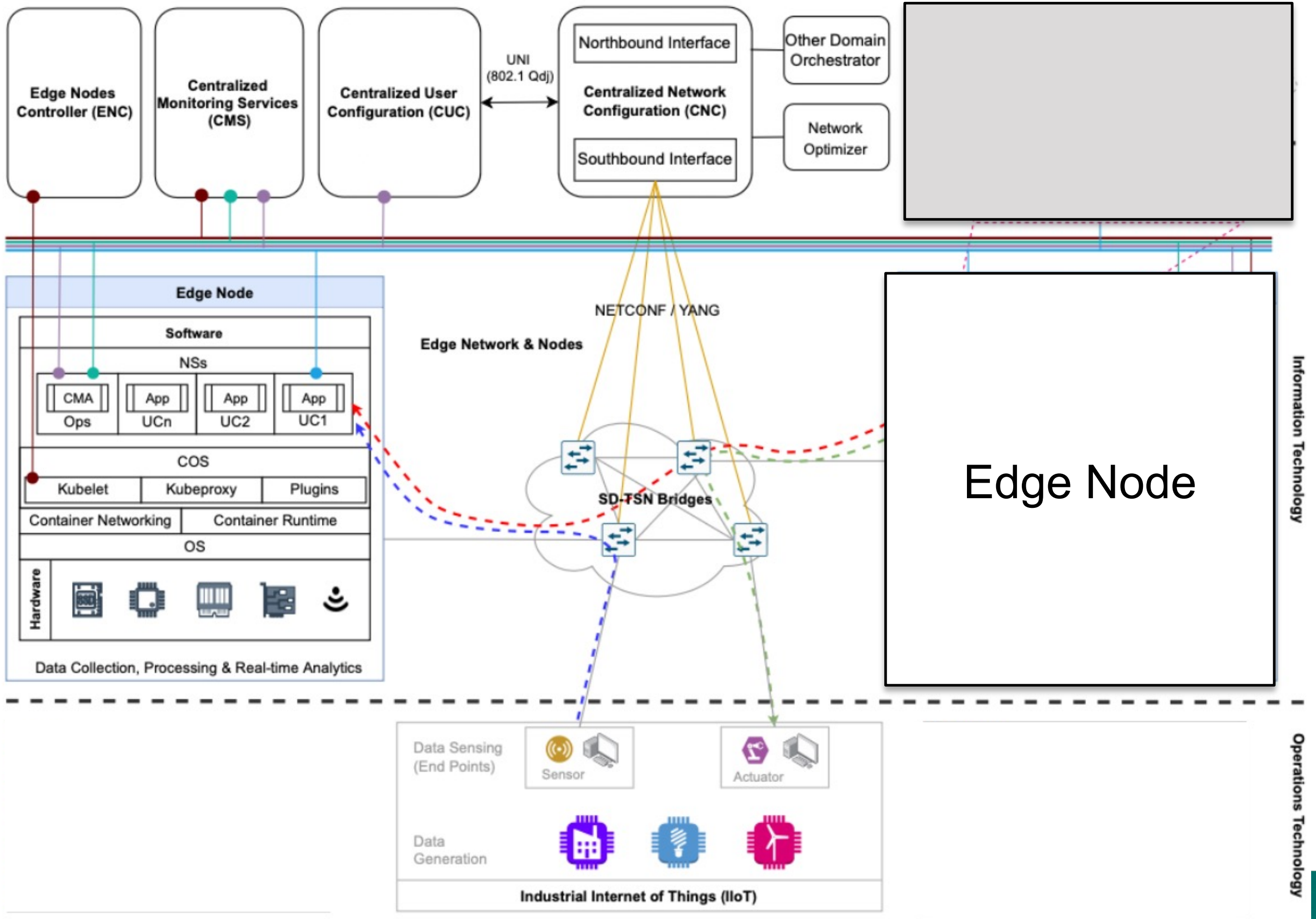
AIDA Architecture



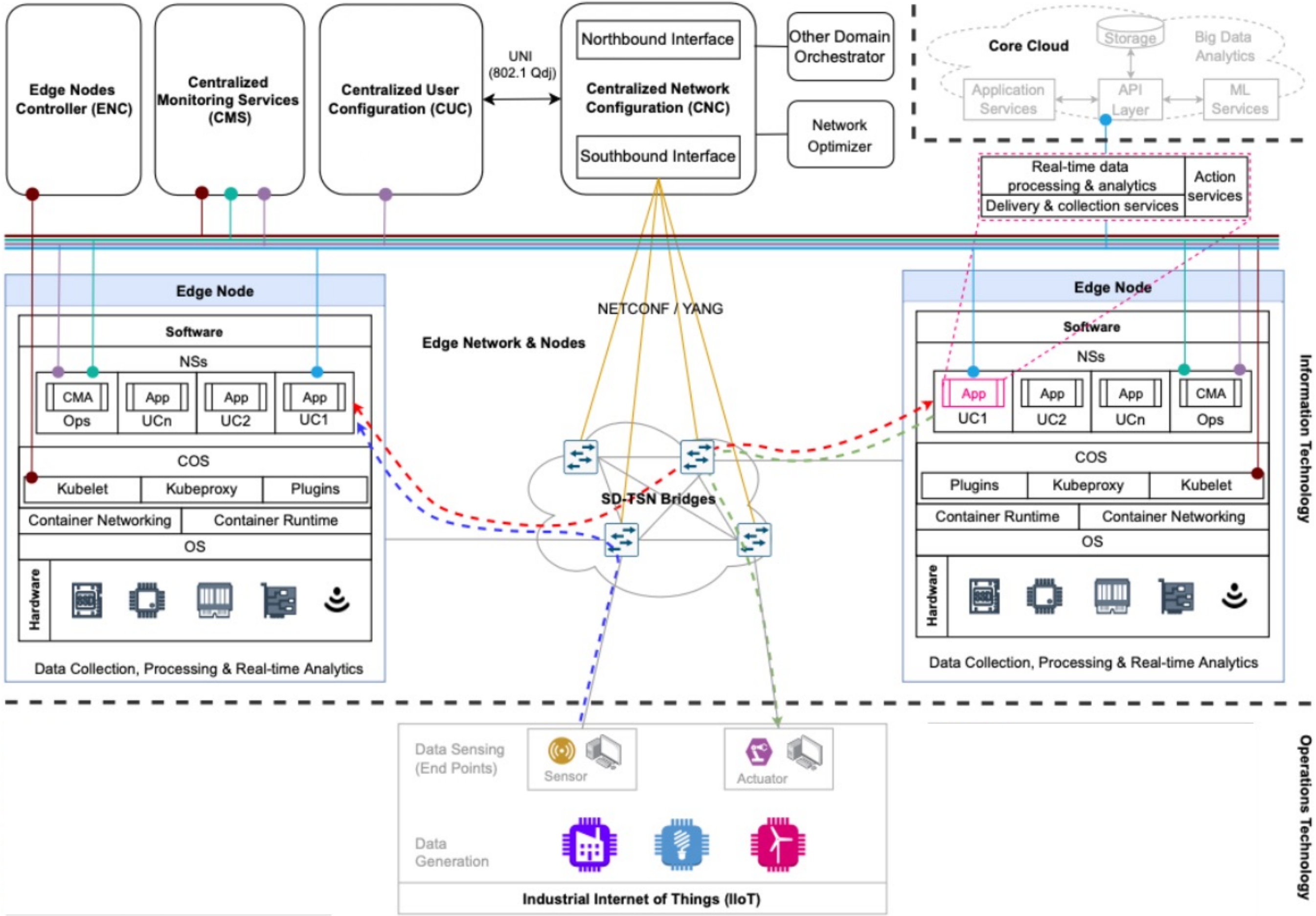
AIDA Architecture



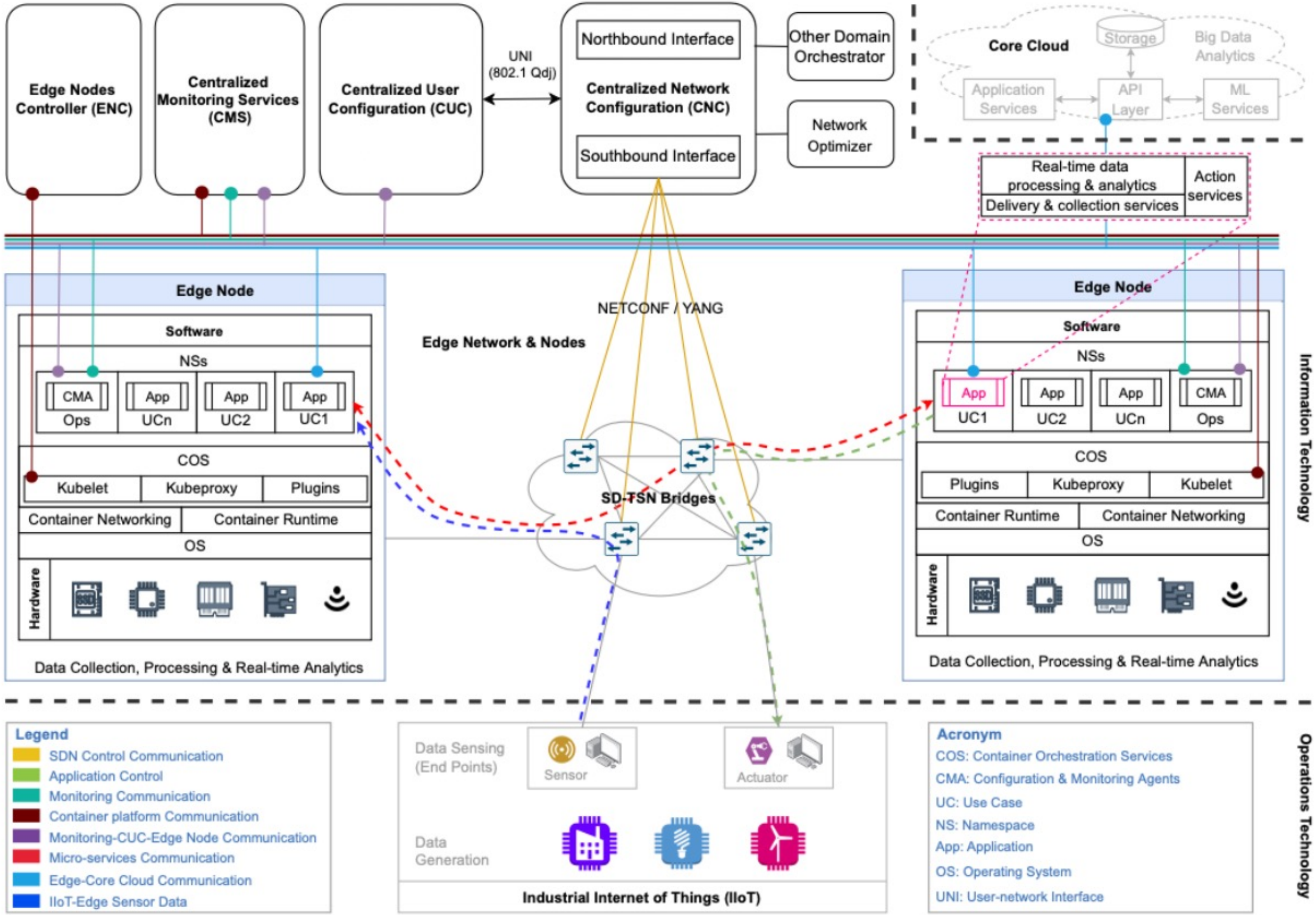
AIDA Architecture



AIDA Architecture



AIDA Architecture



Highlights – TSN Control Plane

- SDN based Control Plane for Time Sensitive Networks (TSN):
  - Microservice based Centralized Network Controller → OpenCNC → Open Source
    - Nortbound: 802.1 Qdj, Southbound: NetConf/Yang for pushing configuration, verified through plugfest
    - Kafka-based Monitoring Backend for Telemetry
  - Endhost support for configuration of i.225/i.226 TSN cards through detd (intel)
  - Joint orchestration of TSN/Talker placement and Network Configuration
- Robust Network (Re-) Configuration
  - Synthesizing TSN configurations using external optimizer
    - Deep reinforcement Learning algorithm design ongoing
    - Digital-Twin based validation approach using simulator in the loop
    - Genetic Algorithm for finding tradeoff between optimality and cost for reconfiguration

Highlights – Real-time Performance Monitoring

- Design of AIDA Distributed Observability Framework (DESK)
  - Based on literature review on observability of distributed edge and containerized microservices
  - Complete implementation based on selected open source tools and metrics
- Experimentation and Analysis of DESK
  - Initial DESK overhead and usability analysis
  - Fault detection and recovery using monitored data at edge nodes
- Latency Monitoring with eBPF
  - Design of ePPing tool for passive RTT measurements
  - Filtering and aggregation for increased efficiency
  - Validation and performance evaluation (PAM 2023)
  - Integration in LibreQoS
  - Measurement study at an ISP in the US is ongoing

Highlights – ML pipeline

- **Trustworthy ML in Production:**
  - New method for using data augmentation for ML testing
  - New methods for ML Testing in production
- **Concept Drift and ML Model Degradation:**
  - Improving scalability of industrial processes using drift handling techniques
  - Proposing an adaptive drift detection mechanism
- **ML pipeline and QA**
  - DQ within MLOps
  - Model versioning and performance degradation
  - Formalizing a holistic robust MLOps framework
- **Data-Centric ML Approach**
  - Data quality scoring approach
  - Evaluation in real-time industrial use cases
  - Improve the overall ML performance is on going
- **System anomaly detection using historical data.**
  - Performing literature study on algorithms and challenged in anomaly detection.
  - Anomaly detection of customers energy consumption using historical consumption data.

Further information – selected pointers

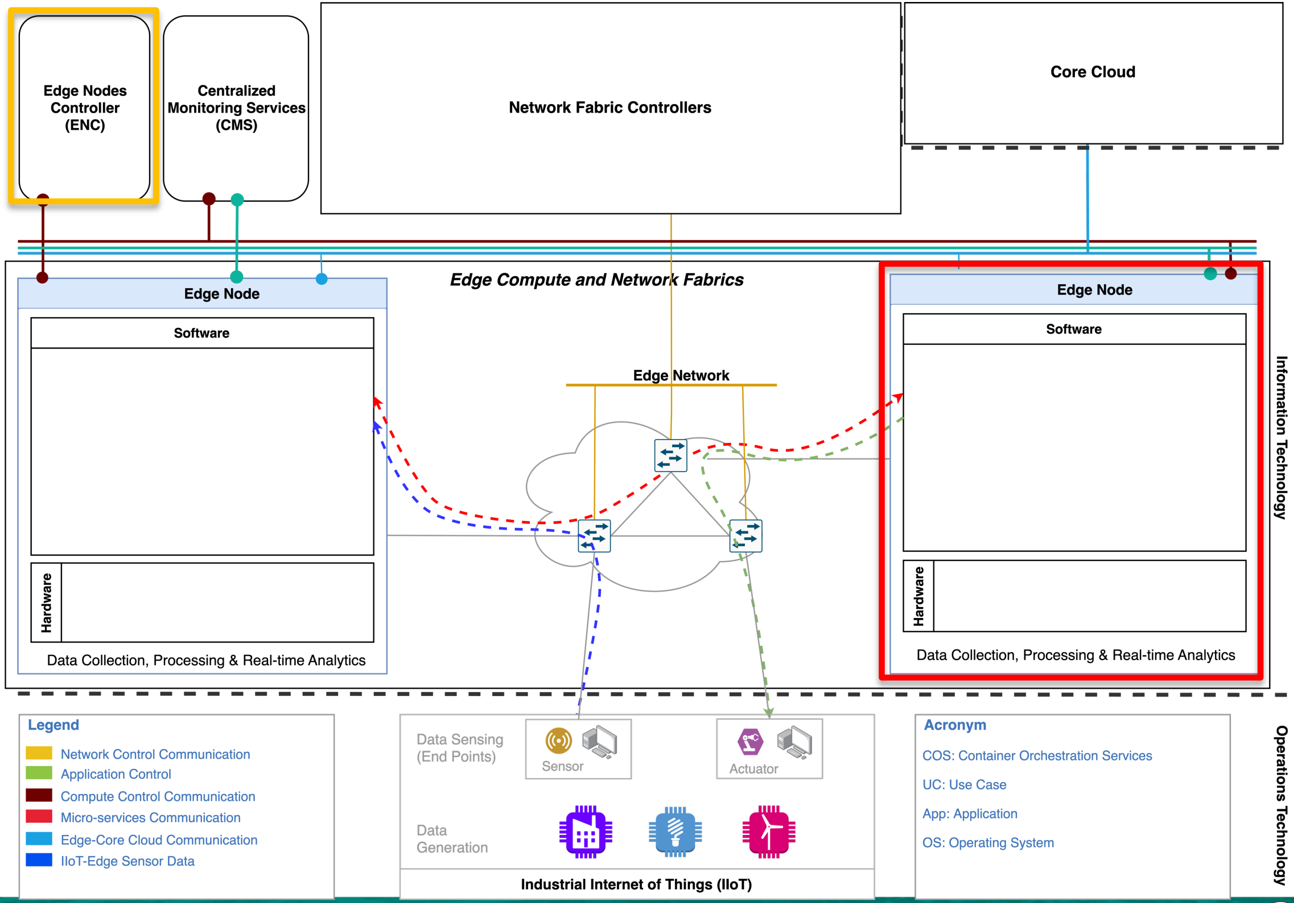
- H. Chahed, et. al., “AIDA—A holistic AI-driven networking and processing framework for industrial IoT applications”, Internet of Things, Volume 22, 2023.
- H. Chahed, S. Oechsle “Closing the configuration loop with OpenCNC and Control TSN Frameworks”, *TSN/A conference*, September 2023.
- H. Chahed, A. Kassler, “Software-Defined Time Sensitive Networks Configuration and Management”, IEEE NFV SDN 2021, 9-11 Nov. 2021.
- M. Usman, et. al., "DESK: Distributed Observability Framework for Edge-Based Containerized Microservices," EuCNC/6G Summit, June 2023.
- S. Sundberg, et. al., ”Efficient Continuous Latency Monitoring with eBPF”. Passive and Active Measurement (PAM), March 2023.
- F. Bayram, et. al., “A Drift Handling Approach for Self-Adaptive ML Software in Scalable Industrial Processes“, IEEE/ACM International Conference on Automated Software Engineering (ASE), Oct. 2022.
- A. Chatterjee, et. al., “Testing of Machine Learning Models with Limited Samples: An Industrial Vacuum Pumping Application“, ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE), November 2022.

- Github: https://github.com/AIDA-KAU
- Web page: https://sola.kau.se/aida/

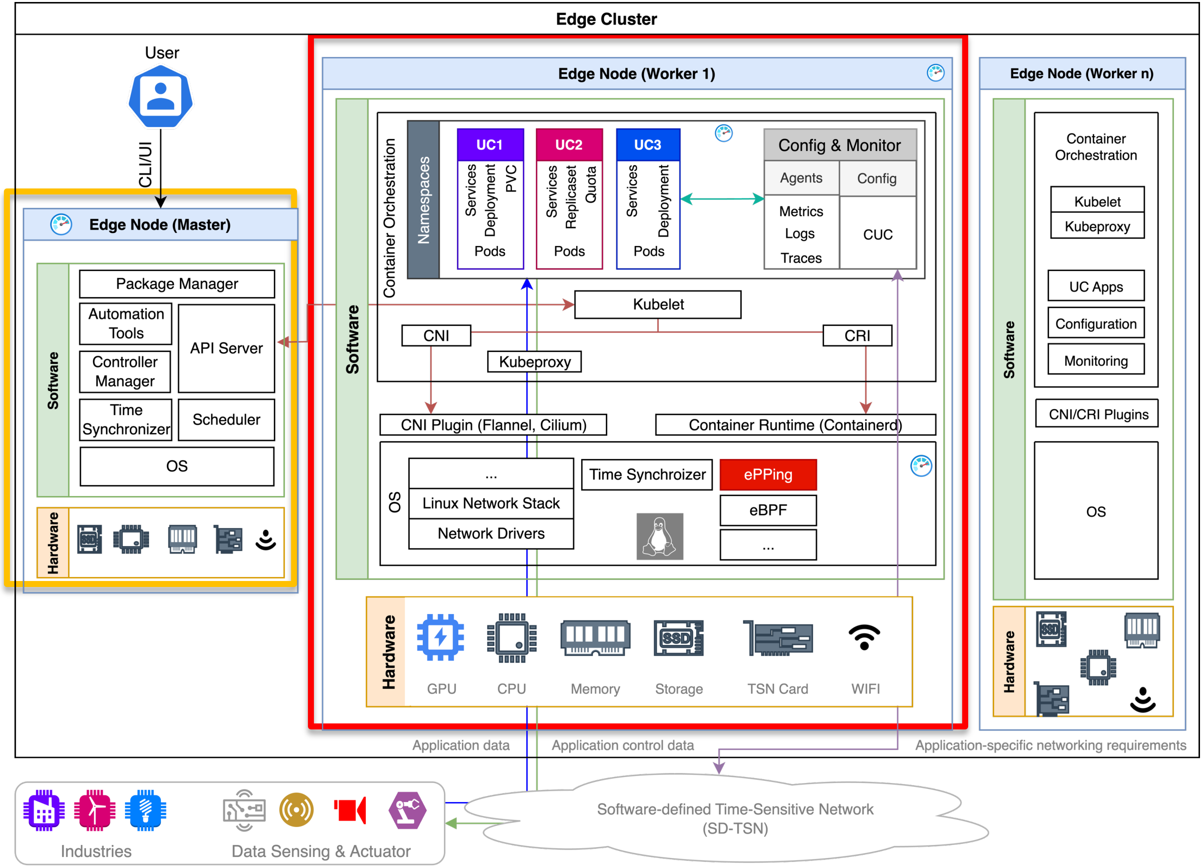
DISTRIBUTED OBSERVABILITY FRAMEWORK

RED HAT RESEARCH DAYS 2023-09-21

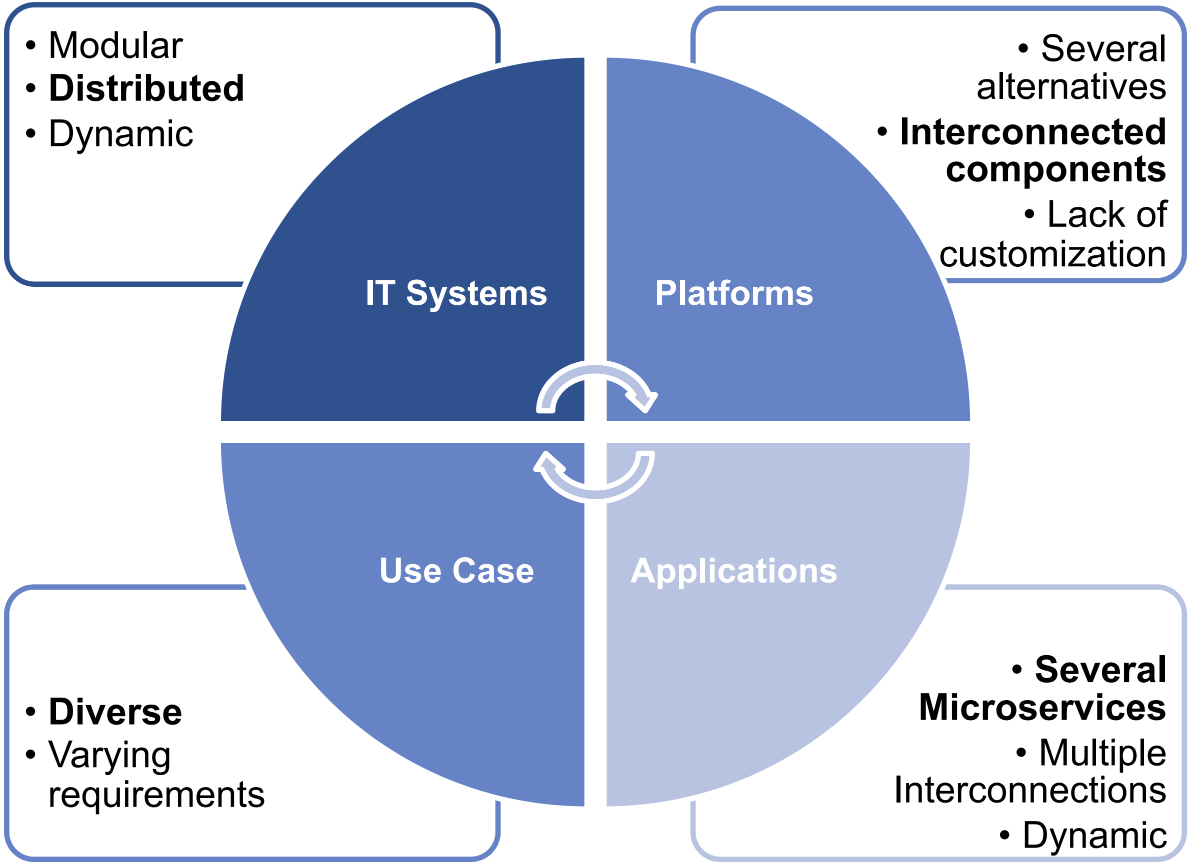
AIDA Overall Architecture



Container-based Edge Computing Platform



Challenges in Monitoring of Distributed Systems

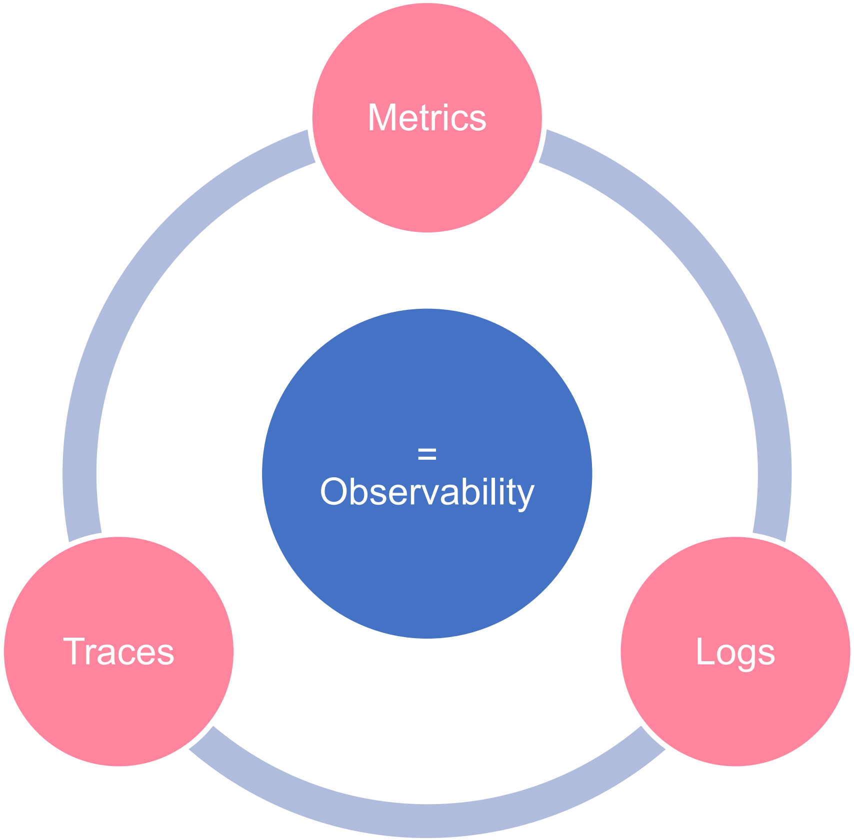


Observability in Distributed Systems

SURVEY

A Survey on Observability of Distributed Edge & Container-Based Microservices

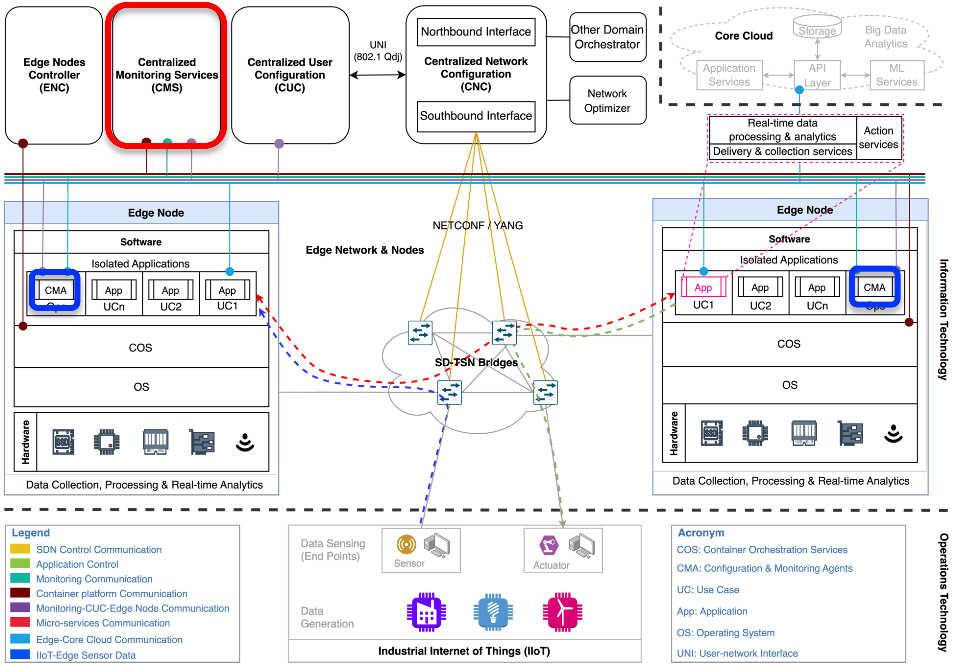
MUHAMMAD USMAN[1], (Member, IEEE), SIMONE FERLIN[1,2], ANNA BRUNSTROM[1], (Member, IEEE), AND JAVID TAHERI[1], (Senior Member, IEEE)
[1]Department of Computer Science, Karlstad University, 65188 Karlstad, Sweden
[2]Red Hat, 16451 Stockholm, Sweden



- Latency (Request service time)
- Traffic (User demand)
- Errors (Rate of failed requests)
- Saturation (Overall system capacity)

Real-time Performance Observability & Optimization Framework

AIDA Overall Architecture



DistributEd obServability frameworK (DESK)

Server-side Components/Services

Measurement Agent(s)

DistributEd obServability frameworK (DESK)

Conferences > 2023 Joint European Conferenc...

DESK: Distributed Observability Framework for Edge-Based Containerized Microservices

Publisher: IEEE   Cite This   PDF

Muhammad Usman ; Simone Ferlin ; Anna Brunstrom ; Javid Taheri   All Authors

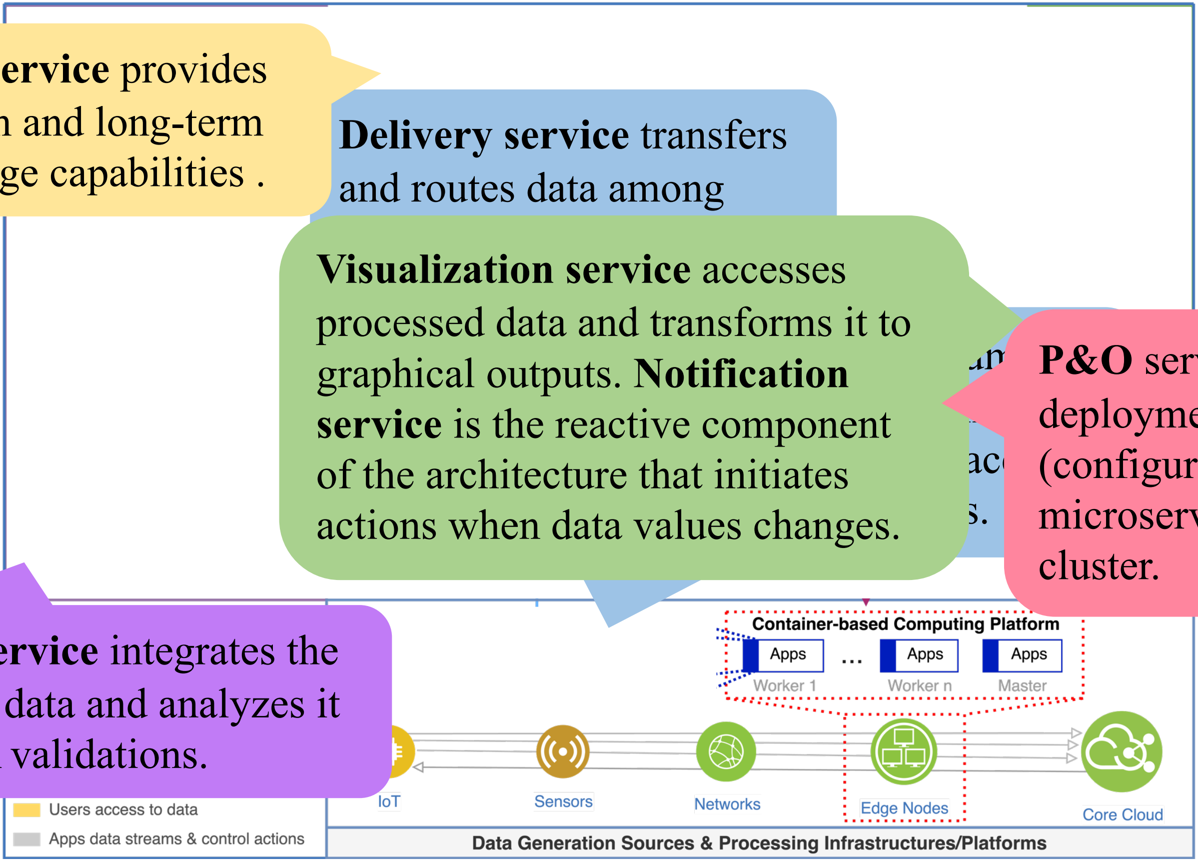
Storage service provides short-term and long-term data storage capabilities .

Delivery service transfers and routes data among

Visualization service accesses processed data and transforms it to graphical outputs. **Notification service** is the reactive component of the architecture that initiates actions when data values changes.

P&O service takes care of deployment and re-(configuration) of deployed microservices in the edge cluster.

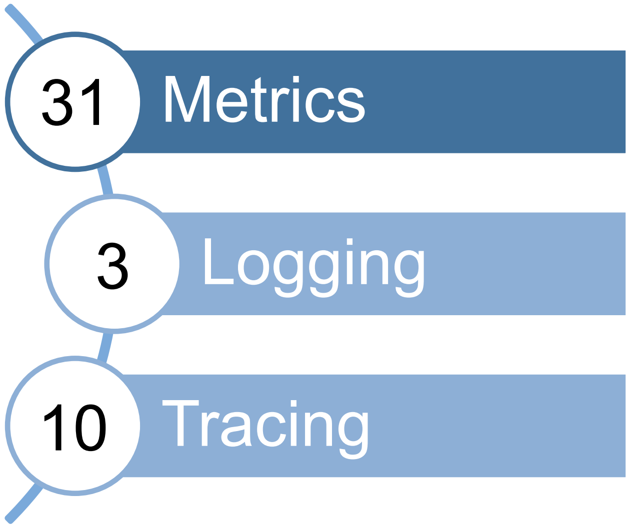
Fusion service integrates the collected data and analyzes it e.g., SLA validations.

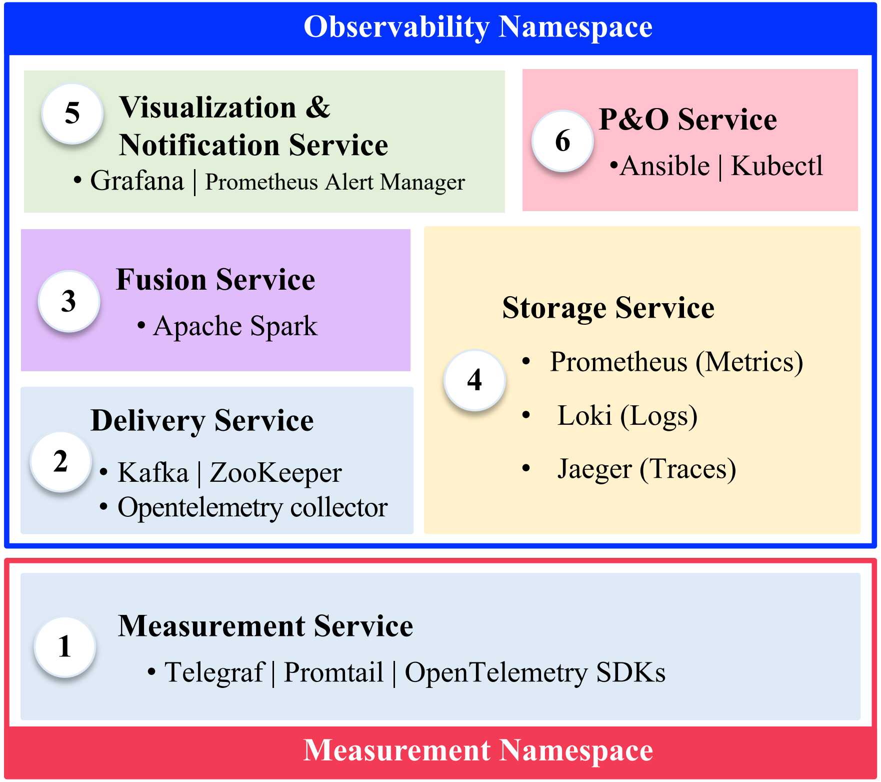


DESK Implementation

https://github.com/AIDA-KAU/Distributed-Observability-Framework.git

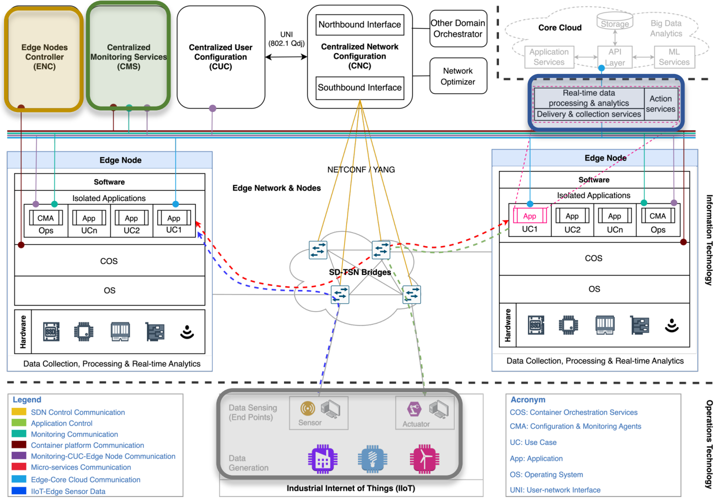
CNCF hosts around 103 projects for observability and analysis (44 projects are open source)

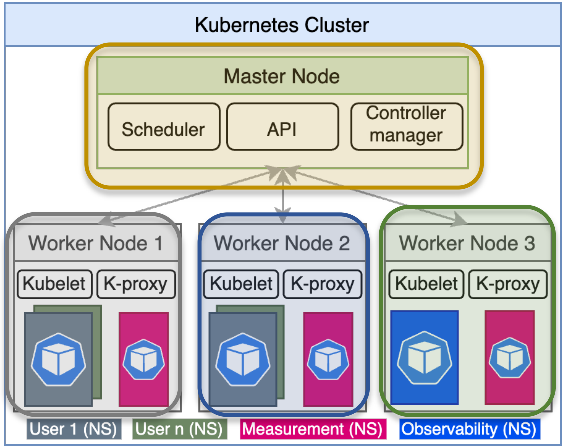




Experimental Setup

- **Hardware:** Desktop-based
- **OS:** Ubuntu 22.04.2 LTS
- **Kernel:** 5.15.0-72-generic
- **Kubernetes version:** v1.26.0
- **Containerd version:** 1.6.21
- **ThingsBoard** edge IoT Platform
- Custom-developed **Simulated** Sensor Data Pipeline





Measurements Overhead Experimentation (Metrics)

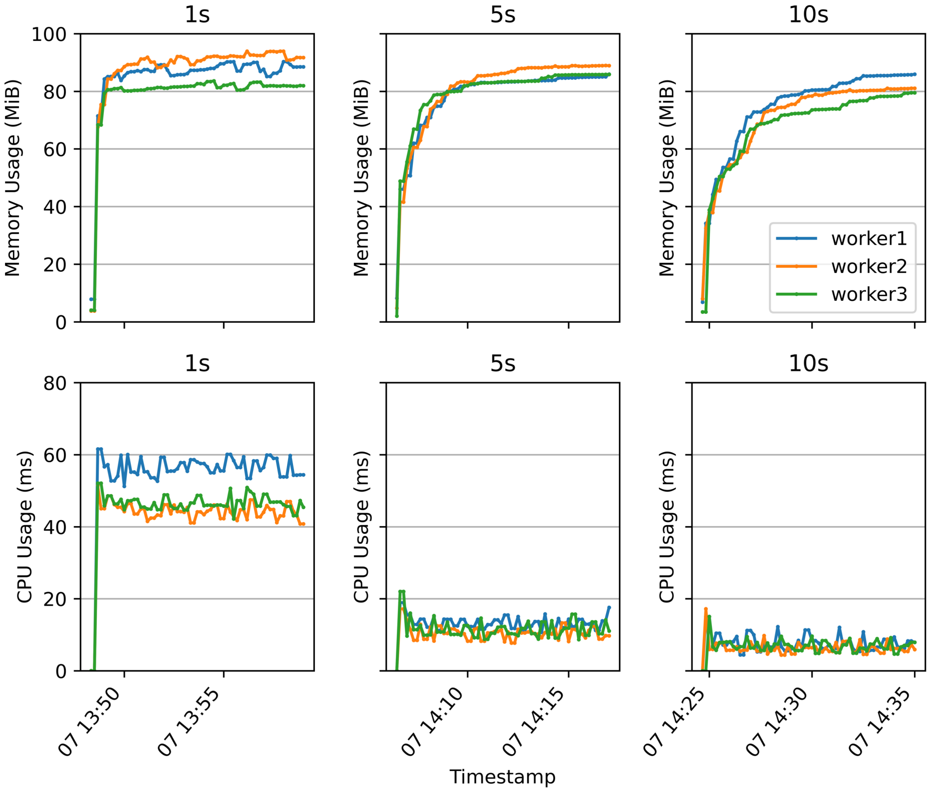
- Agents are created as Daemonset
- Measurement interval: 1s, 5s, 10s
- Number of metrics: 90

Metrics Agent configuration

```yaml
spec:
  serviceAccountName: telegraf
  hostNetwork: true
  containers:
    - <8 keys>

    - name: telegraf
      image: telegraf:latest
      resources:
        limits:
          cpu: 100m
          memory: 80Mi
        requests:
          cpu: 25m
          memory: 25Mi
      env:
        - name: HOSTIP
          valueFrom:
            fieldRef:
              fieldPath: status.hostIP
        - name: HOSTNAME
          valueFrom:
            fieldRef:
              fieldPath: spec.nodeName
```

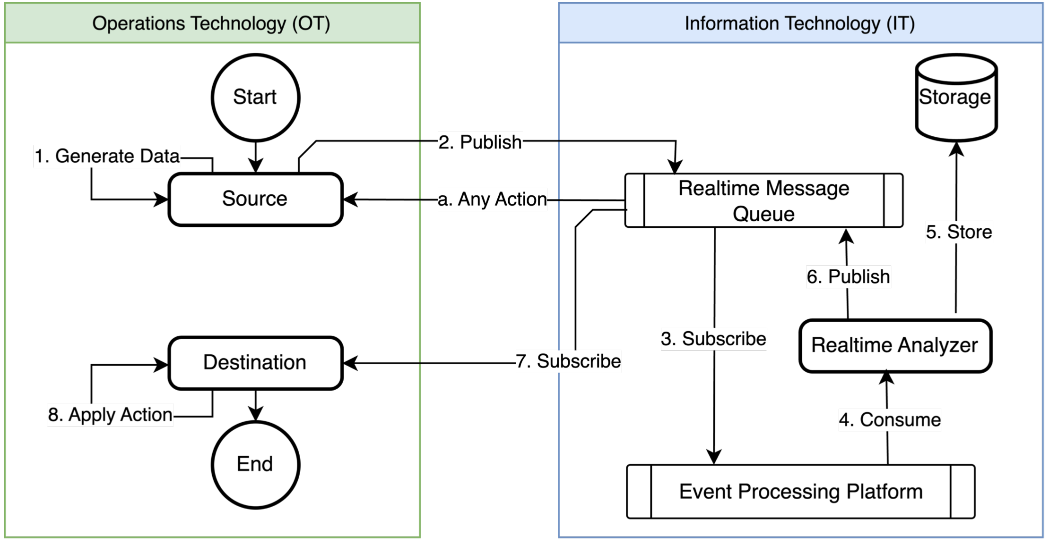
Plugi
Input I
temp
powers
system
cpu
mem
diskio
net
docker
kubern
Outpu
kafka
promet

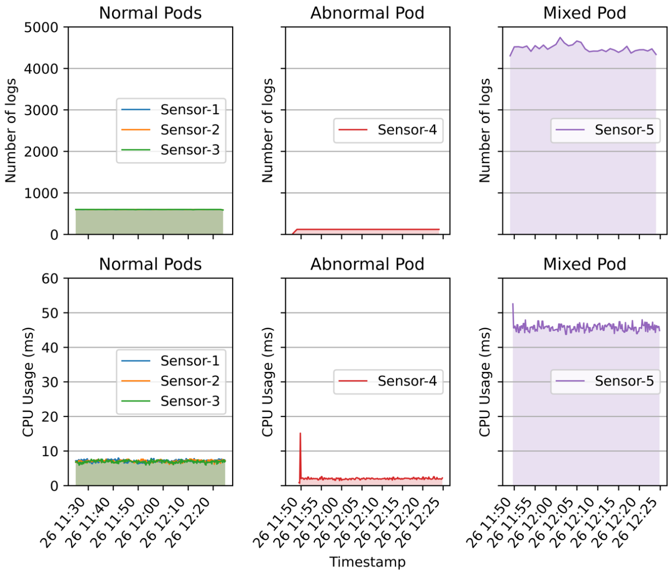


Fault Detection based on Measured Data (1/2)

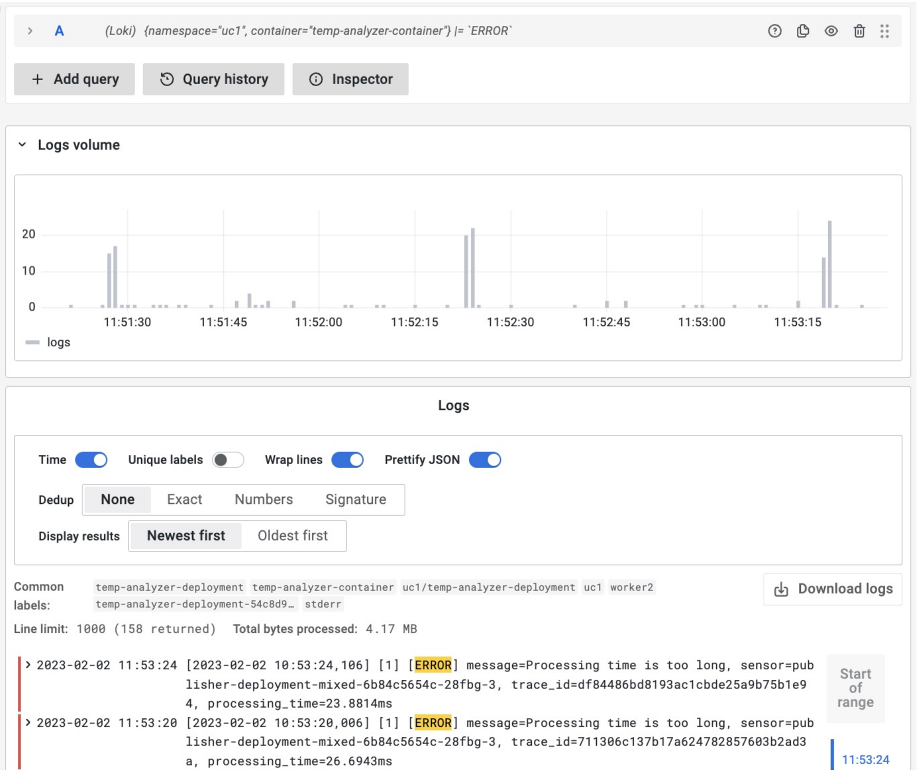
- Develop an end-to-end system:
  - Generate a continuous stream of simulated sensor data

Sensor	Data	Transmission
Normal	Valid	Regular
Abnormal	Valid	Irregular
Mixed	Valid + Invalid	Irregular

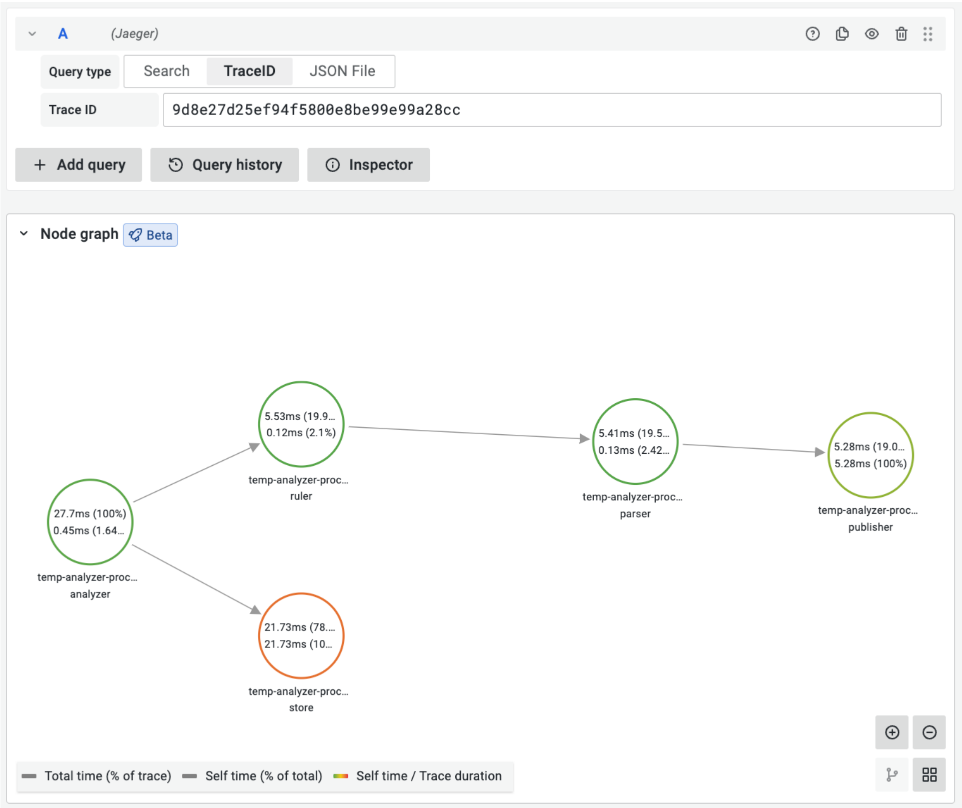




Fault Detection based on Monitored Data (2/2)

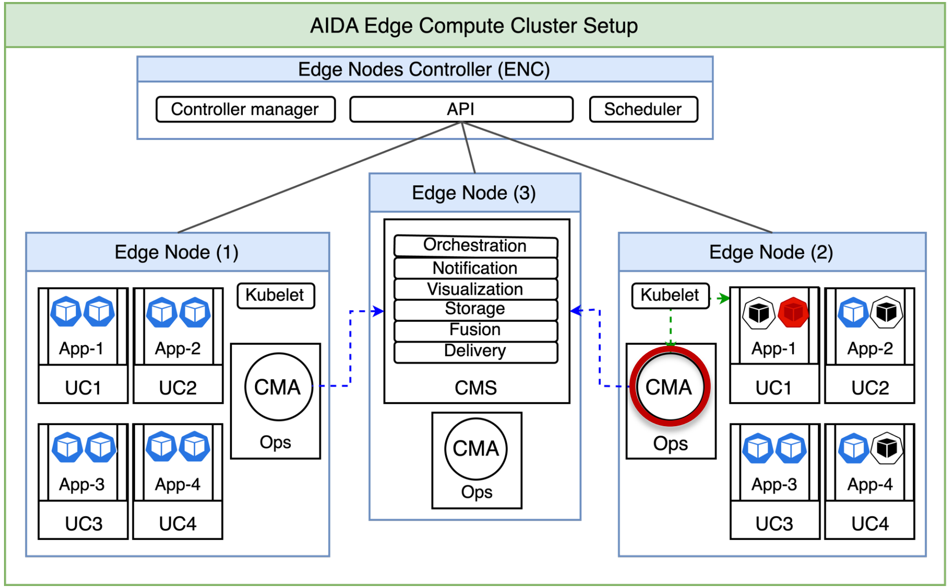


Error Logs

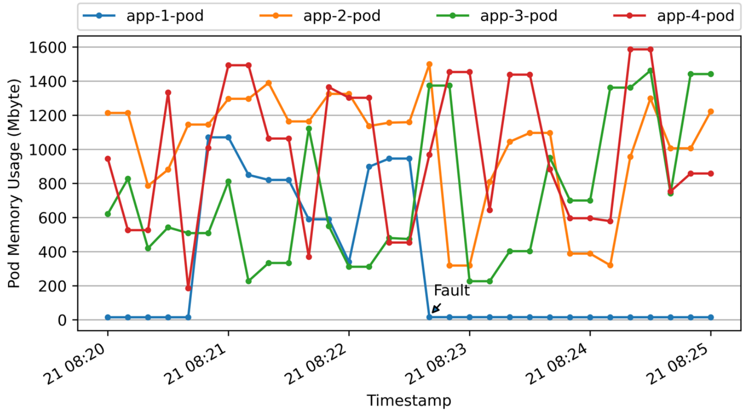


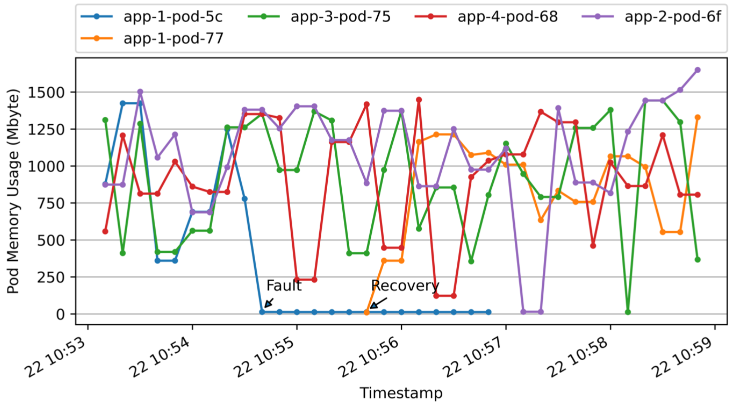
Connected traces

Fault Detection/Recovery Using Metrics at Edge Nodes



Pod memory usage (w/o recovery).





Pod memory usage (with recovery).

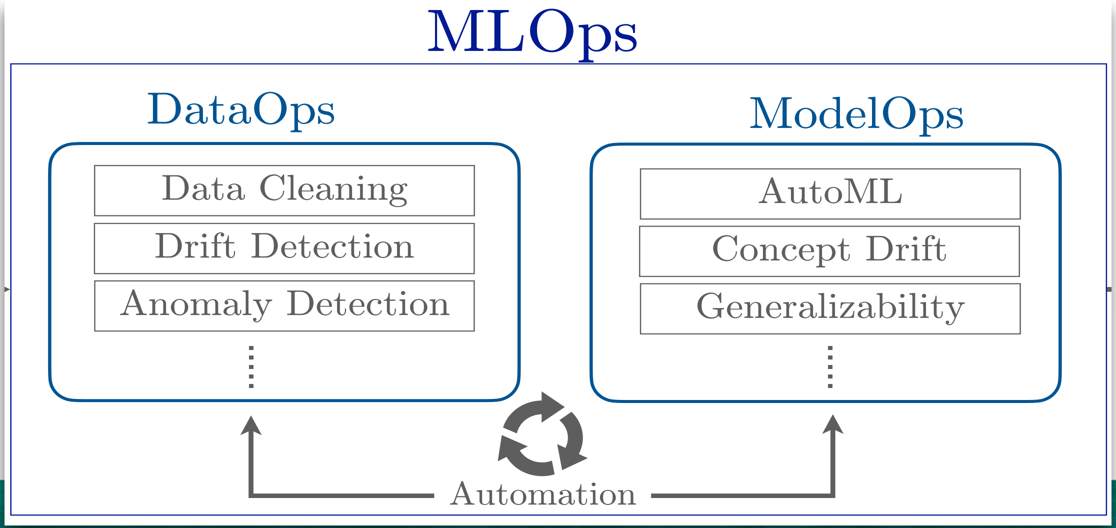
Detect crashed applications inside a container that is reported as operational by Kubernetes

ML PIPELINE

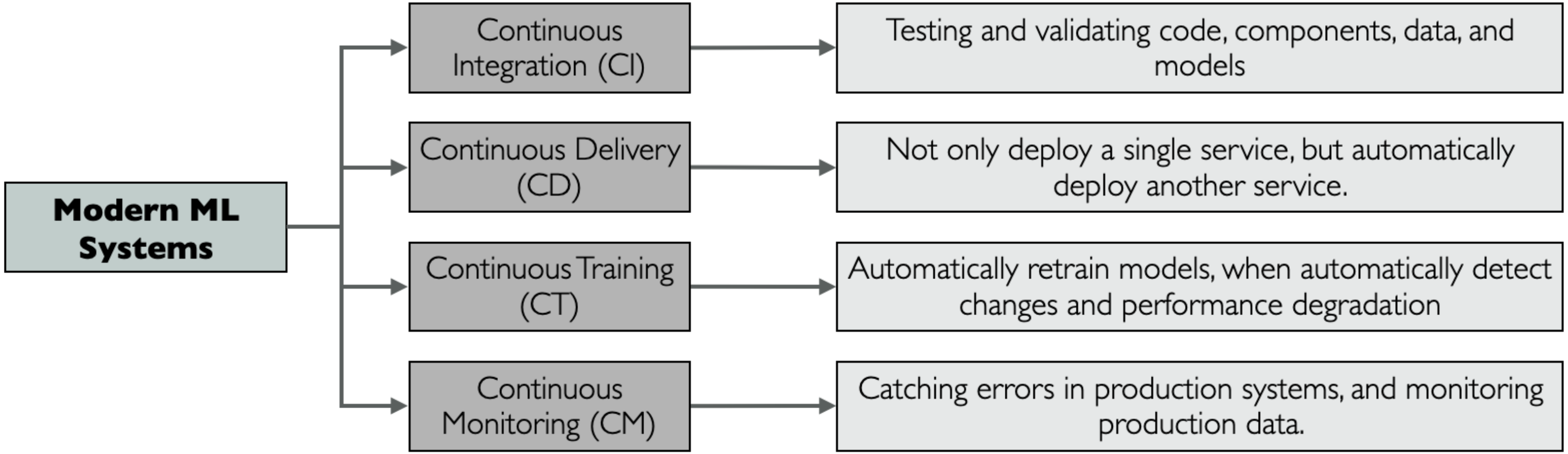
RED HAT RESEARCH DAYS 2023-09-21

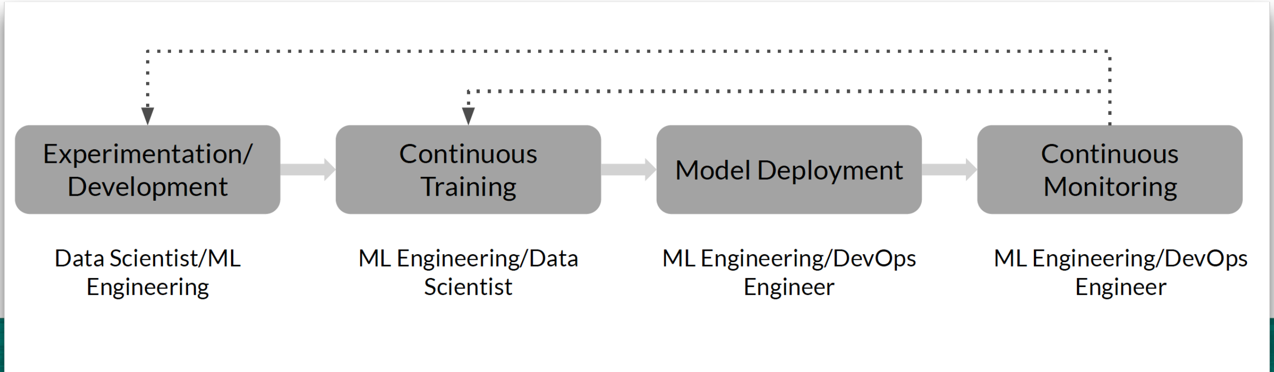
Towards Robust ML Systems in Production

- **Robust performance** is essential for **trustworthy AI systems** (according to EU guidelines)
- **DataOps**: end-to-end data processing operations in production
- **ModelOps**: the set of operations that are performed on the learning task of the ML model
- **Automation**: the engine that drives and coordinates the overall operations



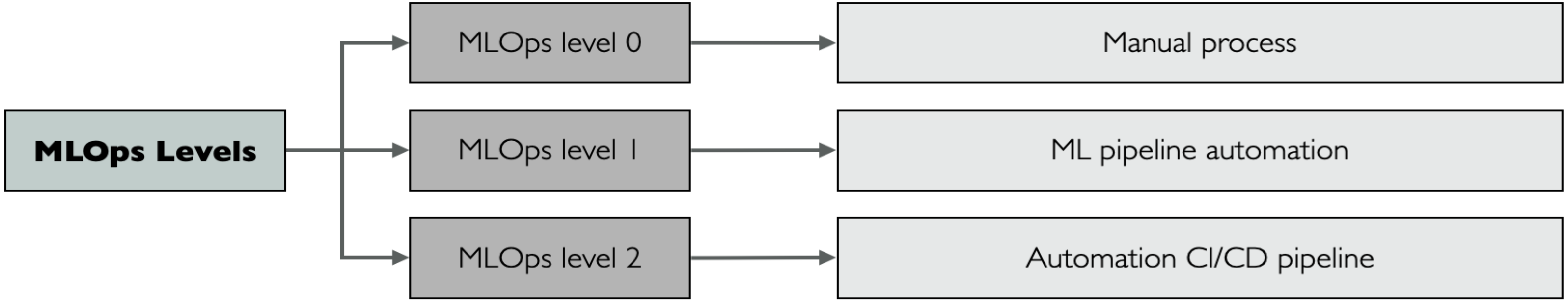
Modern ML Systems in Production





Deploying AI at Scale

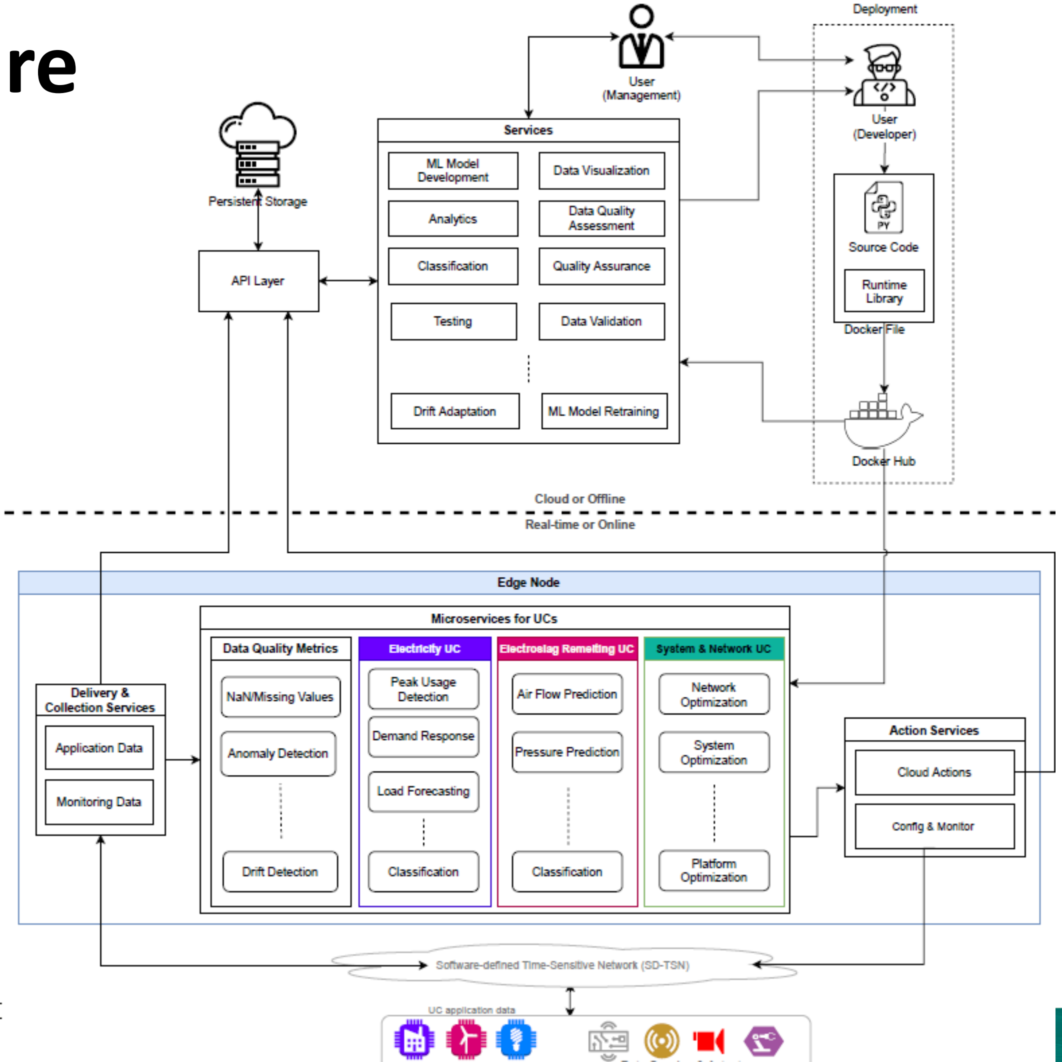
- **Continuity and automation .. Towards continuous everything**
- **Several Challenges:**
  - **Data Quality and Quantity:** Large-scale deployment requires a huge amount of high-quality, labeled data
  - **Model Performance:** AI model degradation
  - **Integration with existing systems:** compatibility and technical difficulties.
  - **Maintenance and Updating:** AI models need to be maintained and updated regularly.



WP3 Architecture

Not many details with the current architectures in the literature, e.g., Google reference Architecture.

Several Software Engineering concepts are missing in the current architectures



Bestoun S. Ahmed, Ph.I

MLOps Lifecycle

- Added value
- Business Requirements
- Key metrics
- Data processing

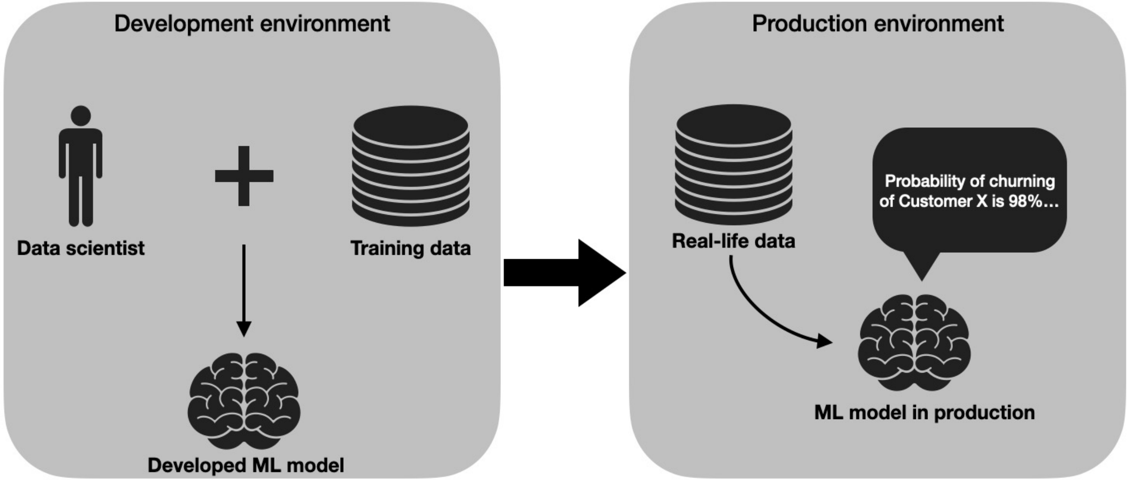
- Feature engineering
- Experiment tracking
- Model training & evaluation

- Runtime environments
- Microservices architecture
- CI/CD pipeline
- Monitoring &retraining

- **ML Lifecycle in MLOps:**
  - **Design:** Project conceptualization and goal-setting.
  - **Development:** Model building, training, and evaluation.
  - **Deployment:** Model goes live in production.

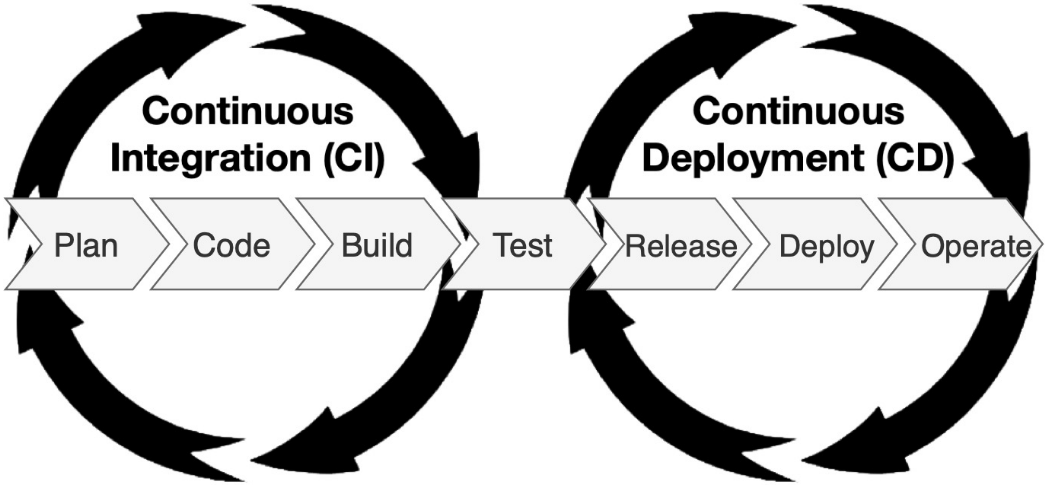
Development to Deployment

- **Development Environment**: Used for model development and testing.
- **Production Environment**: Where the ML model operates in real-time, making predictions based on incoming data.
- **Once the ML model is developed, we need to move the ML model into the production environment.**



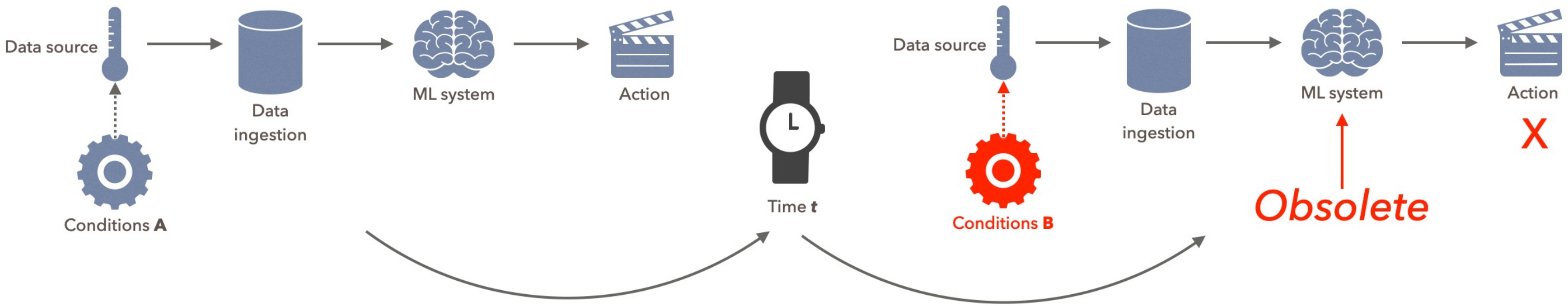
CI/CD

- The use of **continuous integration and continuous deployment**
- **Automating Deployment**: of code, including MLmodels.
- **Series of Steps**: developing, testing, and deploying code, enabling incremental changes and efficient production deployment.



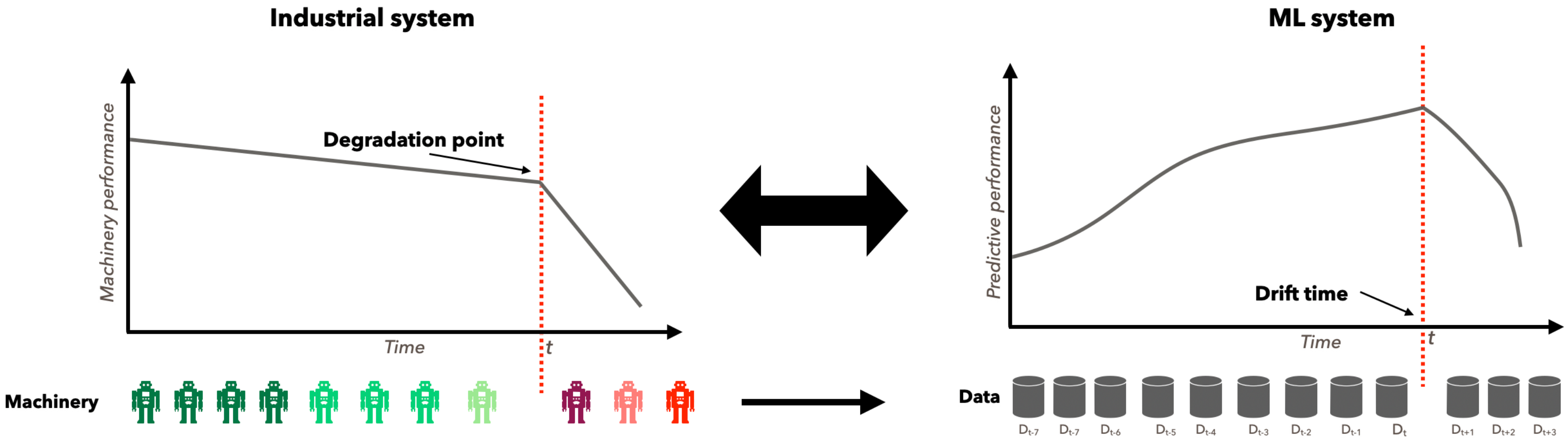
Data-centric problems in Production

- Data verification and training data evolvability
- ML decision making correctness and algorithm testing
- Testing for ML model degradation.
- **Training Data Evolvability** - test the training data against the used model
- **Quality of the data:** Insufficient data, irrelevant features, non-representative training data, overfitting, under-fitting, outliers.



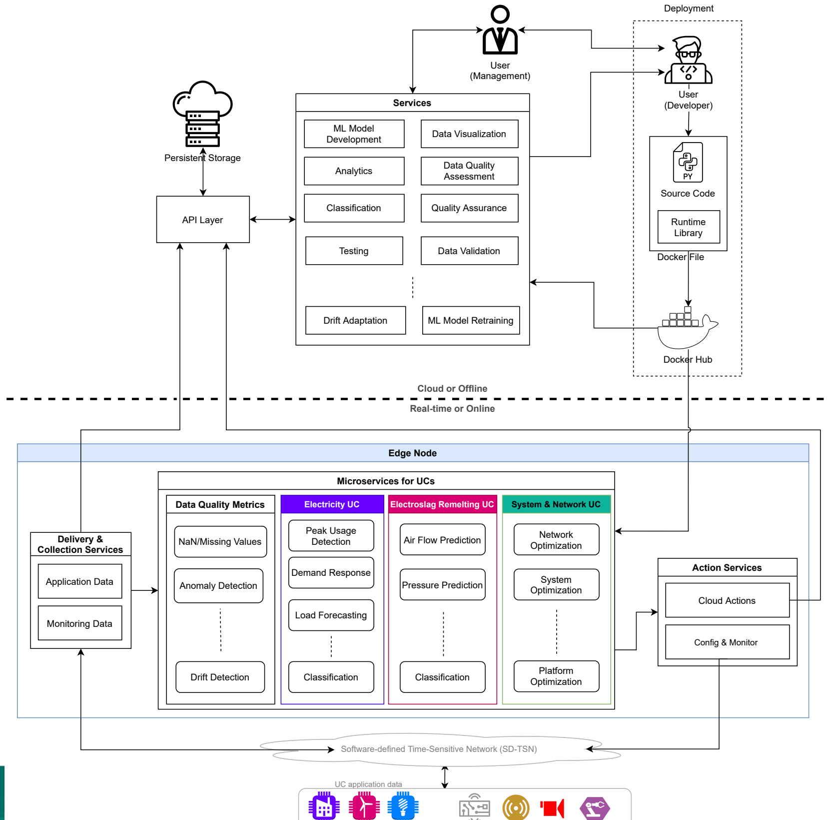
Data Drift and Model Degradation

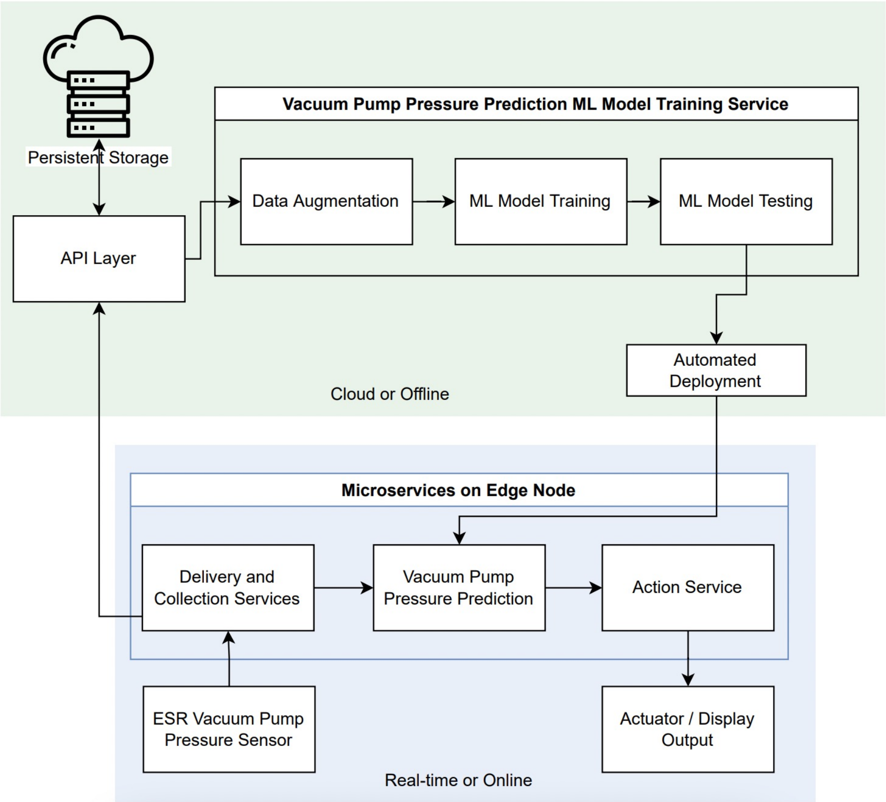
- **Problem Context:** reduce operational costs
- **Task**: prevent **costly** breakdowns
- **Initial Predictive Models:** trained on historical data
- **Machinery degradation -> Drift -> ML model degradation**



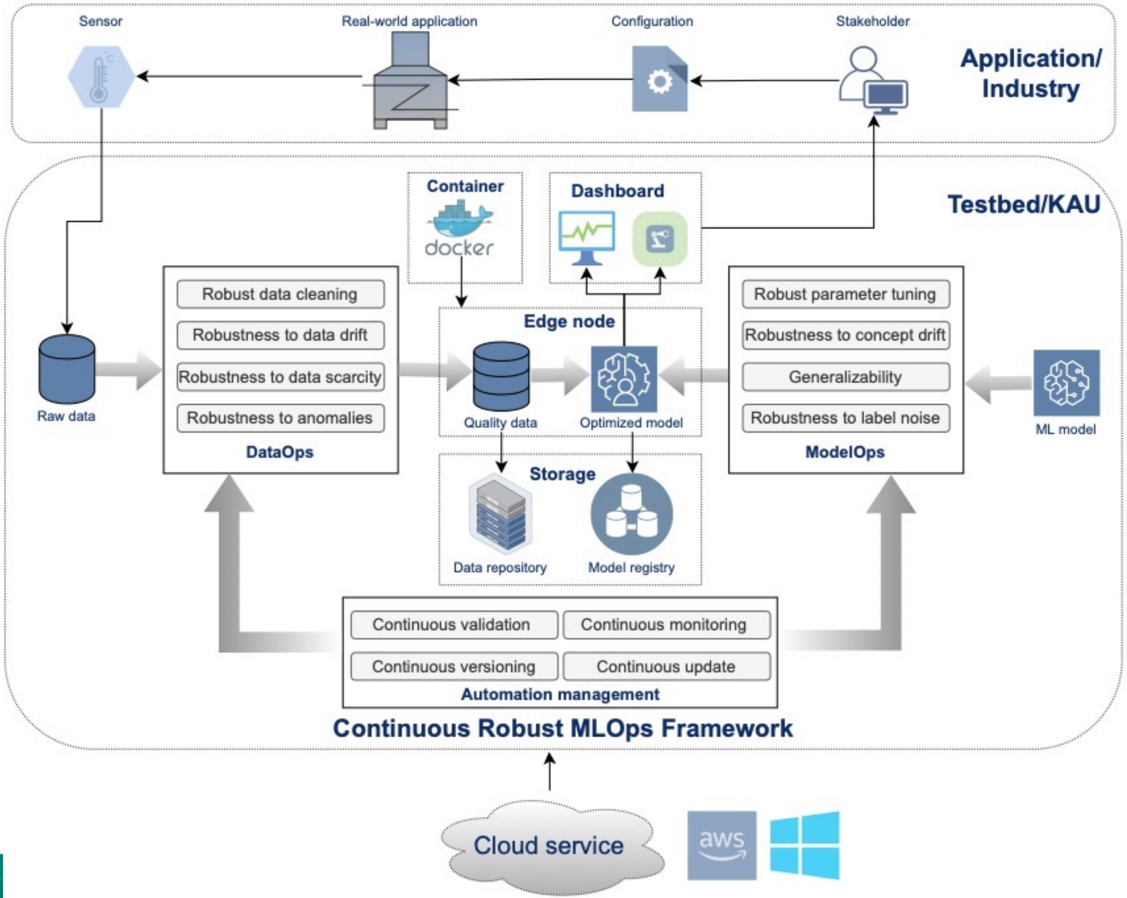
Dealing With the Problems in MLOps and DataOps

- Building customisable micro services to deal with the custom data and ML problems



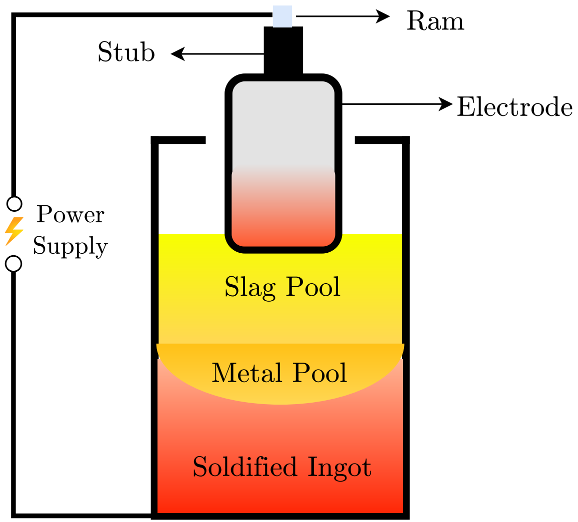


End to End Approach



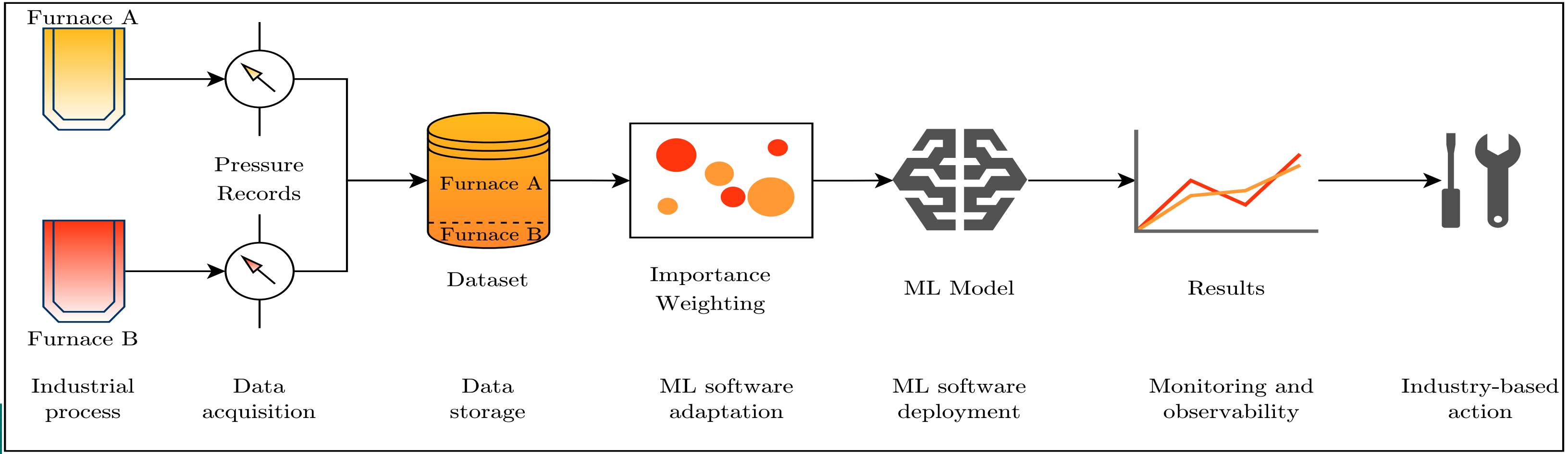
Self-Adaptive Drift Handling

- ML software to **predict the minimum pressure value** of a pumping event
- The minimum pressure value is predicted every **30 seconds** for up to **3 minutes**
- Evaluate: **Predicted value < pressure threshold**
- Benefit: **Early identification of invalid pumping events**
- Industrial process scalability: Introducing **a new furnace** to the industry
- **Fast integration** in the predictive system



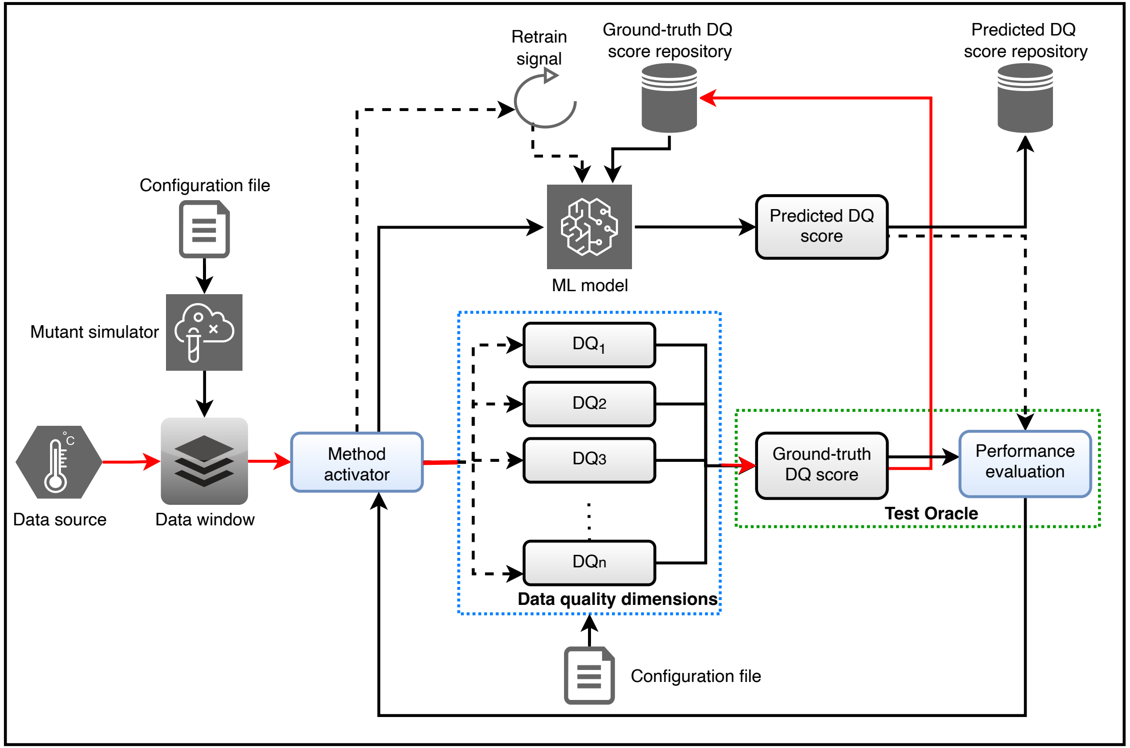
Drift Handling for Self-Adaptive ML in Scalable Industrial Processes

- **Collect —> Adapt —> Deploy —> Monitor —> Decision**
- Shift adaptation: **importance weighting, Kernel Mean Matching (KMM)**
- ML model: **Random Forests(RF)** and **XGBoost**
- Evaluation metric: **mean absolute percentage error (MAPE)**

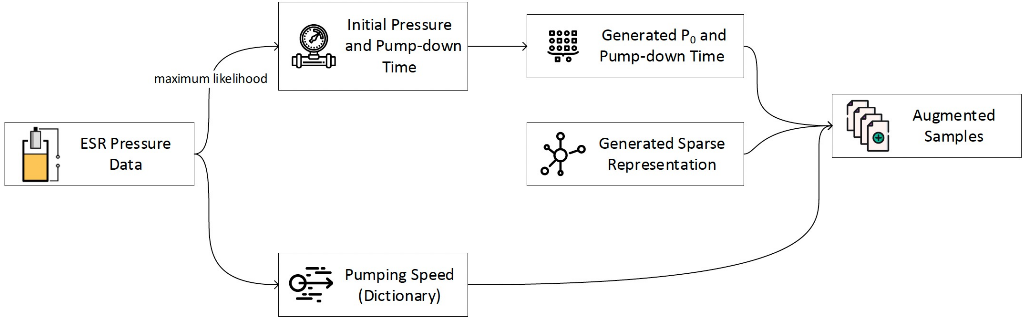


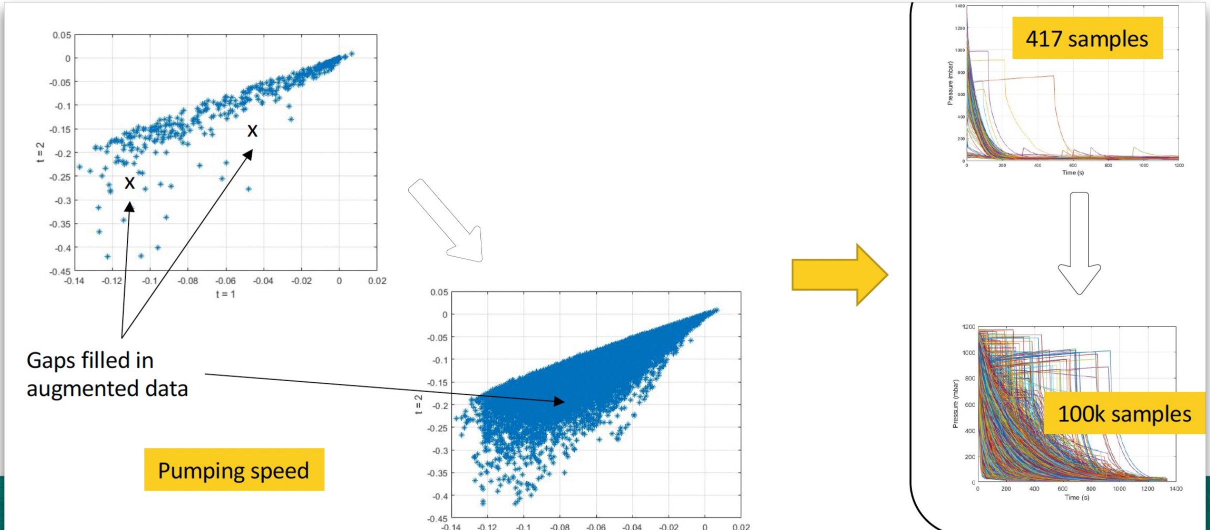
Data Quality Scoring

- ML approach
- Score n data points using the pipeline approach
- Train ML regression on the training datasets of size n
- Predict the score of the testing dataset of size l
- **DQSOps: Continuous Data Quality Scoring Framework for Data-Driven Applications**



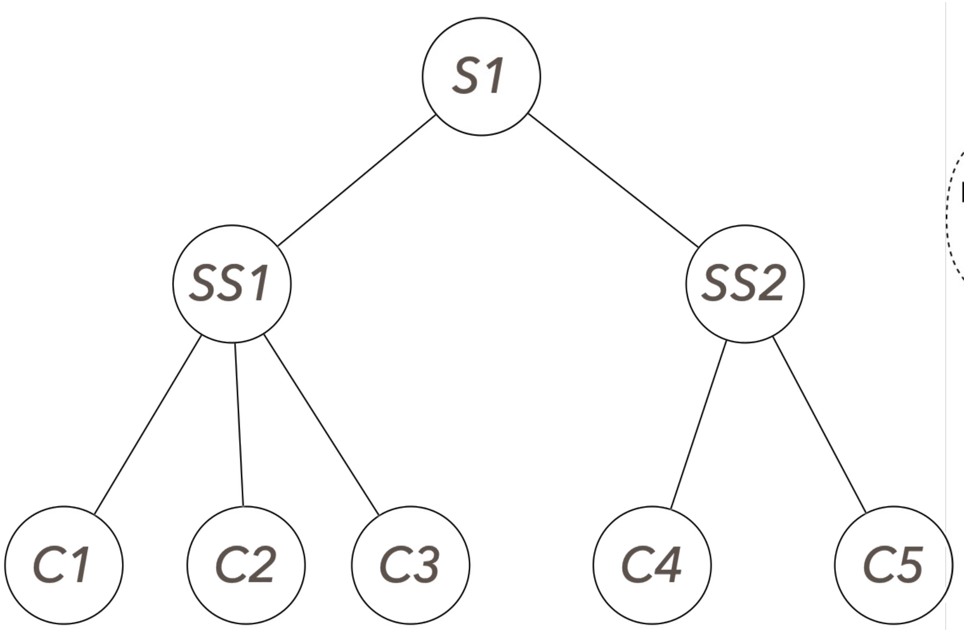
Data Augmentation for Limited Data

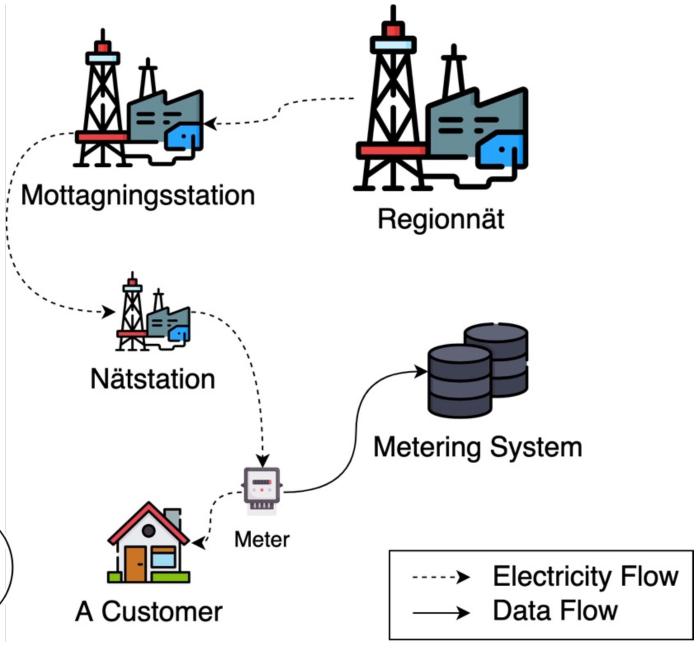




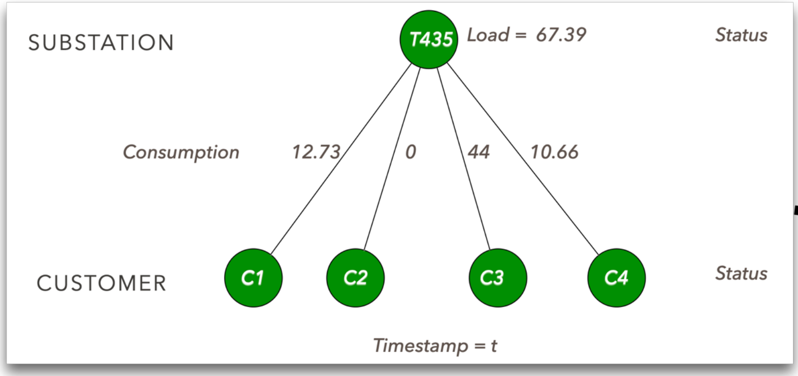
Detecting and Predicting Faults in Electricity Grid Using Customer Data and Topology

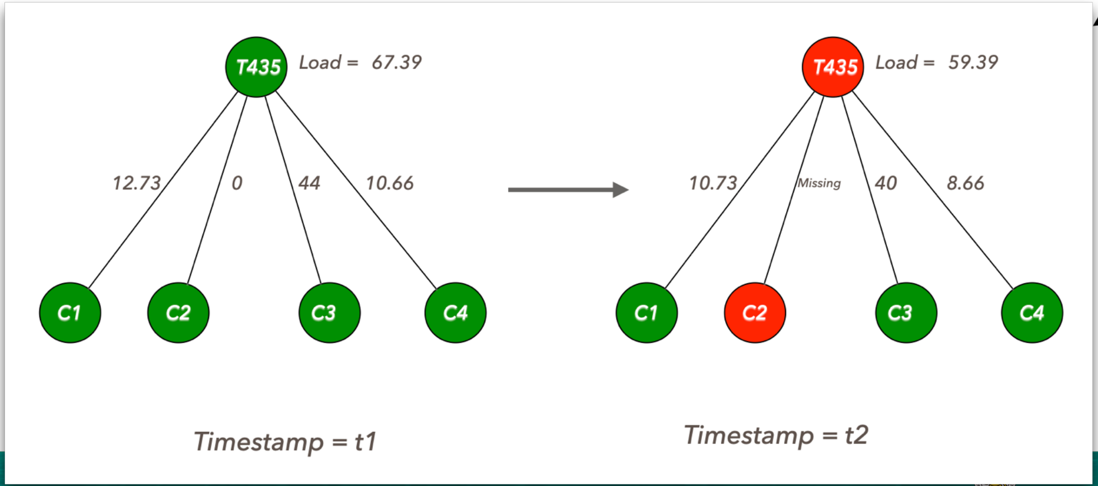
- Simulating the Overall Topology





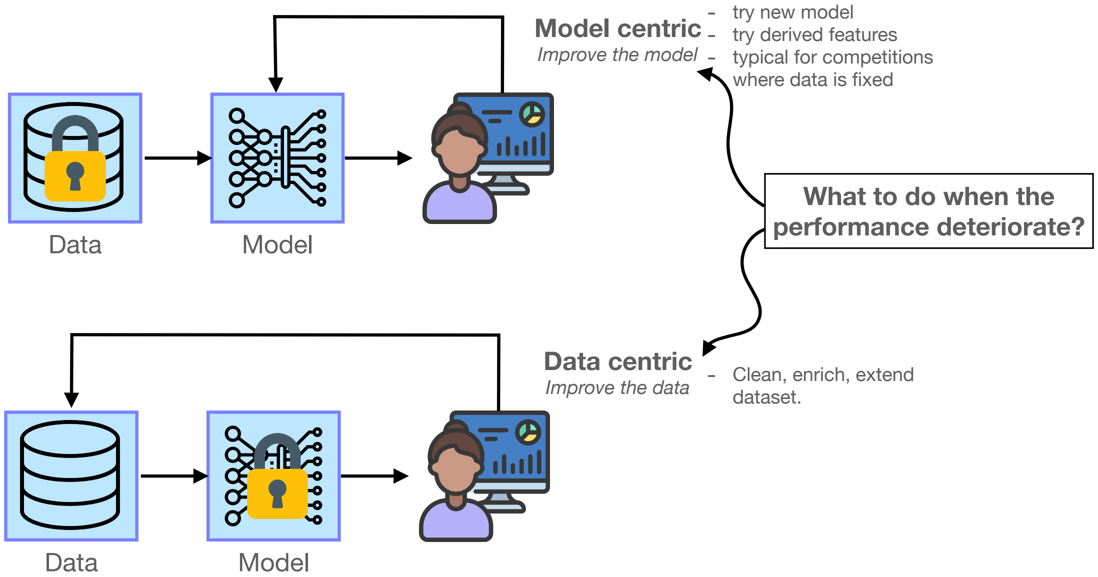
Tracking Changes





What is Next?

- Implementing End-to-End Services in Cloud.
- Continuous testing for model rollback
- Seamless deployment in production and robustness
- Continuous mutation testing for data augmentation
- Continuous anomaly detection and QA.



kau.se/cs

Committed to excellence in distributed systems and communication, security and privacy, and software quality