HIVEMQ

MQTT Standards for Integrating Edge AI Systems

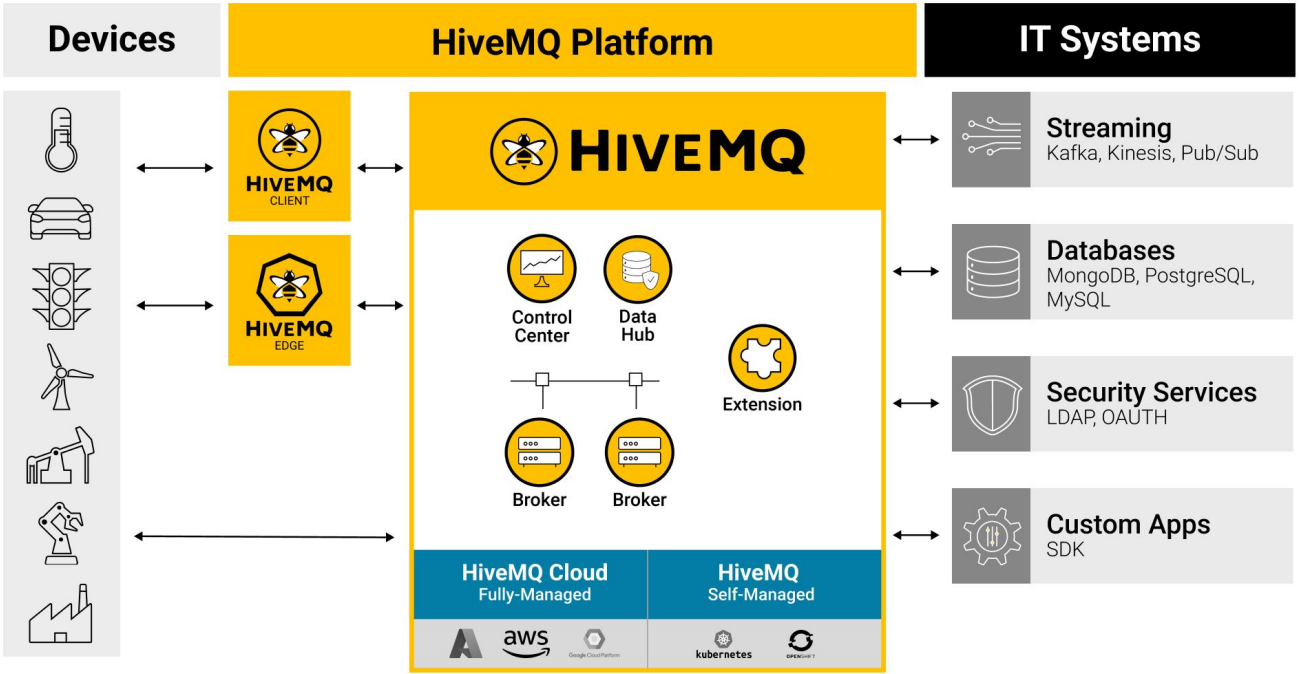
February 2024 Webinar

Speakers

Agenda

- Introduction
- MQTT Standards for Integrating Edge AI Systems
- Demo 1 - Fully Integrated Pattern for Integrating Edge AI into MQTT
- Demo 2 - Unstructured Data Pattern for Integrating Edge AI into MQTT
- Q&A

HIVEMQ | The Enterprise MQTT Platform



Key Industries

Why Standards for MQTT in Edge AI are Critical?

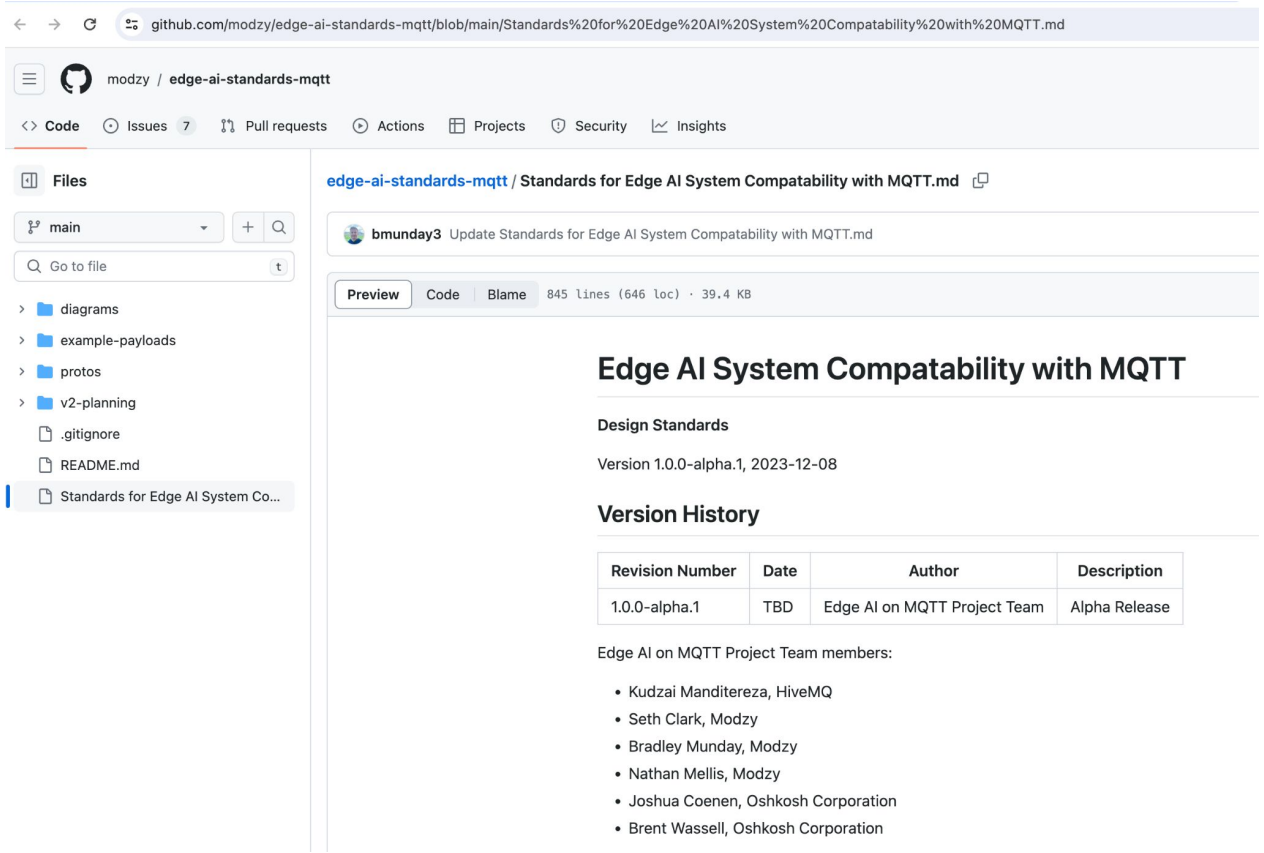
Interoperability and Flexibility

Cost Savings and Accelerated Deployment

Scalability and Consistency

Innovation and Community Engagement

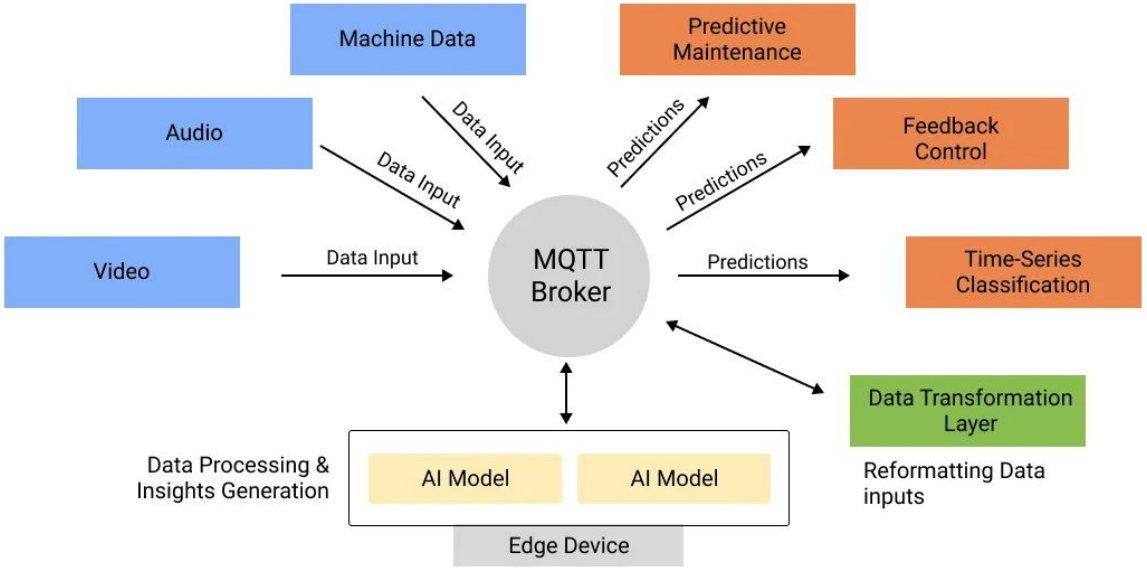
Contributors and Resources



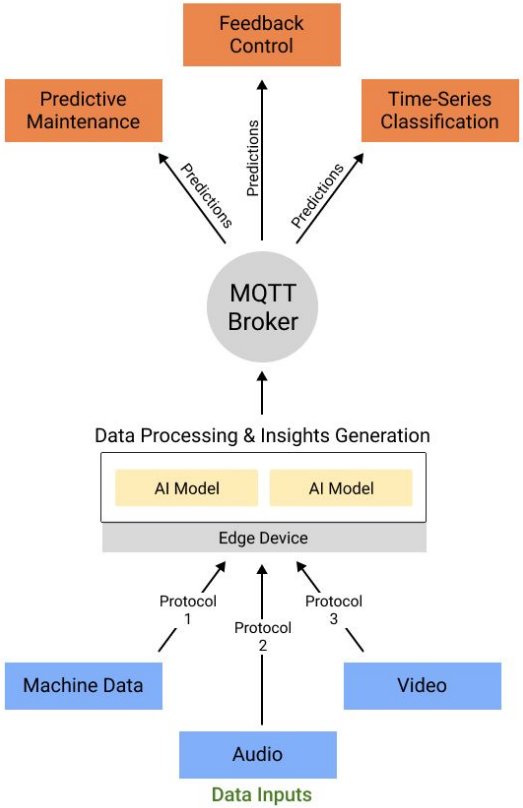
Common Patterns for Edge AI on MQTT

	MODEL INPUT: MQTT TOPIC(S) [SUBSCRIBE]	MODEL INPUT: OTHER PROTOCOLS, DIRECT SENSOR DATA, ETC.
Model Output: MQTT topic(s) [Publish]	The “Fully-integrated” pattern	The “Unstructured Data” pattern
Model Output: Other systems lacking MQTT support	The “Ambassador” pattern	Not covered by this standard

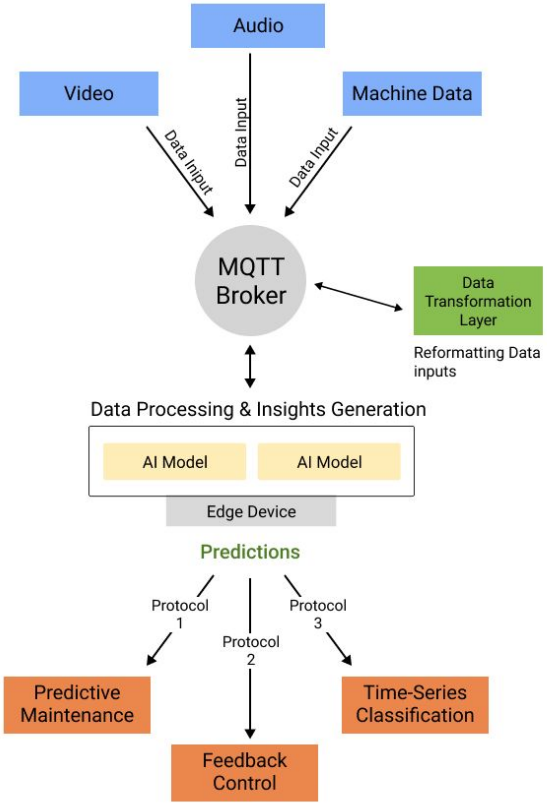
The Fully-Integrated Pattern



The Unstructured Data Pattern



The Ambassador Pattern



Guidelines for MQTT Topic Structure Design

Topic Namespaces for Edge AI

- **Raw Data Namespace** – hold raw sensor data as it is captured. AI model input in the ambassador and fully-connected patterns.
- **Inference Namespace** – hold metrics directly generated by a machine-learning model; do not necessarily provide business value on their own.
- **Insight Namespace** – metrics with innate business value that are composed, calculated, or otherwise built on top of individual inference metrics.

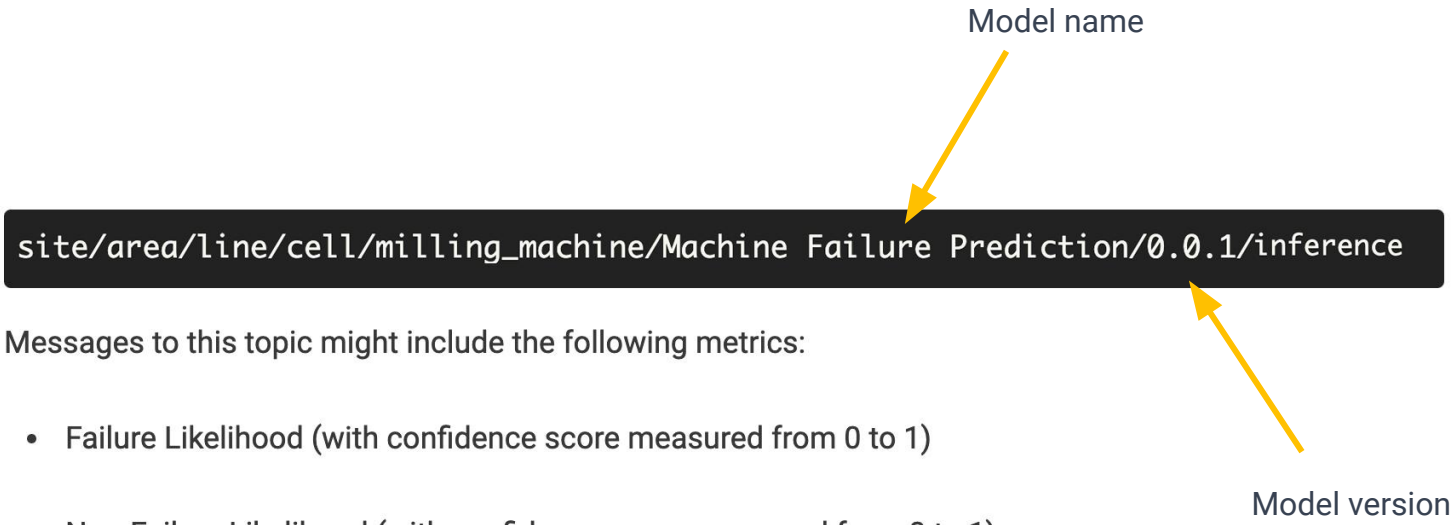
Raw Data namespace Example

```
site/area/line/cell/milling_machine/raw
```

Messages to this topic might include the following metrics:

- Air temperature (K)
- Process temperature (K)
- Rotational speed (rpm)
- Torque (Nm)
- Tool wear (min)

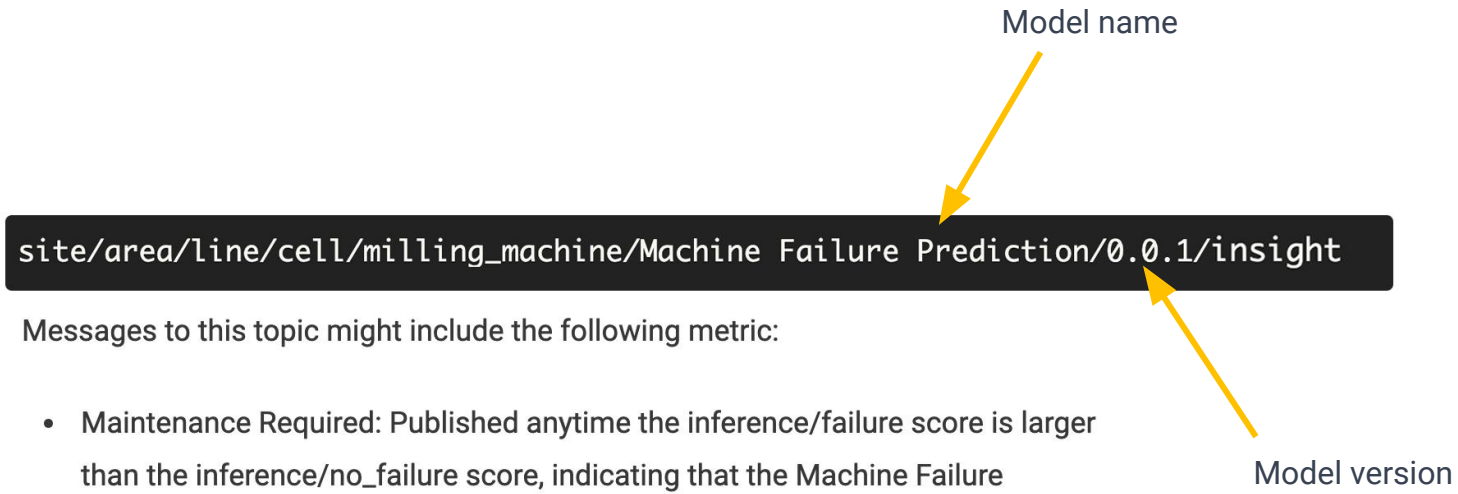
Inference Namespace Example



Messages to this topic might include the following metrics:

- Failure Likelihood (with confidence score measured from 0 to 1)
- Non-Failure Likelihood (with confidence score measured from 0 to 1)

Insight Namespace Example



Messages to this topic might include the following metric:

- Maintenance Required: Published anytime the inference/failure score is larger than the inference/no_failure score, indicating that the Machine Failure Prediction model has found the milling machine to be likely to fail sometime soon.

Flat MQTT Topic Structure

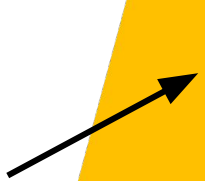
```
[Customized MQTT topic structure]/Edge_DeviceID +/model_name/model_version/inference
```

- **[Customized MQTT topic structure]** – Any existing top-level topic structure for a flat MQTT namespace
- **Edge_DeviceID** – A unique identifier of some kind pointing to the device's location upon which an AI/ML model is running.
- **model_name** – The name of a specific AI/ML model.
- **model_version** – Version of AI/ML model
- **inference** – The namespace element that will contain any messages generated by this version of this model.

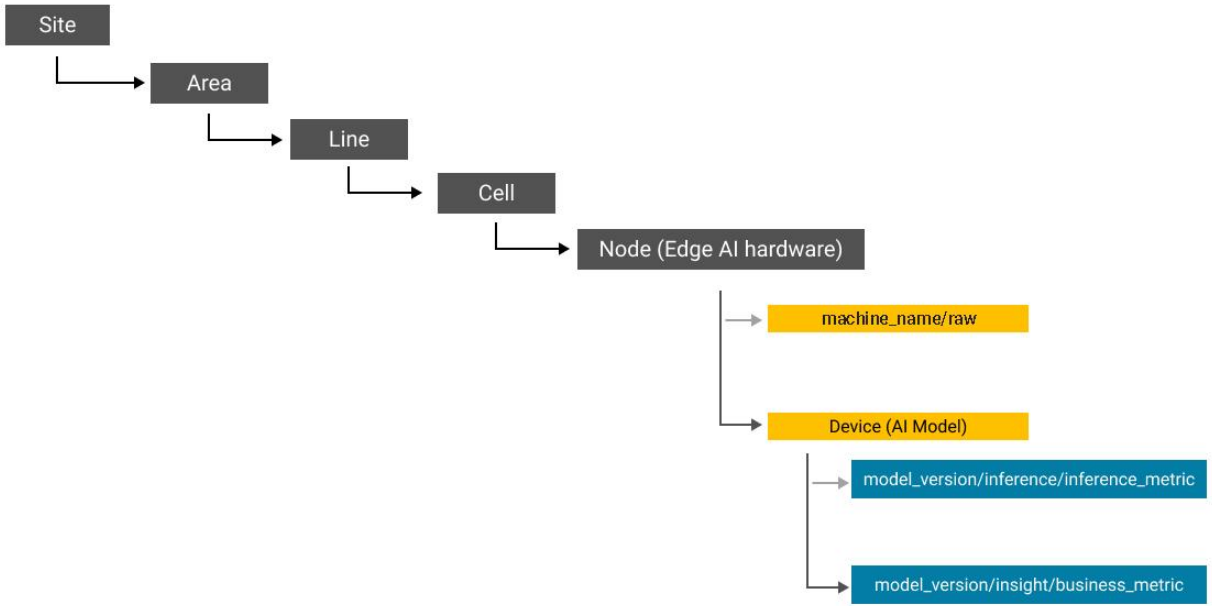
MQTT Sparkplug Topic Structure

```
namespace/group_id/message_type/edge_node_id/[device_id]
```

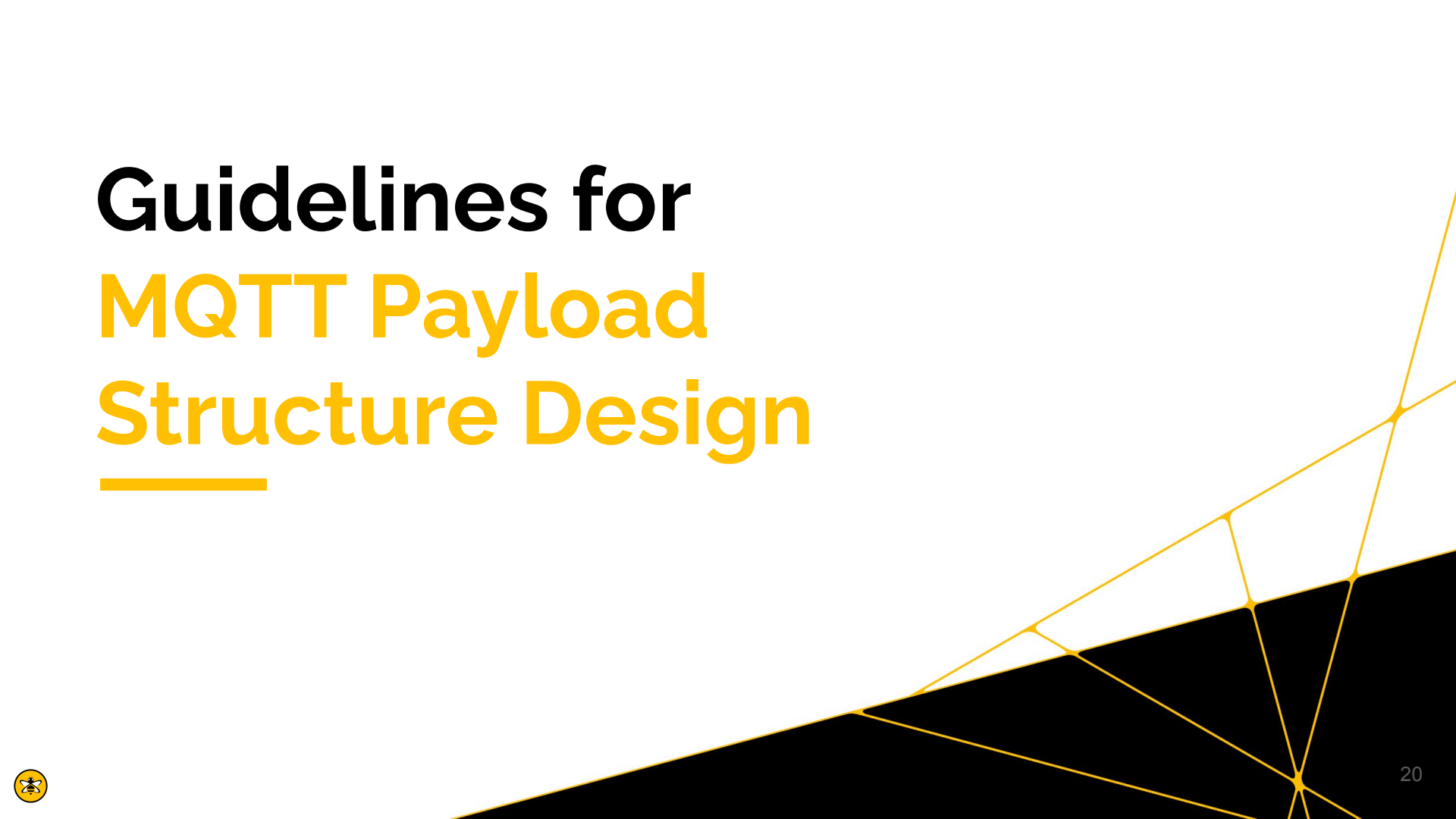
- **namespace** – Sparkplug B root namespace
- **group_id** – Recommends using a concatenation of the ISA-95 format of site:area:line:cell
- **message_type** – Edge AI apps will primarily publish results using the DDATA message type, but other message types supported.
- **edge_node_id** – Identity of the computing hardware that is running models or communicates to the MQTT broker.
- **device_id** – Identity of the AI/ML model that is generating inferences.



Unified namespace Snapshot



Guidelines for MQTT Payload Structure Design



Structured Payloads for Edge AI Outputs

- **Predictions** –regarding potential future events.
- **Structured insights** – (Classification or Detection) that are extracted from unstructured sources like audio and video.

Recommended Formatting

- **Protobuf** –should be used whenever possible for its efficiency and future-proofing qualities
- **JSON** – Should be used for compatibility with non-protobuf capable systems.

Flat MQTT Payload Template


```
{
   "identifier":"inference-2HYZh8a4jtFi3xFc4e3TWRmclff",
   "model":{
      "identifier":"brzrip6cxk",
      "version":"0.0.1",
      "name":"Machine Failure Prediction"
   },
   "tags":{
      "sourceTopic":"site:area:line:cell/node/device/raw/sensor_name",
      "sourceMessageID":"abcd1234",
      "inputSizeInBytes":32,
      "inputSha256Digest":"be01ef104fb88fd151132733e746fe29b997348bf34be875e25ba48c0d7436ca"
   },
   "resultType":"classPredictions",
   "result":{
      "classPredictions":[
         {
            "className":"no_failure",
            "score":0.974
         },
         {
            "className":"failure",
            "score":0.026
         }
      ]
   },
   "explaination":{

   }
}
```

MQTT Sparkplug DBIRTH Payload


```
{
   "timestamp":1486144502122,
   "metrics":[
      {
         "name":"0.0.1/inference/identifier",
         "timestamp":1486144502122,
         "dataType":"string",
         "value":"inference-2HYZh8a4jtFi3xFc4e3TWRmclff"
      },
      {
         "name":"0.0.1/inference/model/identifier",
         "timestamp":1486144502122,
         "dataType":"string",
         "value":"brzrip6cxk"
      },
      {
         "name":"0.0.1/inference/model/version",
         "timestamp":1486144502122,
         "dataType":"string",
         "value":"0.0.1"
      },
      {
         "name":"0.0.1/inference/model/name",
         "timestamp":1486144502122,
         "dataType":"string",
         "value":"Machine Failure Prediction"
      },
      {
         "name":"0.0.1/inference/tags/sourceTopic",
         "timestamp":1486144502122,
         "dataType":"string",
         "value":"site:area:line:cell/node/device/raw/sensor_name"
      },
```

MQTT Sparkplug Host Application

Metric		Value	Data Type
Site:Area:Line:Cell	/group_id		
└*Edge Node ID*	/edge_node_id		
└Edge Node Device ID	/device_id		
└Inference			
└identifier		inference-2HYZh8a4jtFi3xFc4e3TWRmclff	String
└model			
└identifier		brzrip6cxk	String
└version		0.0.1	String
└name		Machine Failure Prediction	String
└tags			
└sourceTopic		site:area:line:cell/node/device/raw/sensor_name	String
└dataType		abcd1234	String
└inputSizeInBytes		32	Integer
└inputSha256Digest		be01ef104fb88fd1...	String
└resultType		classPredictions	String
└result		{"classPredictions":[{"className":"no_ failure","score":0.974}, {"className":"failure","score":0.026}]}	String

MQTT Sparkplug DDATA Payload


```
{
   "timestamp":1486144502122,
   "metrics":[
      {
         "name":"0.0.1/inference/identifier",
         "timestamp":1486144502122,
         "dataType":"string",
         "value":"inference-2HYZh8a4jtFi3xFc4e3TWRmclff"
      },
      {
         "name":"0.0.1/inference/tags/sourceTopic",
         "timestamp":1486144502122,
         "dataType":"string",
         "value":"site:area:line:cell/node/device/raw/sensor_name"
      },
      {
         "name":"0.0.1/inference/tags/sourceMessageID",
         "timestamp":1486144502122,
         "dataType":"string",
         "value":"abcd1234"
      },
      {
         "name":"0.0.1/inference/tags/inputSizeInBytes",
         "timestamp":1486144502122,
         "dataType":"integer",
         "value":32
      },
      {
         "name":"0.0.1/inference/tags/inputSha256Digest",
         "timestamp":1486144502122,
         "dataType":"string",
         "value":"2o3nvi30fh4fb88fd120932733e746fe29b99732ifhi34be875e25ba48c0d7436ca"
      },
      {
         "name":"0.0.1/inference/result",
         "timestamp":1486144502122,
         "dataType":"string",
         "value":'{"classPredictions":[{"className":"no_failure","score":0.087},{"className":"failure","score":0.913
      }
   ],
   "seq":1
}
```

Edge AI/ML
Model Result Formats

Classification

Classification assigns a class to an individual piece of data. This might be useful for classifying an individual image, audio snippet, video frame, or a piece of machine data.

```
{
   "classPredictions":[
      {
         "class":"className",
         "score":1.0
      }
   ]
}
```

Multi-Classification

Similar to classification models, but used when model outputs are grouped into more than two distinct classes.


```
{
   "classifications":[
      {
         "classPredictions":[
            {
               "class":"className",
               "score":1.0
            }
         ]
      }
   ]
}
```

Object Detection

Object detection is used to identify regions of interest within an image or video that are defined by a bounding box. Bounding boxes can have one or more classifications, and images can have one or more bounding boxes.


```
{
   "detections":[
      {
         "class":"className",
         "score":1.0,
         "boundingBox":{
            "x":100,
            "y":200,
            "width":300,
            "height":400
         }
      }
   ]
}
```


Named Entity Recognition

Named entity recognition is used to identify unique entities, such as names, organizations, and locations, within a larger corpus of text.


```
{
   "entities":[
      {
         "entityGroup":"B-LOC",
         "score":1.0,
         "textSpan":{
            "start":0,
            "end":5,
            "text":"Paris is a city."
         }
      }
   ]
}
```

Demos

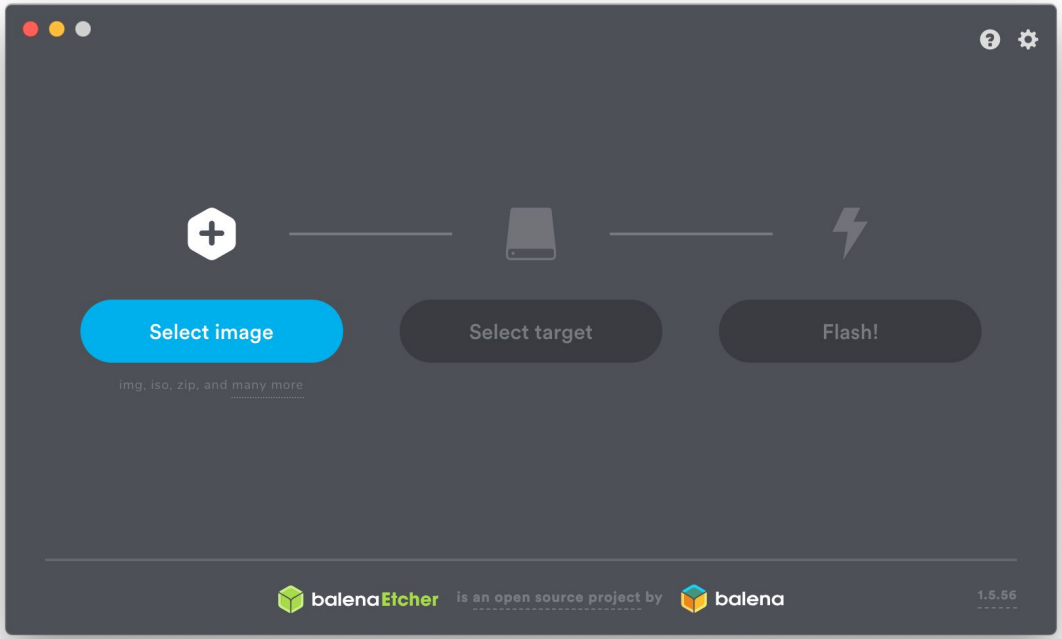
Fully Integrated Pattern - Marc Pous, balena.io

Unstructured Data Pattern - Magnus McCune, HiveMQ

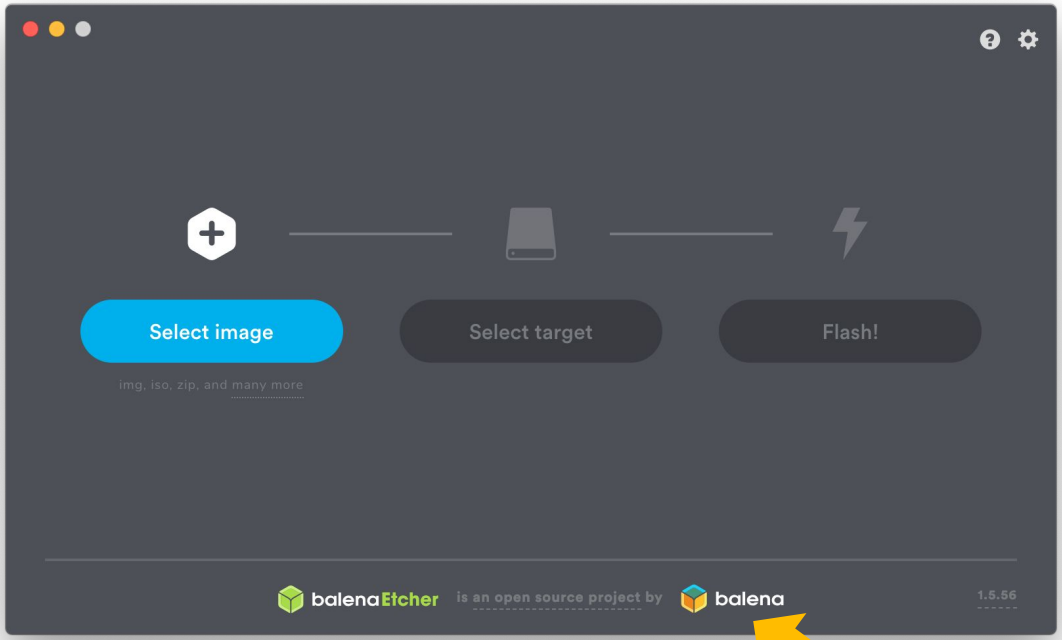
Demo 1: The Fully-Integrated Pattern with HiveMQ Edge AI balena

Marc Pous
IoT Giant & Developer Advocate

Anyone?


Select image
img, iso, zip, and many more
Select target
Flash!
balenaEtcher
is an open source project by
balena
1.5.56

Anyone?

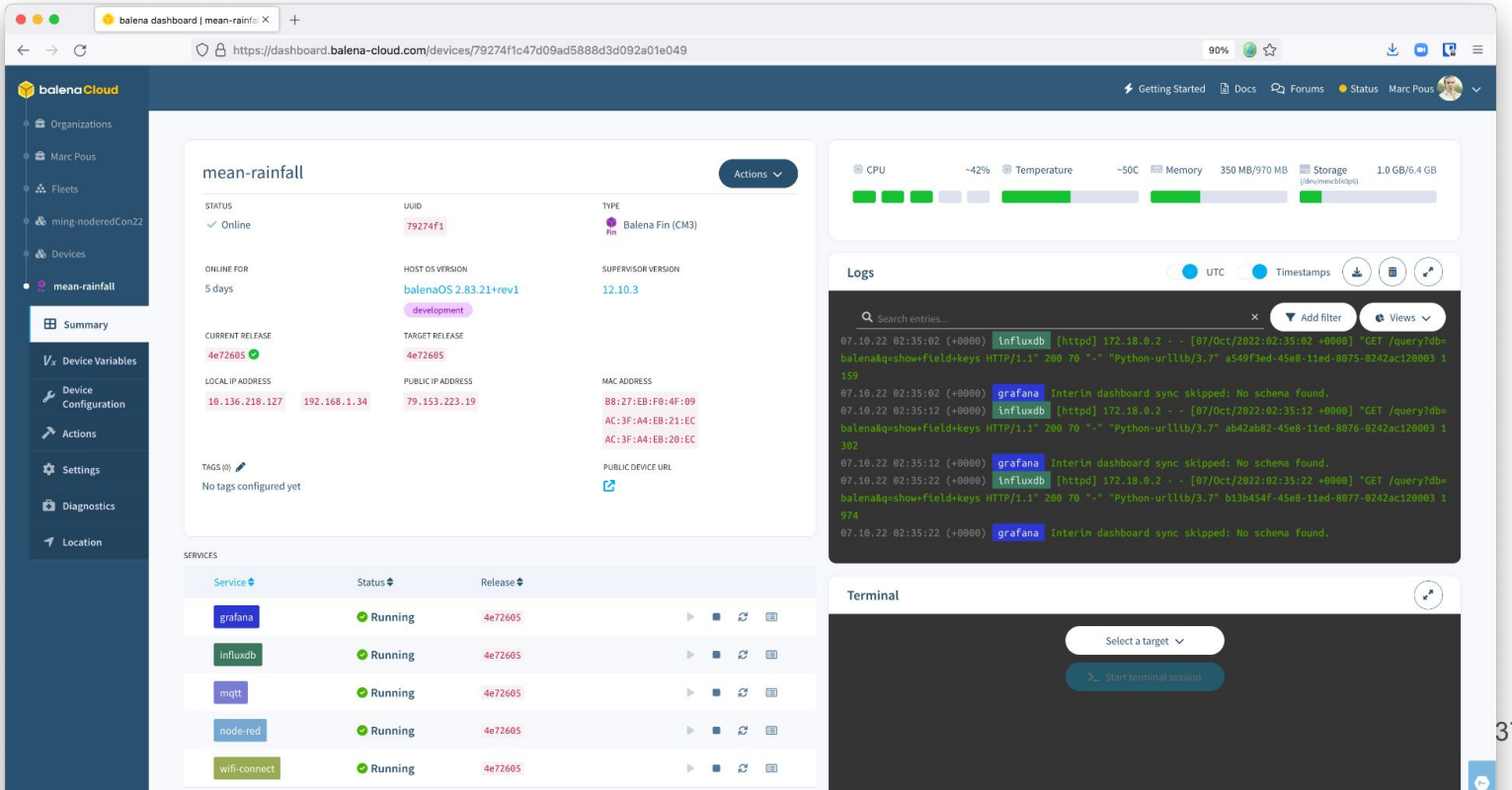

Select image
img, iso, zip, and many more
Select target
Flash!
balenaEtcher
is an open source project by
balena
1.5.56

What is balena?

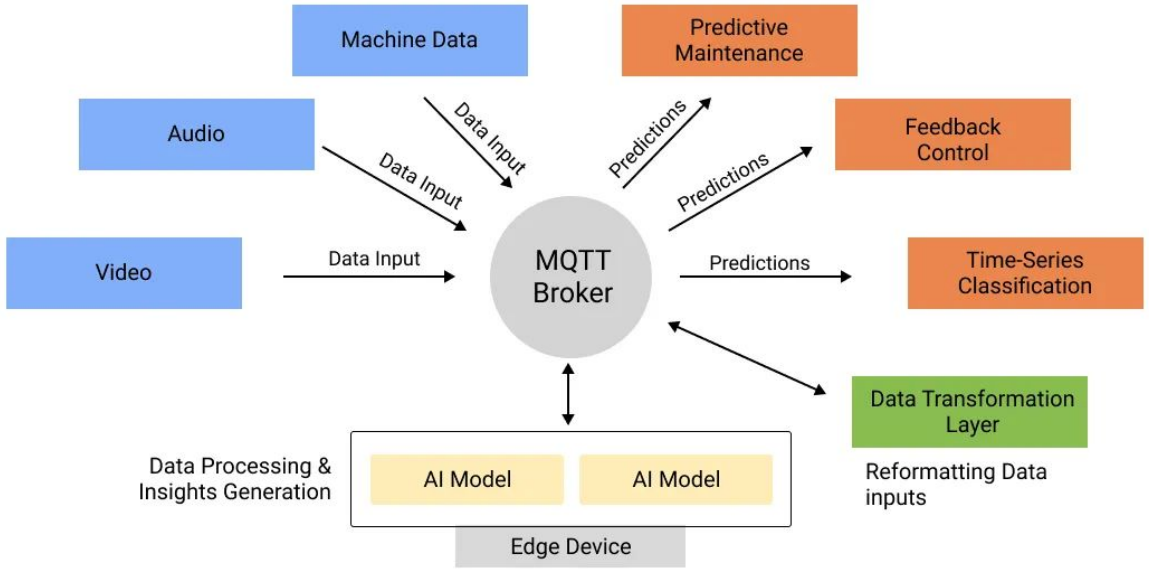
balena

A secure container-based technology stack that enables you to **develop**, **deploy**, **manage** and **scale** large fleets of IoT Linux devices at any stage.

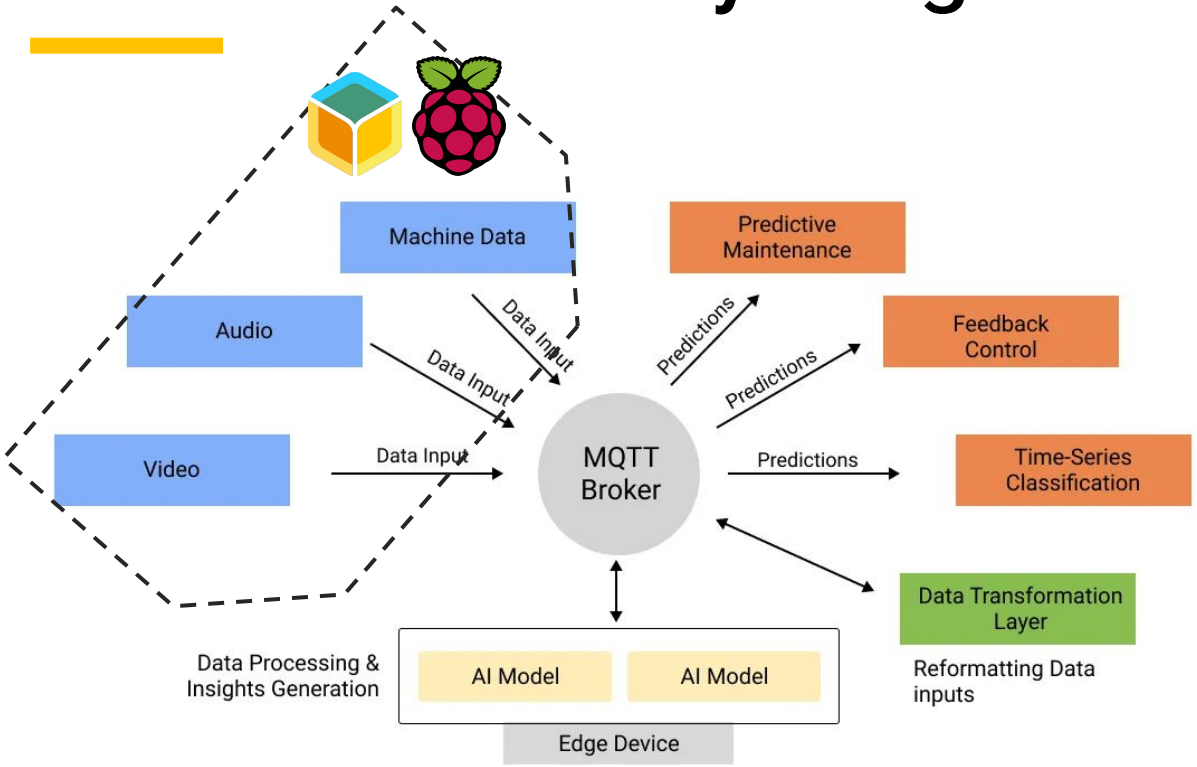
What is balena?



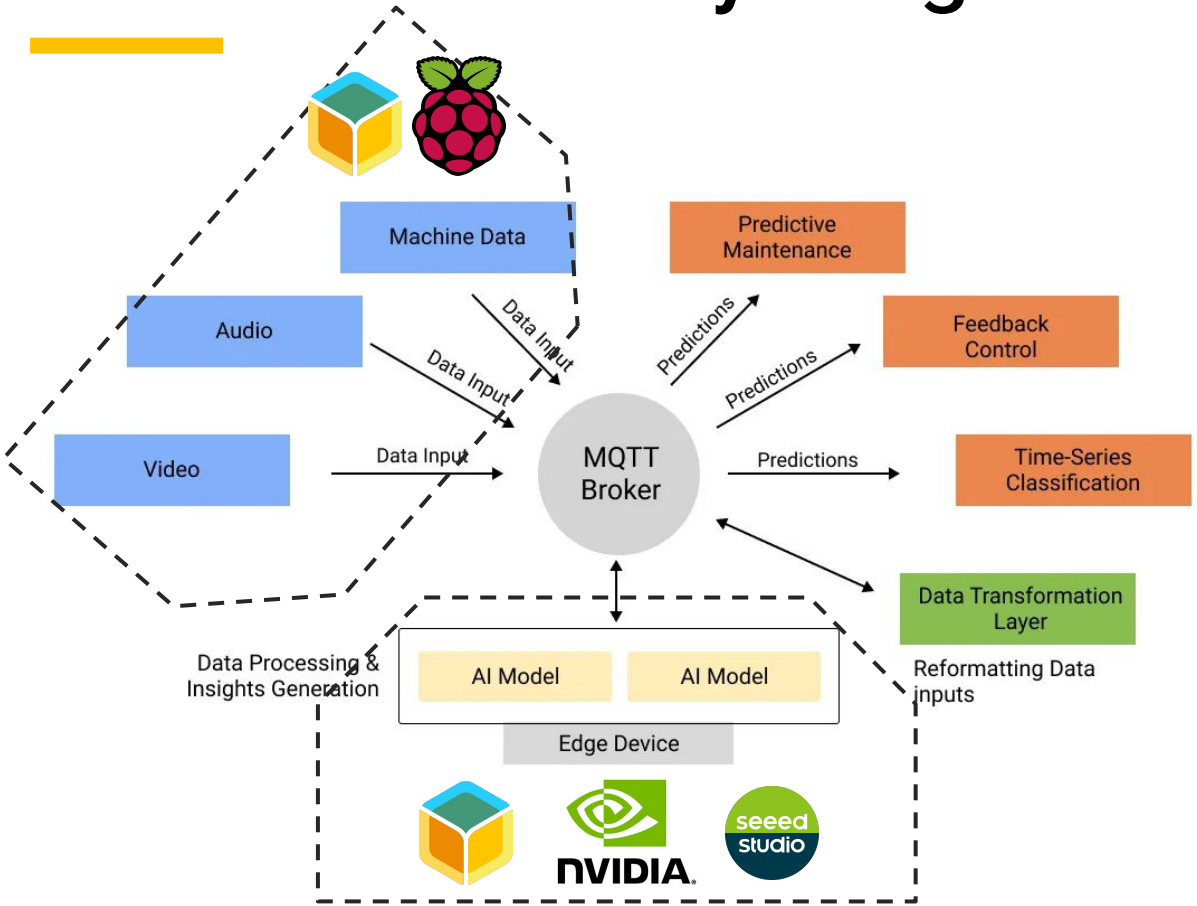
balena Demo - Fully-Integrated Pattern



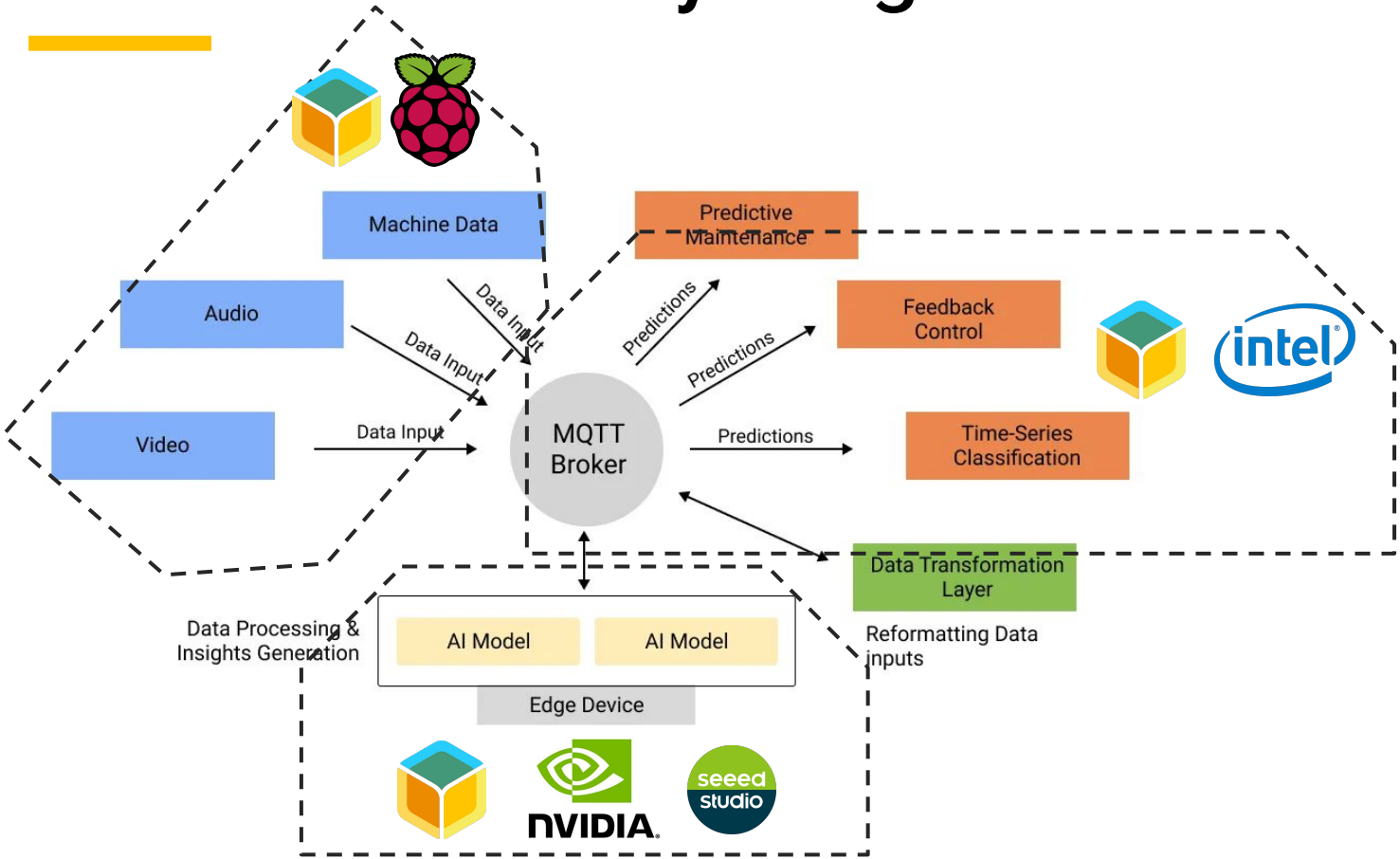
balena Demo - Fully-Integrated Pattern



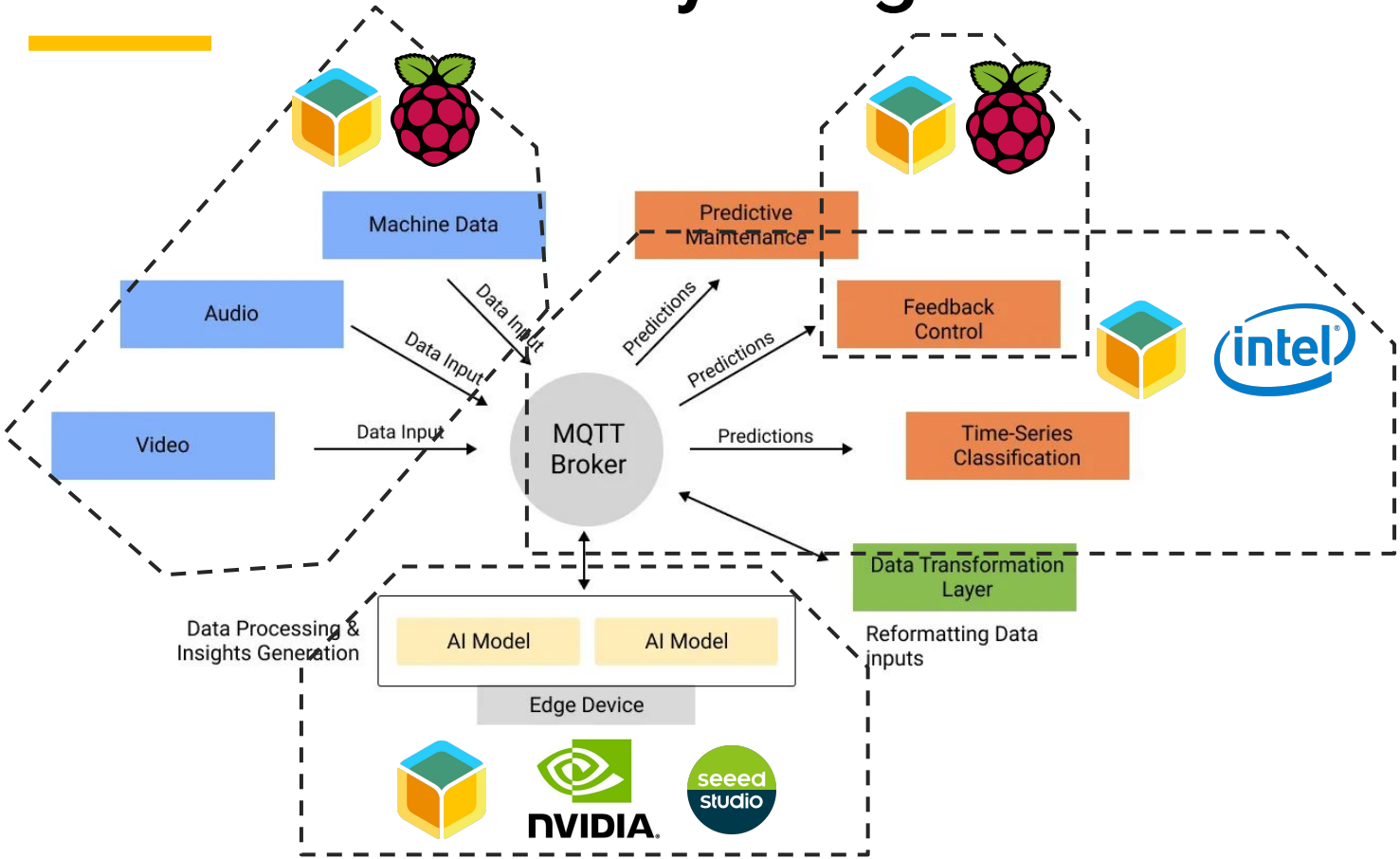
balena Demo - Fully-Integrated Pattern



balena Demo - Fully-Integrated Pattern



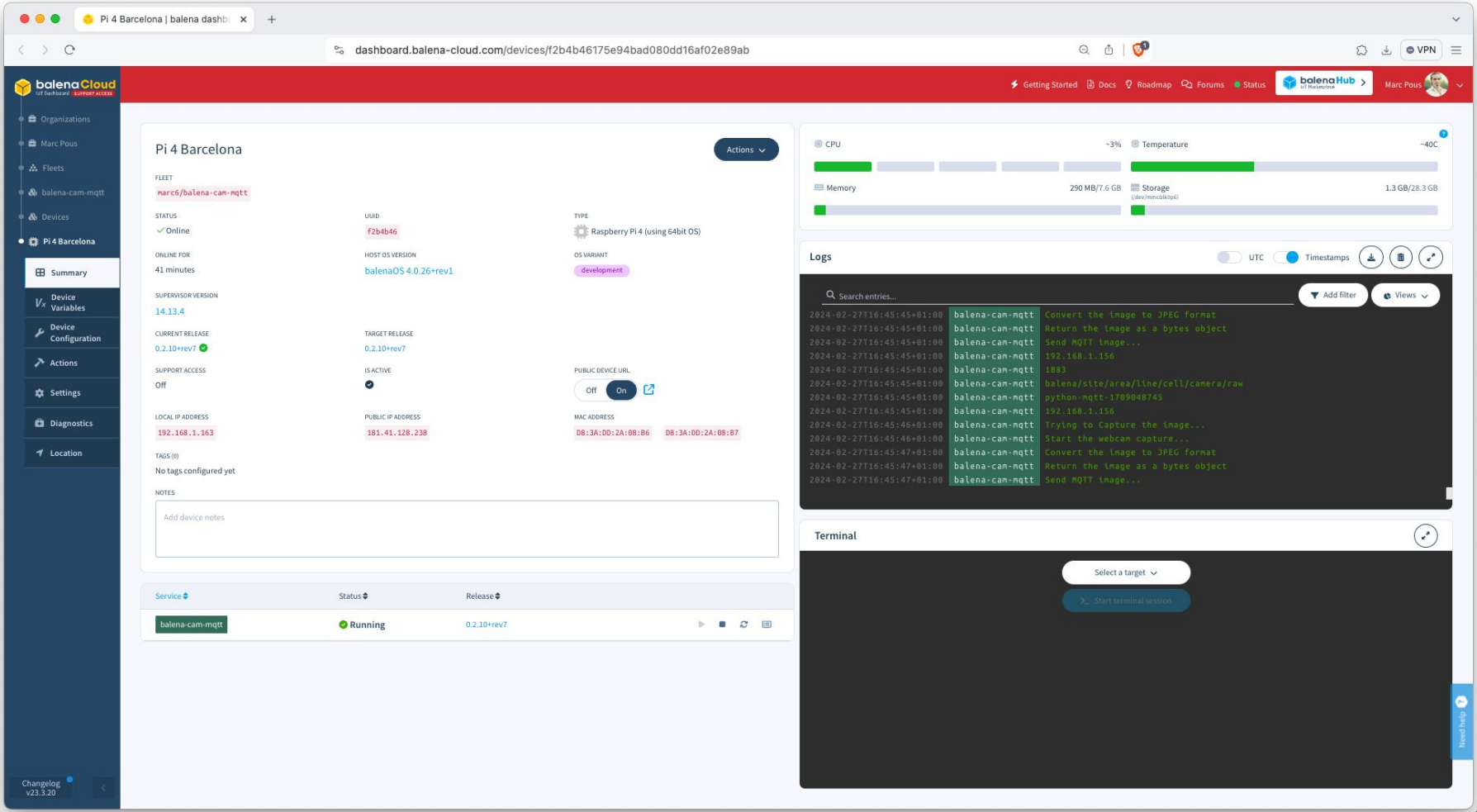
balena Demo - Fully-Integrated Pattern

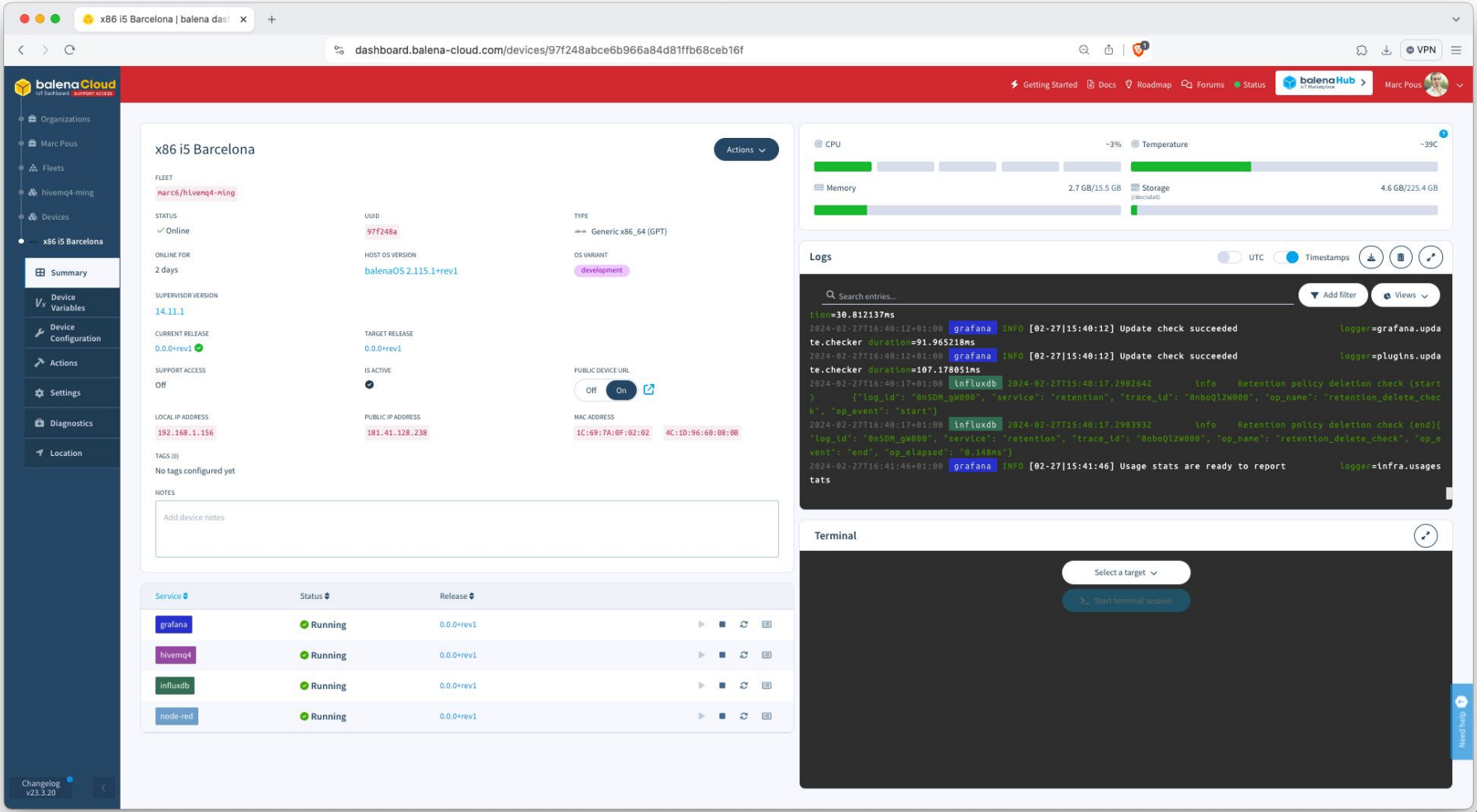


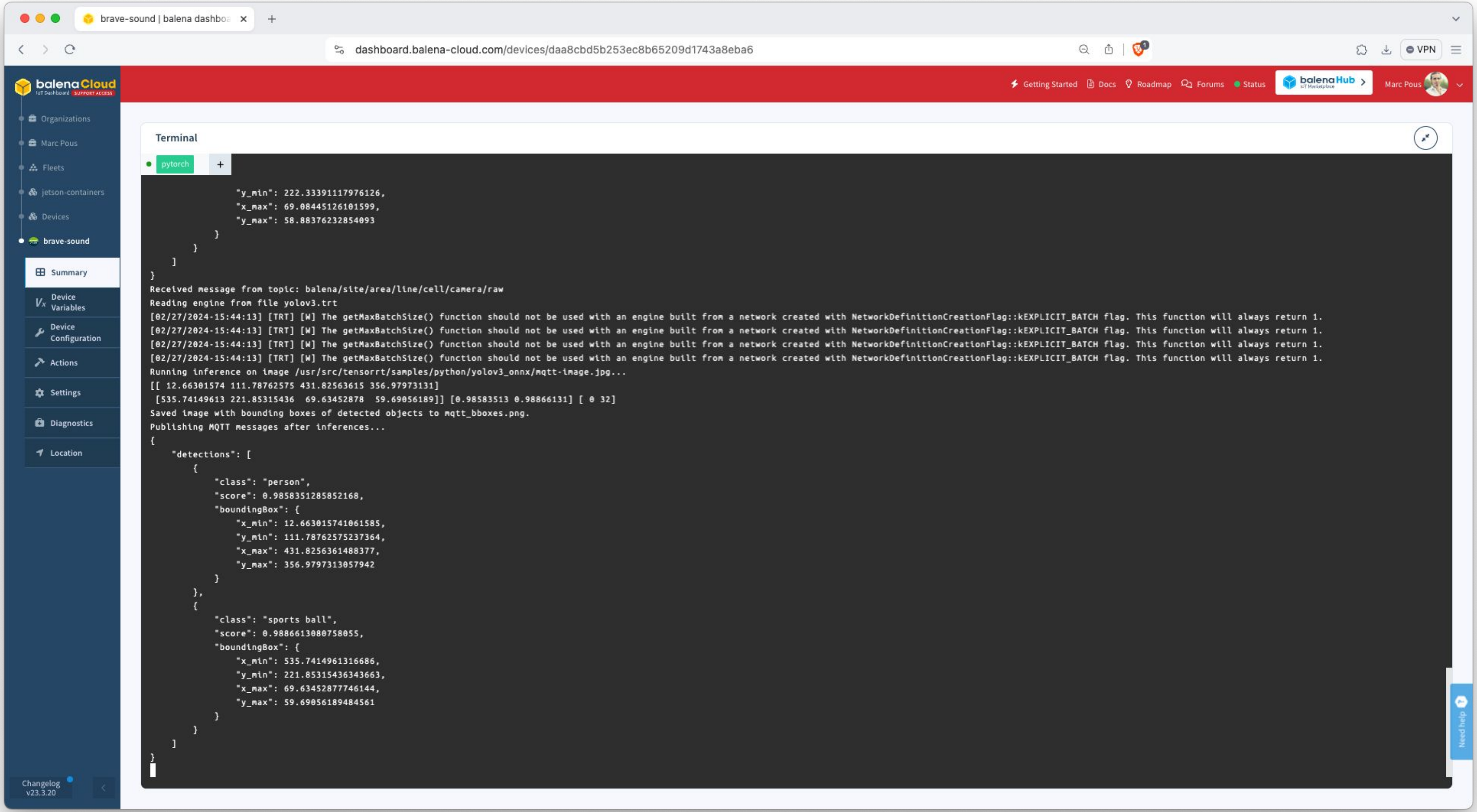
balena Demo

- **balenaCam MQTT** – Raspberry Pi Camera publishing over MQTT.
  https://github.com/mpous/balena-cam-mqtt
- **MING broker** – Intel NUC with the MING stack.
  https://github.com/mpous/ming
- **Seeed Studio ReComputer J4012 / NVIDIA Jetson Orin 16GB** – Edge AI device with TensorRT AI model to recognize real-time images.
  https://github.com/mpous/J4012-pytorch-mqtt

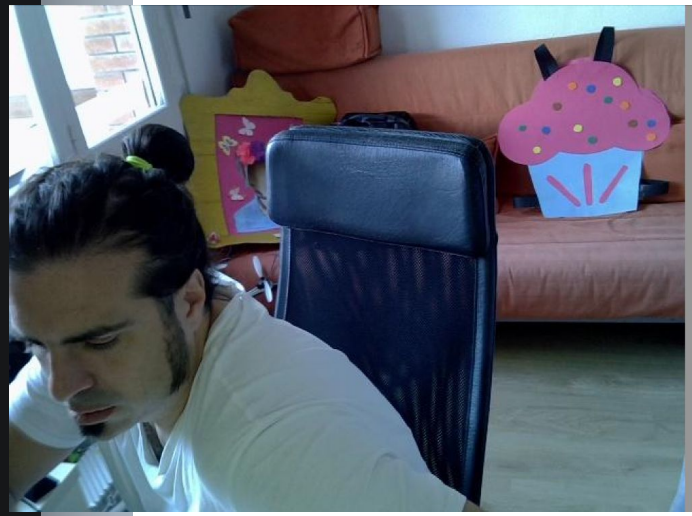
Let me show you more …


Pi 4 Barcelona | balena dashb
dashboard.balena-cloud.com/devices/f2b4b46175e94bad080dd16af02e89ab
VPN
balenaCloud
Getting Started
Docs
Roadmap
Forums
Status
balenaHub
Marc Pous
Organizations
Marc Pous
Fleets
balena-cam-mqtt
Devices
Pi 4 Barcelona
Summary
Device Variables
Device Configuration
Actions
Settings
Diagnostics
Location
Changelog v23.3.20
Pi 4 Barcelona
Actions
FLEET
marc6/balena-cam-mqtt
STATUS
Online
UUID
f2b4b46
TYPE
Raspberry Pi 4 (using 64bit OS)
ONLINE FOR
41 minutes
HOST OS VERSION
balenaOS 4.0.26+rev1
OS VARIANT
development
SUPERVISOR VERSION
14.13.4
CURRENT RELEASE
0.2.10+rev7
TARGET RELEASE
0.2.10+rev7
SUPPORT ACCESS
Off
IS ACTIVE
PUBLIC DEVICE URL
Off
On
LOCAL IP ADDRESS
192.168.1.163
PUBLIC IP ADDRESS
181.41.128.238
MAC ADDRESS
D8:3A:DD:2A:08:B6
D8:3A:DD:2A:08:B7
TAGS (0)
No tags configured yet
NOTES
Add device notes
Service
Status
Release
balena-cam-mqtt
Running
0.2.10+rev7
CPU
~3%
Temperature
~40C
Memory
290 MB/7.6 GB
Storage
1.3 GB/28.3 GB
Logs
UTC
Timestamps
Search entries...
Add filter
Views
2024-02-27T16:45:45+01:00 balena-cam-mqtt Convert the image to JPEG format
2024-02-27T16:45:45+01:00 balena-cam-mqtt Return the image as a bytes object
2024-02-27T16:45:45+01:00 balena-cam-mqtt Send MQTT image...
2024-02-27T16:45:45+01:00 balena-cam-mqtt 192.168.1.156
2024-02-27T16:45:45+01:00 balena-cam-mqtt 1883
2024-02-27T16:45:45+01:00 balena-cam-mqtt balena/site/area/line/cell/camera/raw
2024-02-27T16:45:45+01:00 balena-cam-mqtt python-mqtt-1709048745
2024-02-27T16:45:45+01:00 balena-cam-mqtt 192.168.1.156
2024-02-27T16:45:46+01:00 balena-cam-mqtt Trying to Capture the image...
2024-02-27T16:45:46+01:00 balena-cam-mqtt Start the webcam capture...
2024-02-27T16:45:47+01:00 balena-cam-mqtt Convert the image to JPEG format
2024-02-27T16:45:47+01:00 balena-cam-mqtt Return the image as a bytes object
2024-02-27T16:45:47+01:00 balena-cam-mqtt Send MQTT image...
Terminal
Select a target
Start terminal session
Need help


x86 i5 Barcelona | balena das
dashboard.balena-cloud.com/devices/97f248abce6b966a84d81ffb68ceb16f
VPN
balenaCloud
Getting Started
Docs
Roadmap
Forums
Status
balenaHub
Marc Pous
Organizations
Marc Pous
Fleets
hivemq4-ming
Devices
x86 i5 Barcelona
Summary
Device Variables
Device Configuration
Actions
Settings
Diagnostics
Location
Changelog v23.3.20
x86 i5 Barcelona
Actions
FLEET
marc6/hivemq4-ming
STATUS
Online
UUID
97f248a
TYPE
Generic x86_64 (GPT)
ONLINE FOR
2 days
HOST OS VERSION
balenaOS 2.115.1+rev1
OS VARIANT
development
SUPERVISOR VERSION
14.11.1
CURRENT RELEASE
0.0.0+rev1
TARGET RELEASE
0.0.0+rev1
SUPPORT ACCESS
Off
IS ACTIVE
PUBLIC DEVICE URL
Off
On
LOCAL IP ADDRESS
192.168.1.156
PUBLIC IP ADDRESS
181.41.128.238
MAC ADDRESS
1C:69:7A:0F:02:02
4C:1D:96:60:08:0B
TAGS (0)
No tags configured yet
NOTES
Add device notes
Service
Status
Release
grafana
Running
0.0.0+rev1
hivemq4
Running
0.0.0+rev1
influxdb
Running
0.0.0+rev1
node-red
Running
0.0.0+rev1
CPU
~3%
Temperature
~39C
Memory
2.7 GB/15.5 GB
Storage
4.6 GB/225.4 GB
Logs
UTC
Timestamps
Search entries...
Add filter
Views
tion=30.812137ms
2024-02-27T16:40:12+01:00 grafana INFO [02-27|15:40:12] Update check succeeded logger=grafana.upda
te.checker duration=91.965218ms
2024-02-27T16:40:12+01:00 grafana INFO [02-27|15:40:12] Update check succeeded logger=plugins.upda
te.checker duration=107.178051ms
2024-02-27T16:40:17+01:00 influxdb 2024-02-27T15:40:17.290264Z info Retention policy deletion check (start
} {"log_id": "0nSDM_gW000", "service": "retention", "trace_id": "0nboQl2W000", "op_name": "retention_delete_chec
k", "op_event": "start"}
2024-02-27T16:40:17+01:00 influxdb 2024-02-27T15:40:17.290393Z info Retention policy deletion check (end){
"log_id": "0nSDM_gW000", "service": "retention", "trace_id": "0nboQl2W000", "op_name": "retention_delete_check", "op_e
vent": "end", "op_elapsed": "0.148ms"}
2024-02-27T16:41:46+01:00 grafana INFO [02-27|15:41:46] Usage stats are ready to report logger=infra.usages
tats
Terminal
Select a target
Start terminal session
Need help


brave-sound | balena dashboard
dashboard.balena-cloud.com/devices/daa8cbd5b253ec8b65209d1743a8eba6
balenaCloud
Getting Started
Docs
Roadmap
Forums
Status
balenaHub
Marc Pous
Organizations
Marc Pous
Fleets
jetson-containers
Devices
brave-sound
Summary
Device Variables
Device Configuration
Actions
Settings
Diagnostics
Location
Changelog v23.3.20
Terminal
pytorch
"y_min": 222.33391117976126,
"x_max": 69.08445126101599,
"y_max": 58.88376232854093
}
}
]
}
Received message from topic: balena/site/area/line/cell/camera/raw
Reading engine from file yolov3.trt
[02/27/2024-15:44:13] [TRT] [W] The getMaxBatchSize() function should not be used with an engine built from a network created with NetworkDefinitionCreationFlag::kEXPLICIT_BATCH flag. This function will always return 1.
[02/27/2024-15:44:13] [TRT] [W] The getMaxBatchSize() function should not be used with an engine built from a network created with NetworkDefinitionCreationFlag::kEXPLICIT_BATCH flag. This function will always return 1.
[02/27/2024-15:44:13] [TRT] [W] The getMaxBatchSize() function should not be used with an engine built from a network created with NetworkDefinitionCreationFlag::kEXPLICIT_BATCH flag. This function will always return 1.
[02/27/2024-15:44:13] [TRT] [W] The getMaxBatchSize() function should not be used with an engine built from a network created with NetworkDefinitionCreationFlag::kEXPLICIT_BATCH flag. This function will always return 1.
Running inference on image /usr/src/tensorrt/samples/python/yolov3_onnx/mqtt-image.jpg...
[[12.66301574 111.78762575 431.82563615 356.97973131]
[535.74149613 221.85315436 69.63452878 59.69056189]] [0.98583513 0.98866131] [0 32]
Saved image with bounding boxes of detected objects to mqtt_bboxes.png.
Publishing MQTT messages after inferences...
{
"detections": [
{
"class": "person",
"score": 0.9858351285852168,
"boundingBox": {
"x_min": 12.663015741061585,
"y_min": 111.78762575237364,
"x_max": 431.8256361488377,
"y_max": 356.9797313057942
}
},
{
"class": "sports ball",
"score": 0.9886613080758055,
"boundingBox": {
"x_min": 535.7414961316686,
"y_min": 221.85315436343663,
"x_max": 69.63452877746144,
"y_max": 59.69056189484561
}
}
]
}

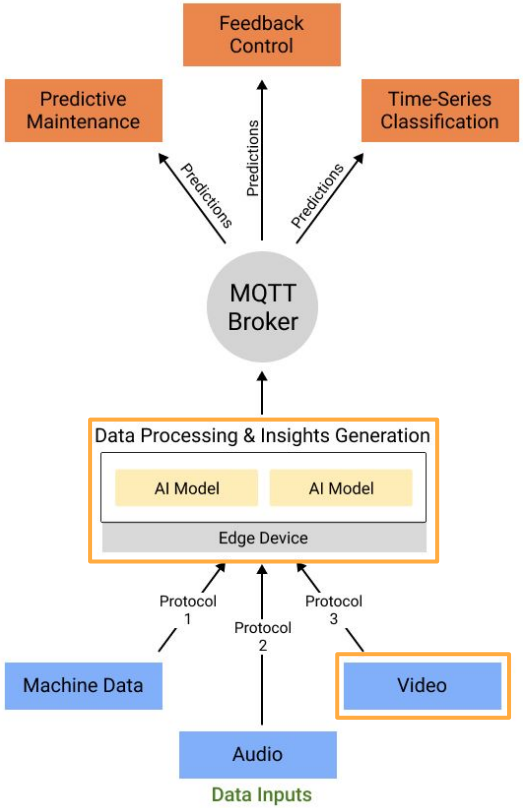
```
{
    "detections": [
        {
            "class": "person",
            "score": 0.9657025785242721,
            "boundingBox": {
                "x_min": 3.7658301071171607,
                "y_min": 150.20850711183255,
                "x_max": 447.67390142228925,
                "y_max": 313.51492320512546
            }
        },
        {
            "class": "sports ball",
            "score": 0.9867270698056363,
            "boundingBox": {
                "x_min": 536.148606279469,
                "y_min": 222.20140379521948,
                "x_max": 68.18516925408174,
                "y_max": 59.01444510192577
            }
        }
    ]
}
```

Demo 2: The Unstructured Data Pattern with HiveMQ

Magnus McCune
Senior IoT Solutions Architect

The Unstructured Data Pattern



Demo - Unstructured Data Pattern

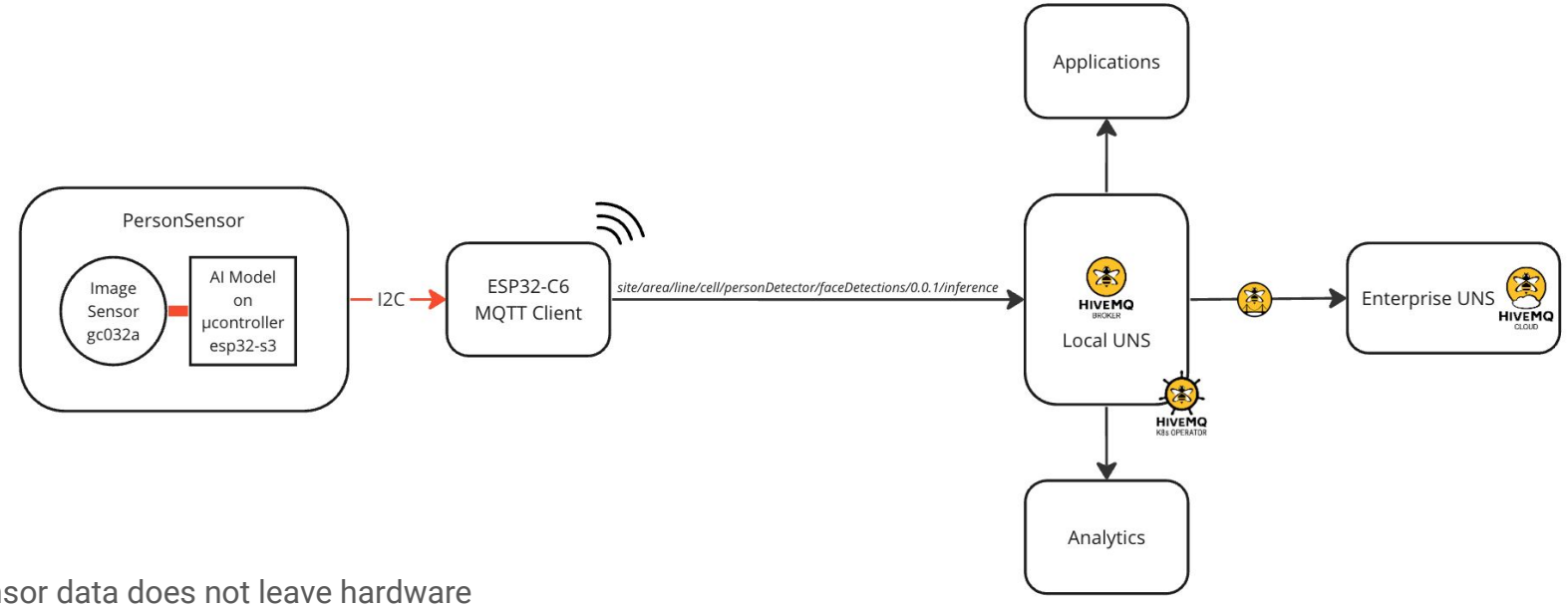


*Image sensor data does not leave hardware

	ESP32-S3	ESP32-C6
Processor	Tensilica Xtensa 32 bit 240MHz dual-core	RISC V 32 bit 160MHz
SRAM	520KB	512KB
ROM	384KB	320KB
WiFi	WiFi4 - Absent	WiFi6

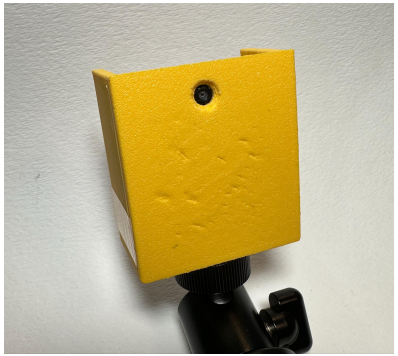
Output from Useful Sensor PersonSensor

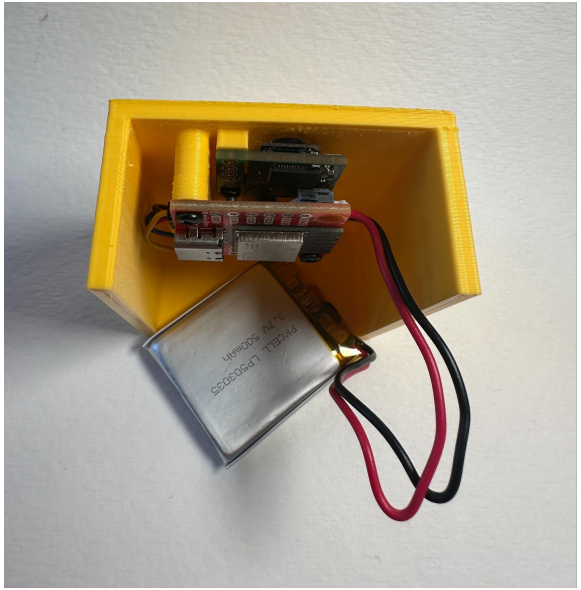

```
{
  "facesDetected": 2,
  "detections": [
    {
      "class": "face0",
      "score": 99,
      "facing": false,
      "boundingBox": {
        "x": 92,
        "y": 59,
        "width": 43,
        "height": 62
      }
    },
    {
      "class": "face1",
      "score": 96,
      "facing": true,
      "boundingBox": {
        "x": 148,
        "y": 74,
        "width": 34,
        "height": 57
      }
    }
  ]
}
```

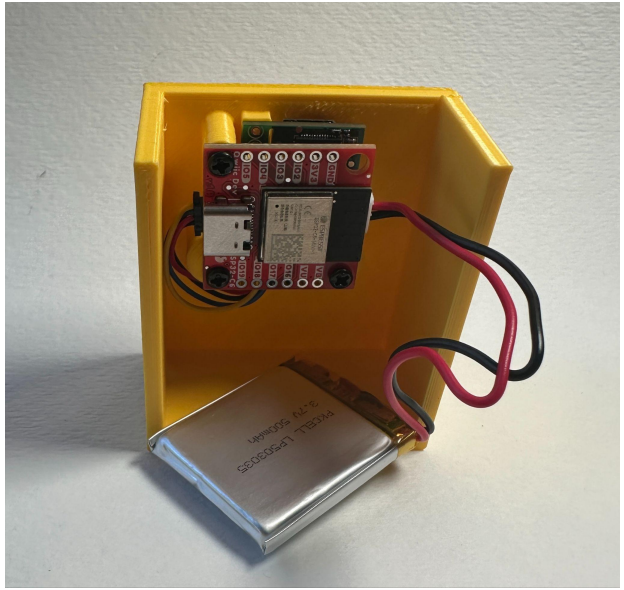

Inference - JSON payload following the format defined in spec with the full output from the AI Model

Insight - An example insight using this data might involve simple business logic calculation that compares the number of faces currently present to an expected number of faces.

Demo







Demo

Not part of the EdgeAI spec, but it is always a good practice to include device state in our namespace.

Demo

demo/halifax/area/line/cell/personDetector/faceDetections/0.0.1/inference


```
{
  "detections": [],
  "facesDetected": 0
}
```


When no one is in frame, the `detections` array is empty and the `facesDetected` metric reads 0

Report By Exception written into the logic ensures that no repeat messages are published if the payloads would otherwise be identical

Demo

demo/halifax/area/line/cell/personDetector/faceDetections/0.0.1/inference


```
{
  "detections": [
    {
      "class": "face0",
      "score": 65,
      "facing": false,
      "boundingBox": {
        "x": 112,
        "y": 91,
        "width": 36,
        "height": 62
      }
    }
  ],
  "facesDetected": 1
}
```


When a person is within frame the *facesDetected* metric indicates the number of identified faces and the *detections* array contains an object for each detected face.

The face0 object contains a confidence score, the *facing* bool and a nested object describing the bounding box.

Thank You!

Q&A