Ask the Expert

How Do I Streamline AI Project Workflows?

Daniele Cazzari, Sr Manager, IoT Product Innovation & Engineering
Rik de Ruiter, Senior Systems Architect

Copyright © SAS Institute Inc. All rights reserved.

Daniele Cazzari

Sr Manager, IoT Product Innovation & Engineering

Daniele brings over 10 years of experience in IoT architecture. He currently serves as a technical lead for cloud-native SAS IoT solutions and products. His primary objective is to streamline the processing and analysis of data derived from edge devices. Daniele has driven several key initiatives including ESP-ONNX integration, computer vision edge architecture and cloud marketplaces deployment.

Copyright © SAS Institute Inc. All rights reserved.

Rik de Ruiter

Senior Systems Architect

Rik is a technical product manager in the SAS IoT division, where he is dedicated to improving real-time inferencing and open source integration for IoT applications. His focus is on making it easier to solve real business cases using AI.

Copyright © SAS Institute Inc. All rights reserved.

ASK THE EXPERT

How Do I Streamline AI Project Workflows?

Daniele Cazzari
Sr Manager, IoT Product Innovation & Engineering

Rik de Ruiter
Sr Systems Architect

Company Confidential – For Internal Use Only
Copyright © SAS Institute Inc. All rights reserved.

Agenda

Introduction to ONNX and ESP

ONNX Integration with ESP

Demo:
ONNX Object Detection

Azure Marketplace deployment

Q&A

Introduction to ONNX and ESP

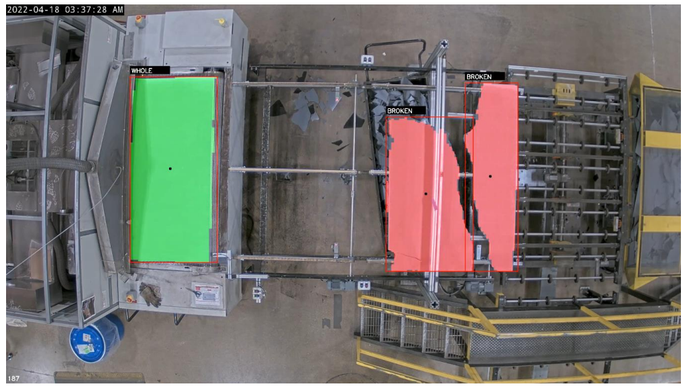
What do these technologies help me with?

Company Confidential – For Internal Use Only
Copyright © SAS Institute Inc. All rights reserved.

Real-world Computer Vision use cases

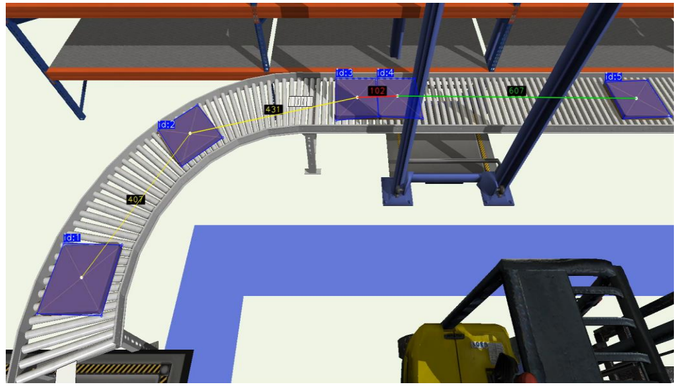
Applicable to many industries

Product Quality

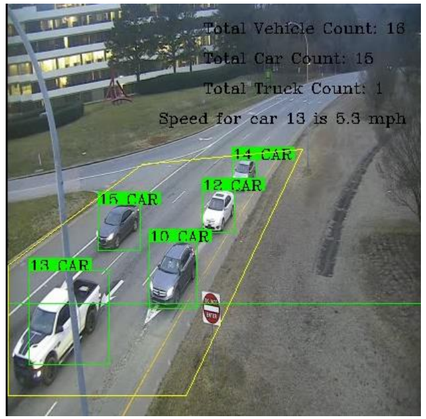


Safety in Workspace

Conveyor belt monitoring

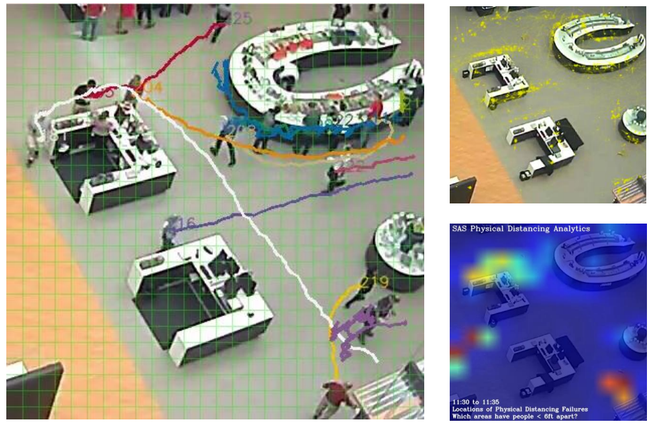


Traffic & Vehicle Monitoring



Traffic Light Monitoring

In-Store monitoring



Company Confidential – For Internal Use Only
Copyright © SAS Institute Inc. All rights reserved.

What is ONNX?

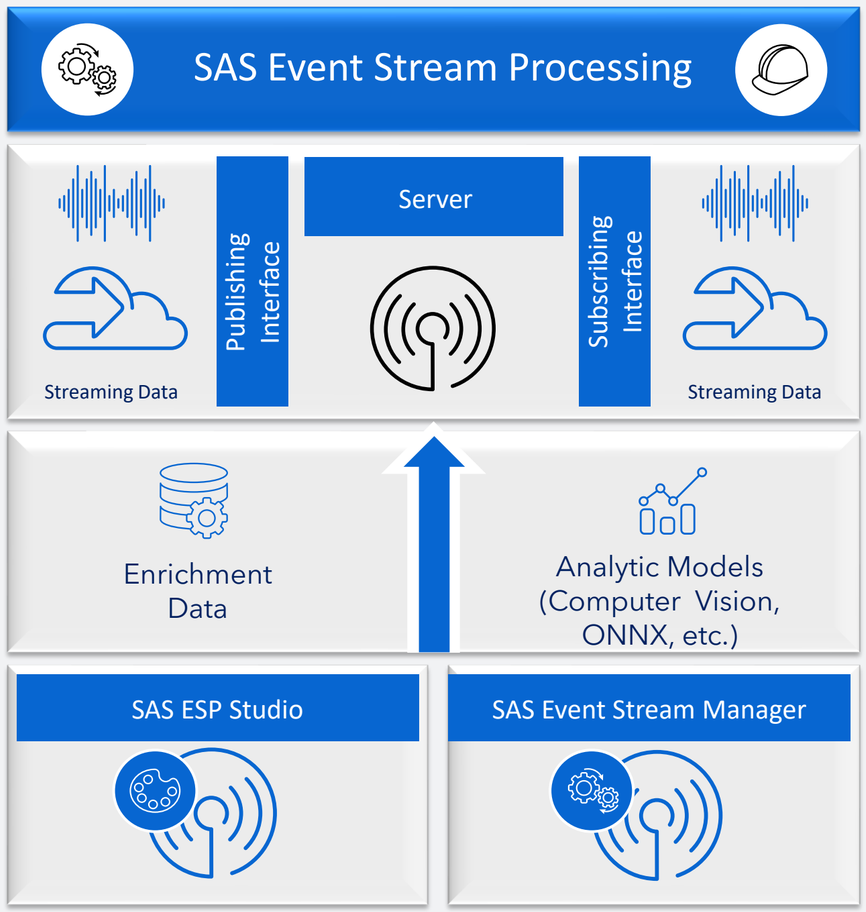
- The **Open Neural Network Exchange (ONNX)** is open standards for representing machine learning algorithms to promote innovation and collaboration in the AI sector.

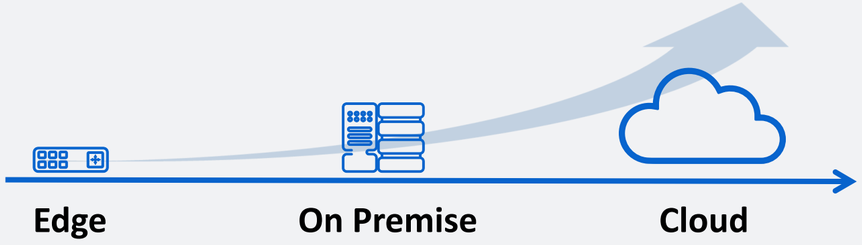
- Framework interoperability
  - Allow developers to more easily move between frameworks, some of which may be more desirable for specific phases of the development process, such as fast training, network architecture flexibility or inferencing on mobile devices.

- Shared optimization
  - *Allow hardware vendors and others to improve the performance of artificial neural networks of multiple frameworks at once by targeting the ONNX representation*

Company Confidential – For Internal Use Only
Copyright © SAS Institute Inc. All rights reserved.

Source: https://onnx.ai





SAS Event Stream Processing

Empowering Computer Vision

- ✓ **Fast streaming data analysis:** Processes structured or unstructured data continuously, on the move, in-memory with very **high speed** and **low latency.**
- ✓ **End-to-end, streamlined, low-code solution** for **real-time** Computer Vison inferencing. Enable end-to-end use case from video ingestion to image annotation into a **single, cohesive solution.**
- ✓ **Optimized for deep leaning inferencing:** Provide **build-in capabilities** to inference computer vision models, leveraging native ONNX runtime integration and **hardware acceleration** frameworks like **Intel Openvino**.
- ✓ **Operational Efficiency:** Enhances the ability to monitor, analyze, and optimize operations in real time, leading to **significant improvements in productivity** and cost savings.

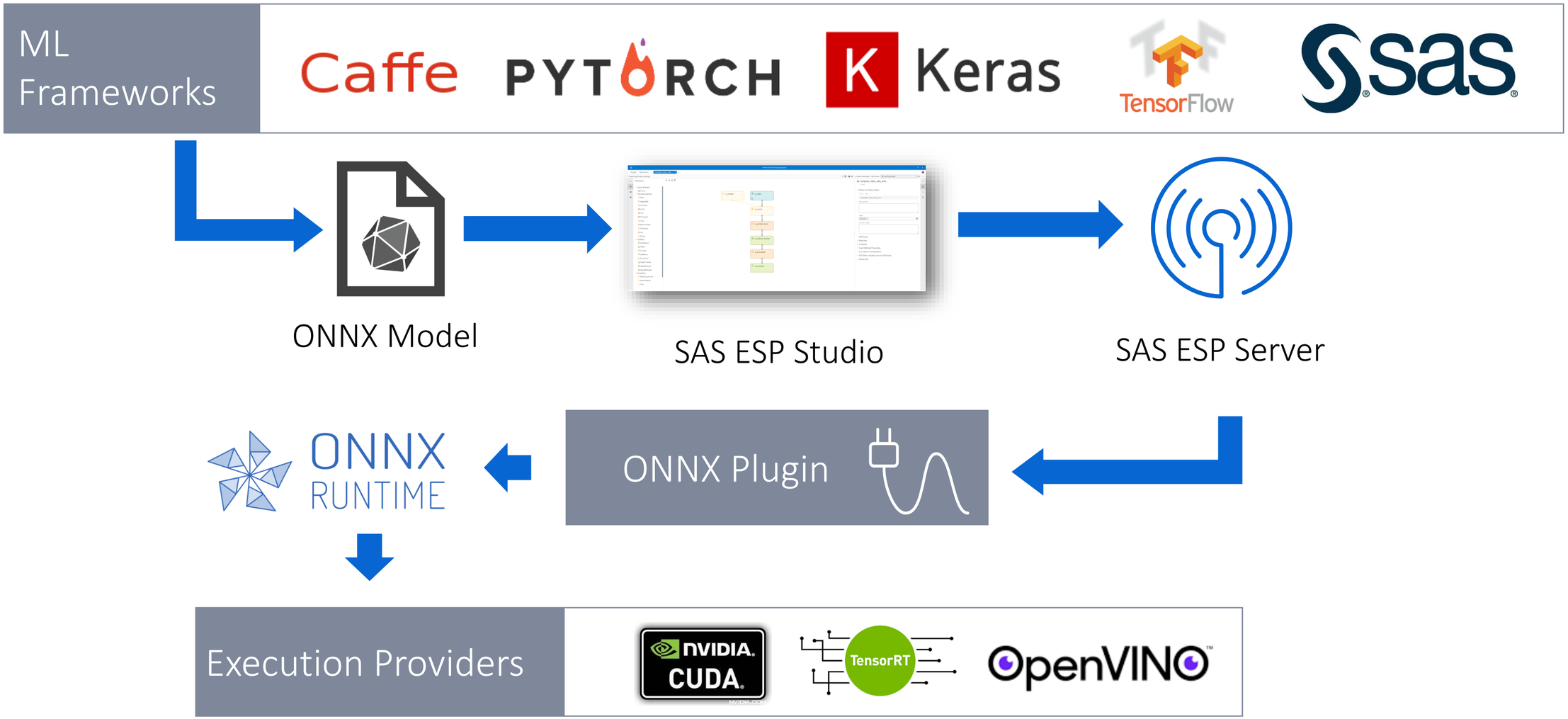
ONNX Integration with ESP

How does it work?

Company Confidential – For Internal Use Only
Copyright © SAS Institute Inc. All rights reserved.

ESP and ONNX

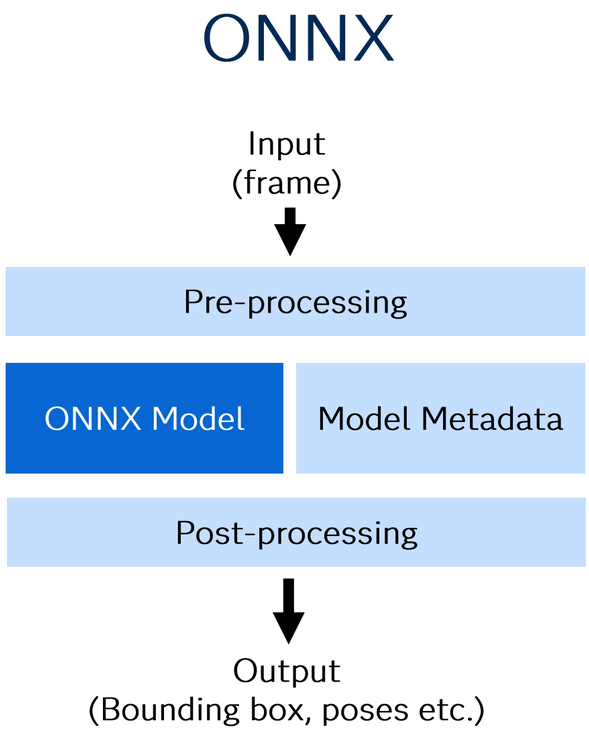
Components



Company Confidential – For Internal Use Only
Copyright © SAS Institute Inc. All rights reserved.

ONNX scoring pipeline

ONNX model is only a portion of the overall flow

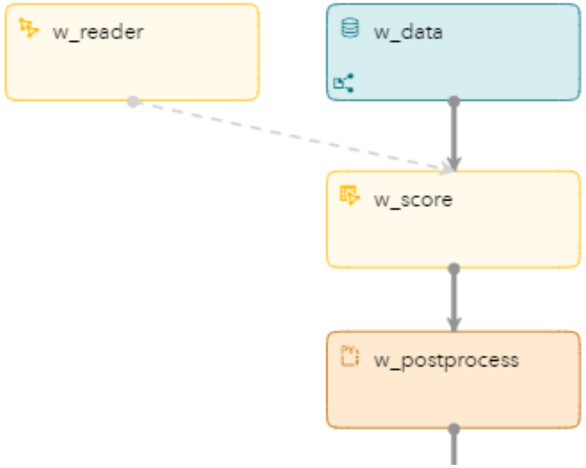


- ONNX file is only the model and does not include pre-processing, post-processing or metadata
- Examples of preprocessing steps are resizing, BGR2RGB conversion, normalization, NCHW encoding
- Postprocessing includes decoding the output tensor of the ONNX model, and sometimes additional steps such as non-max suppression (NMS)
- Model metadata often includes a labels.txt file to map IDs back to class names later

Company Confidential – For Internal Use Only
Copyright © SAS Institute Inc. All rights reserved.

ESP ONNX Pipeline

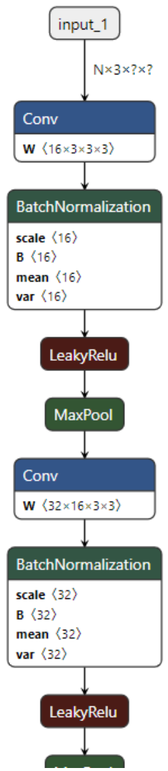
Low-code pipeline for real-time computer vision



- Reading video data via the Video Capture connector **(new!)**
- Preprocessing configuration via the UI **(new!)**
- Scoring the ONNX model via the Score window
- Postprocessing handled via a Python window **(new!)**
- Python window contains utility functions to make it easier to work with tensors and images **(new!)**

Company Confidential – For Internal Use Only
Copyright © SAS Institute Inc. All rights reserved.

ONNX Scoring: Retrieve Model Tensor Input and Output



- Each model could one or more tensors as input or output
- A tool like Netron provides graphical information about input and output formats (https://netron.app/)
- ESP provides functionality to retrieve this information via Studio or command line
- Each input/output is a tensor that need to be encoded/decoded during the inferencing process

Company Confidential – For Internal Use Only
Copyright © SAS Institute Inc. All rights reserved.

UI-based pre-processing

Steps:

resize — Advanced properties

Property	Value
resizeType	letterbox
width	@MODEL_WIDTH@
height	@MODEL_HEIGHT@

color

Property	Value
function*	BGR2RGB

normalize

Property	Value
type*	zero_one

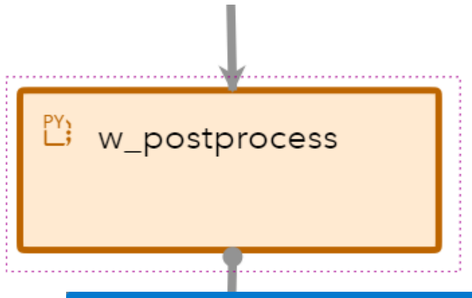
encode

Property	Value
shape*	NCHW

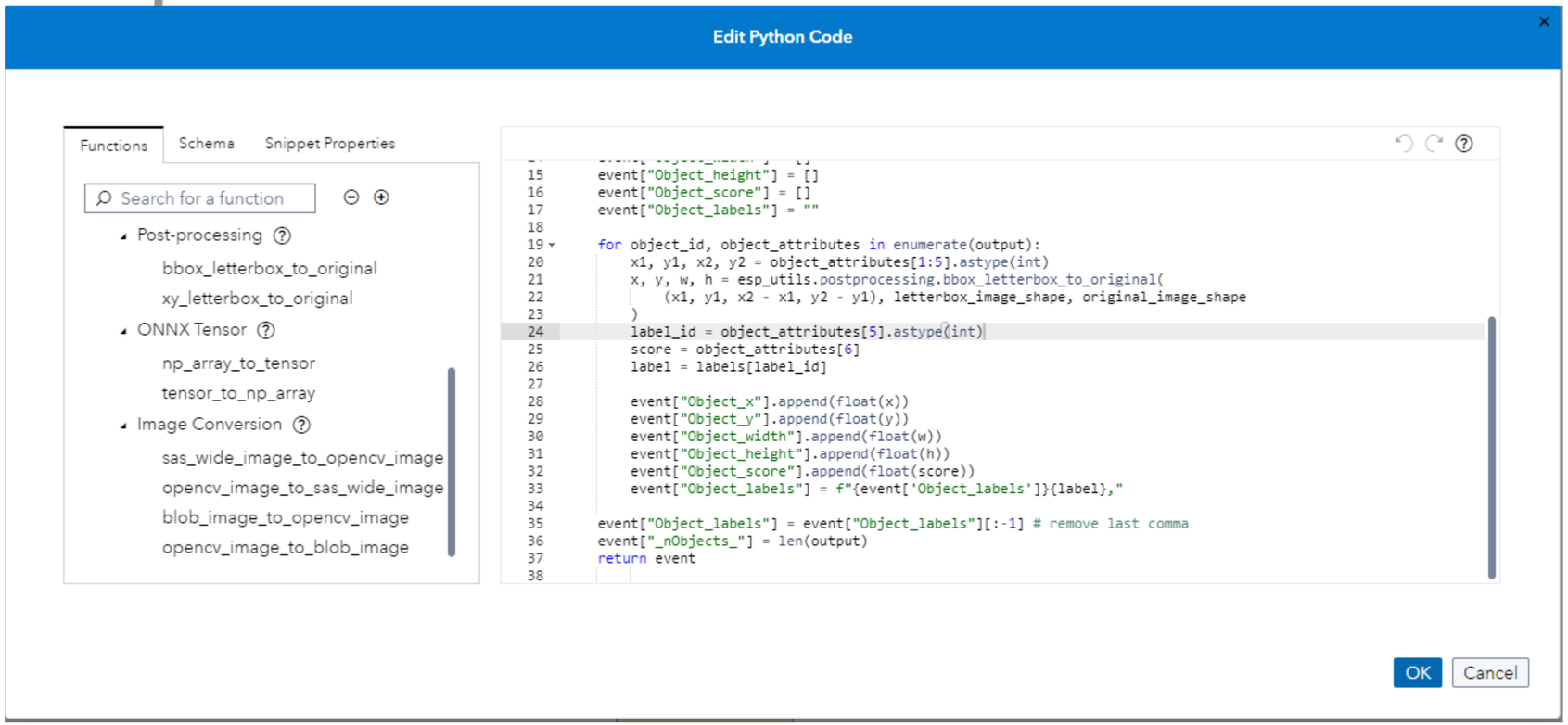
- UI based preprocessing:
  - Resize: basic, letterbox, ratio resize
  - Normalization: 0-1, mean vector, standard deviation vect.
  - Color Encoding: BGR, RGB
  - Other transformations: Image Encoding (NCHW), Padding/ Stride, Expand Dimension

Company Confidential – For Internal Use Only
Copyright © SAS Institute Inc. All rights reserved.

Post-processing



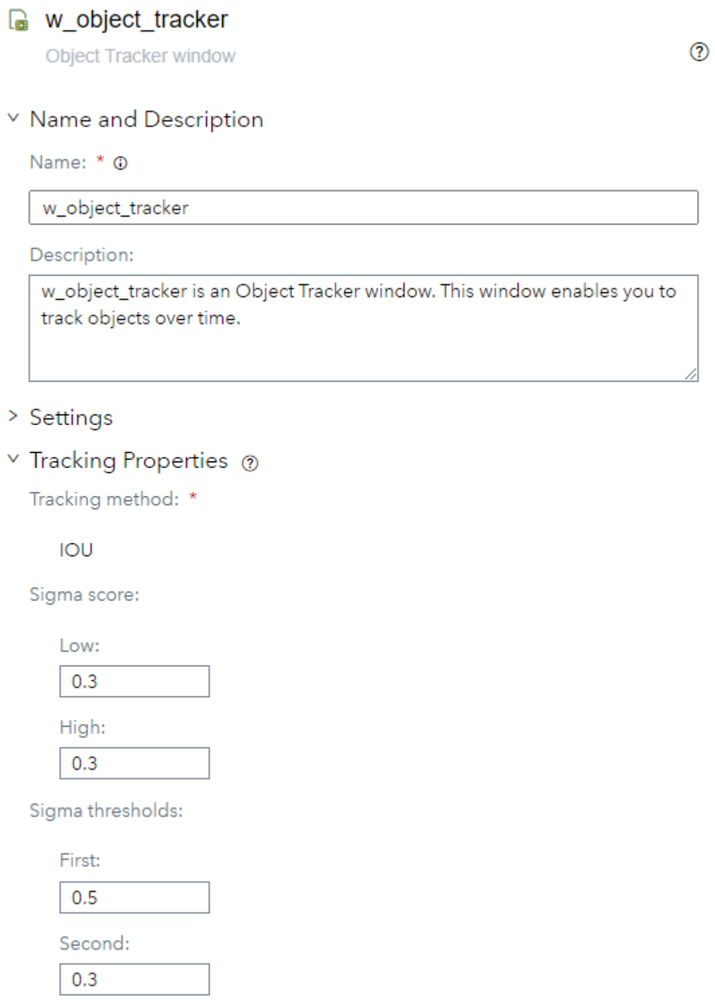
- Process the ONNX output tensors into ESP variables using a Python window
- Utility functions are available



Company Confidential – For Internal Use Only
Copyright © SAS Institute Inc. All rights reserved.

What's next?

- Further processing could include:
  - Object tracking
  - Geofencing
  - Annotation
  - Decision logic
  - Notifications

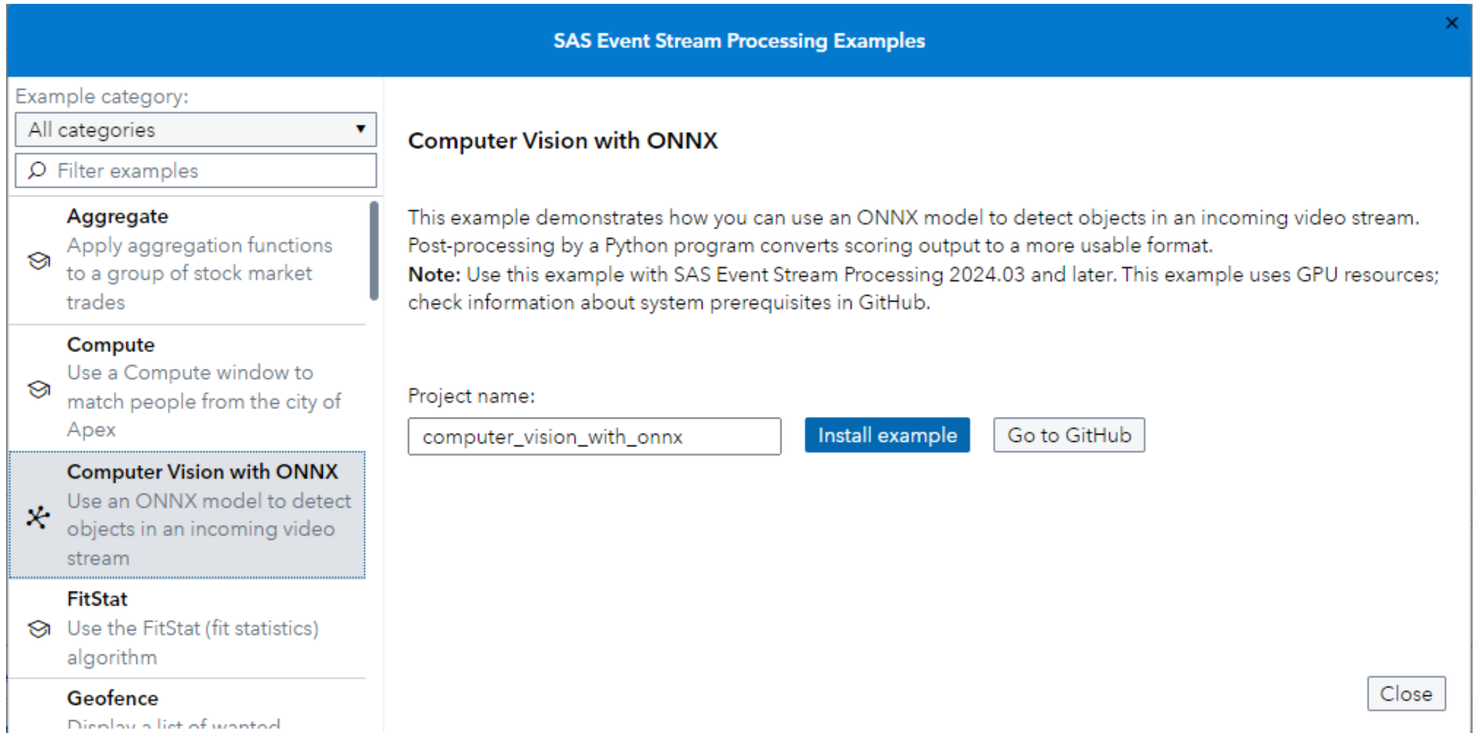


Company Confidential – For Internal Use Only
Copyright © SAS Institute Inc. All rights reserved.

Installable example for Computer Vision

Runs out of the box, comes with all scripts and files to run

- Object detection example



Company Confidential – For Internal Use Only
Copyright © SAS Institute Inc. All rights reserved.

Demo ONNX Object Detection

Company Confidential – For Internal Use Only
Copyright © SAS Institute Inc. All rights reserved.

Azure Marketplace

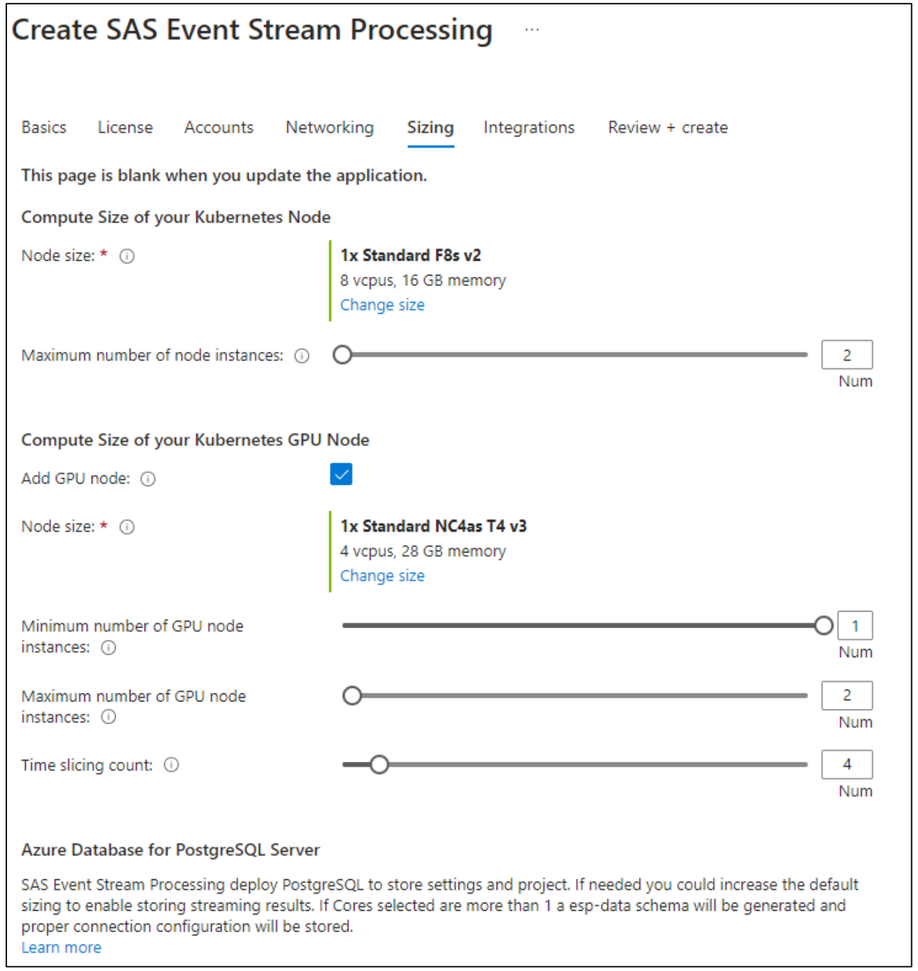
Enablement and deployment

Company Confidential – For Internal Use Only
Copyright © SAS Institute Inc. All rights reserved.

Event Stream Processing

Azure Marketplace Deployment

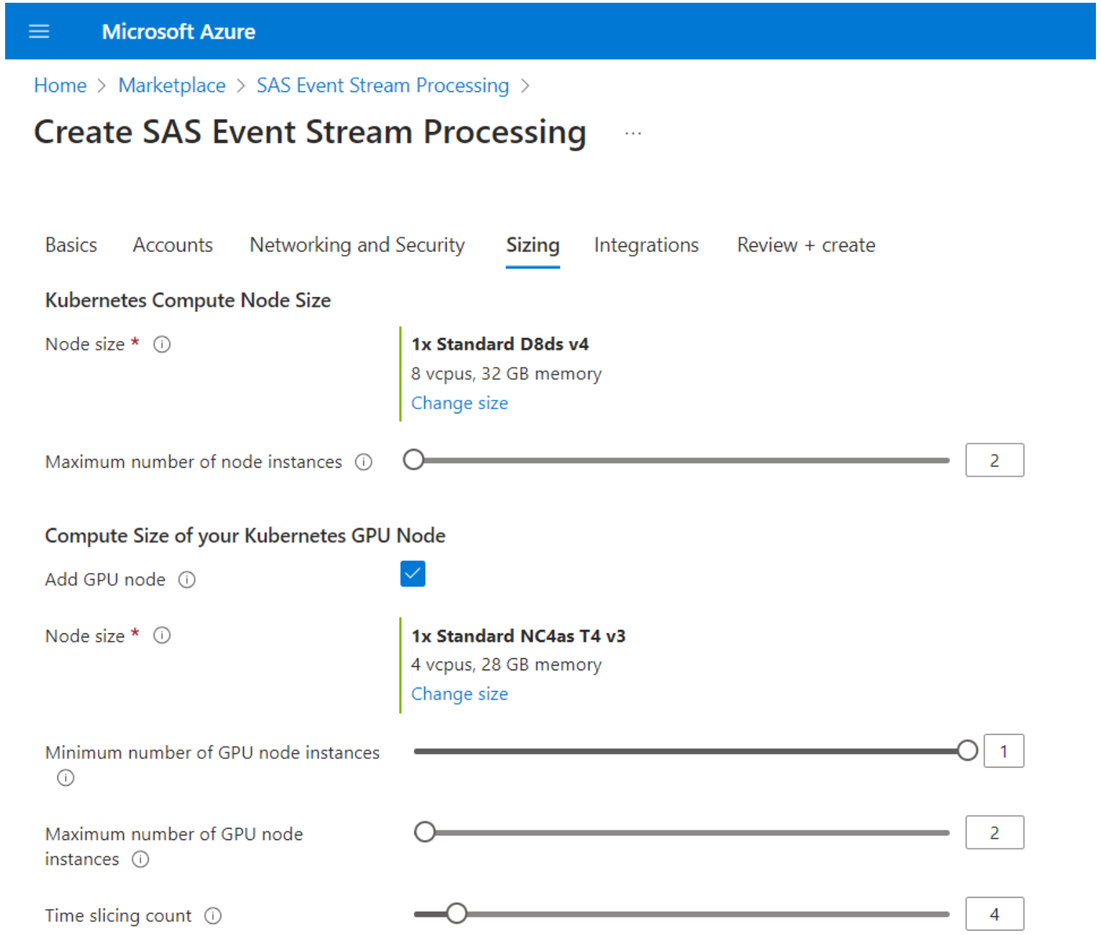
- Simple and intuitive provisioning of a cost-effective Kubernetes infrastructure that fully support SAS Event Stream Processing enabled with Git and Grafana
- Support for scalable CPU and GPU workload tailored for you application needs
- Automatic configuration of managed Postgres DB that could be also used to store streaming analysis outcome
- Optional integration with EventHub for data ingestion and external blob storage



Company Confidential – For Internal Use Only
Copyright © SAS Institute Inc. All rights reserved.

GPU support in Azure Marketplace

- Easy to deploy a GPU, just a checkbox
- Reduce costs by autoscaling to 0 when the GPU is not used
- Multiple projects can share the GPU (time slicing support) to fully utilize a GPU



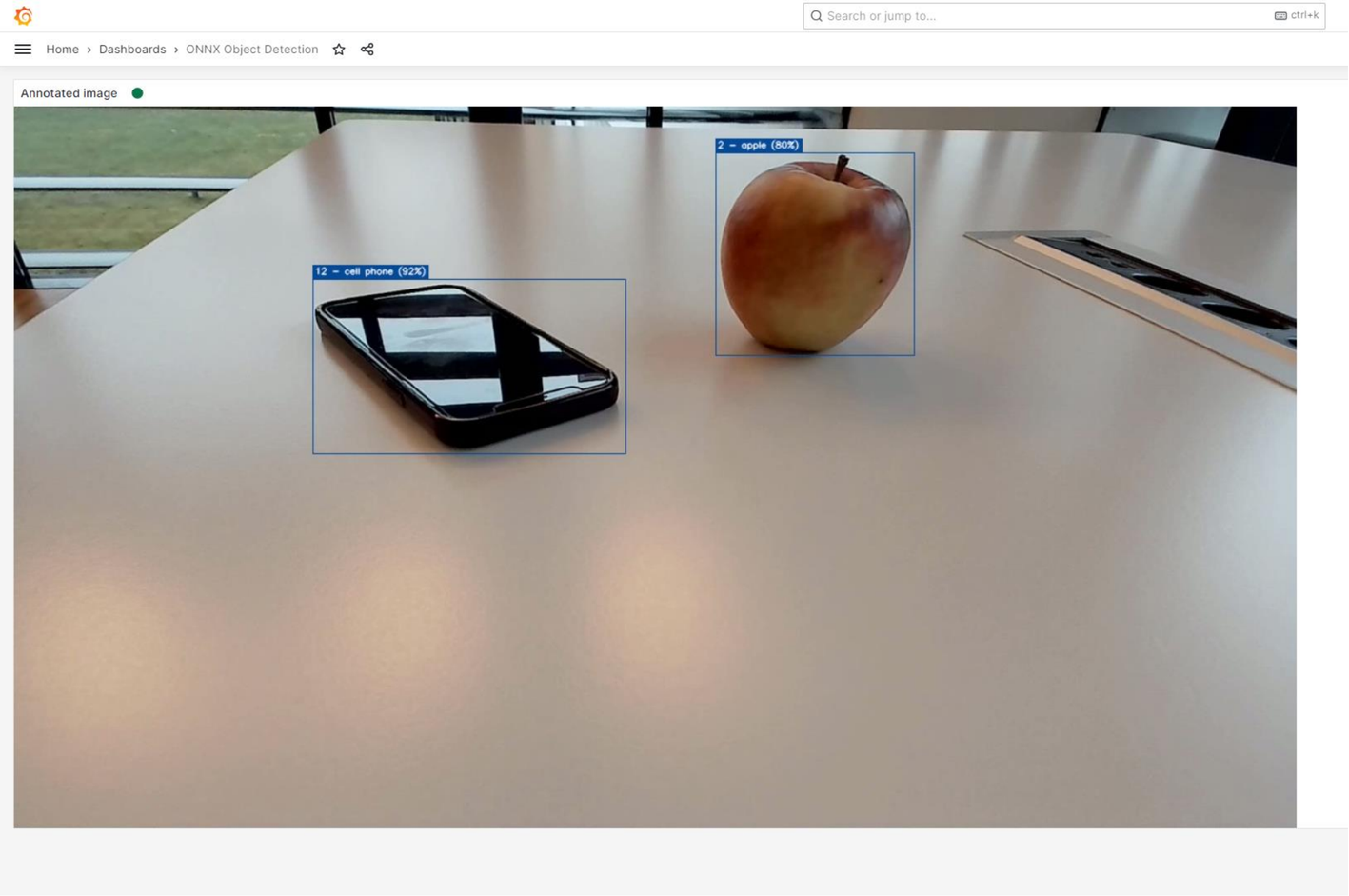
Company Confidential – For Internal Use Only
Copyright © SAS Institute Inc. All rights reserved.

Grafana integration

Deployed automatically with ESP on Azure Marketplace

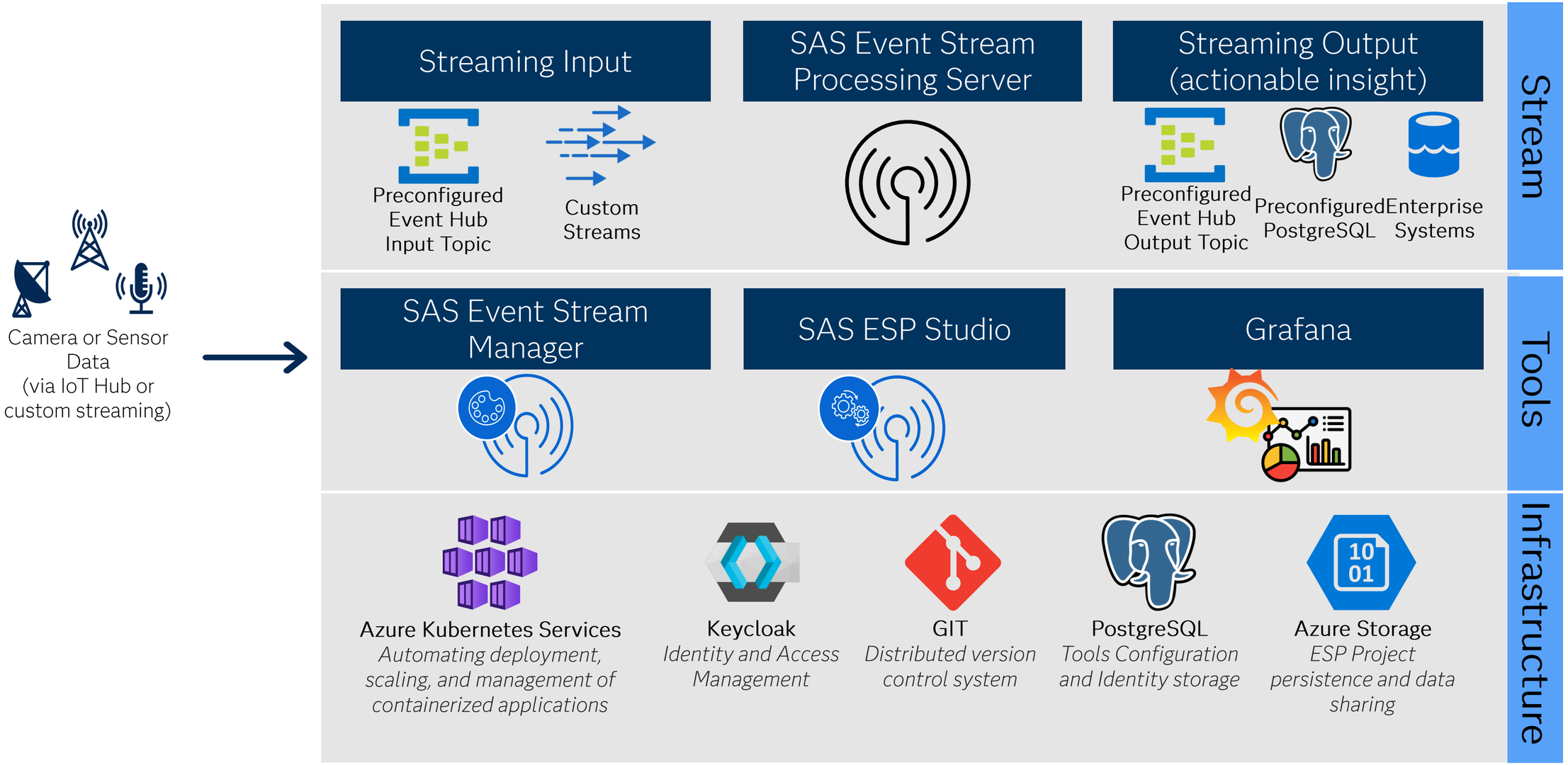
The Grafana ESP plugin…

- Allows to visualize and report on running ESP projects in Grafana
- Discovers ESP server pods in the cluster
- Makes efficient connections to the running projects
- Supports Computer Vision use cases



Company Confidential – For Internal Use Only
Copyright © SAS Institute Inc. All rights reserved.

Azure Marketplace Architecture



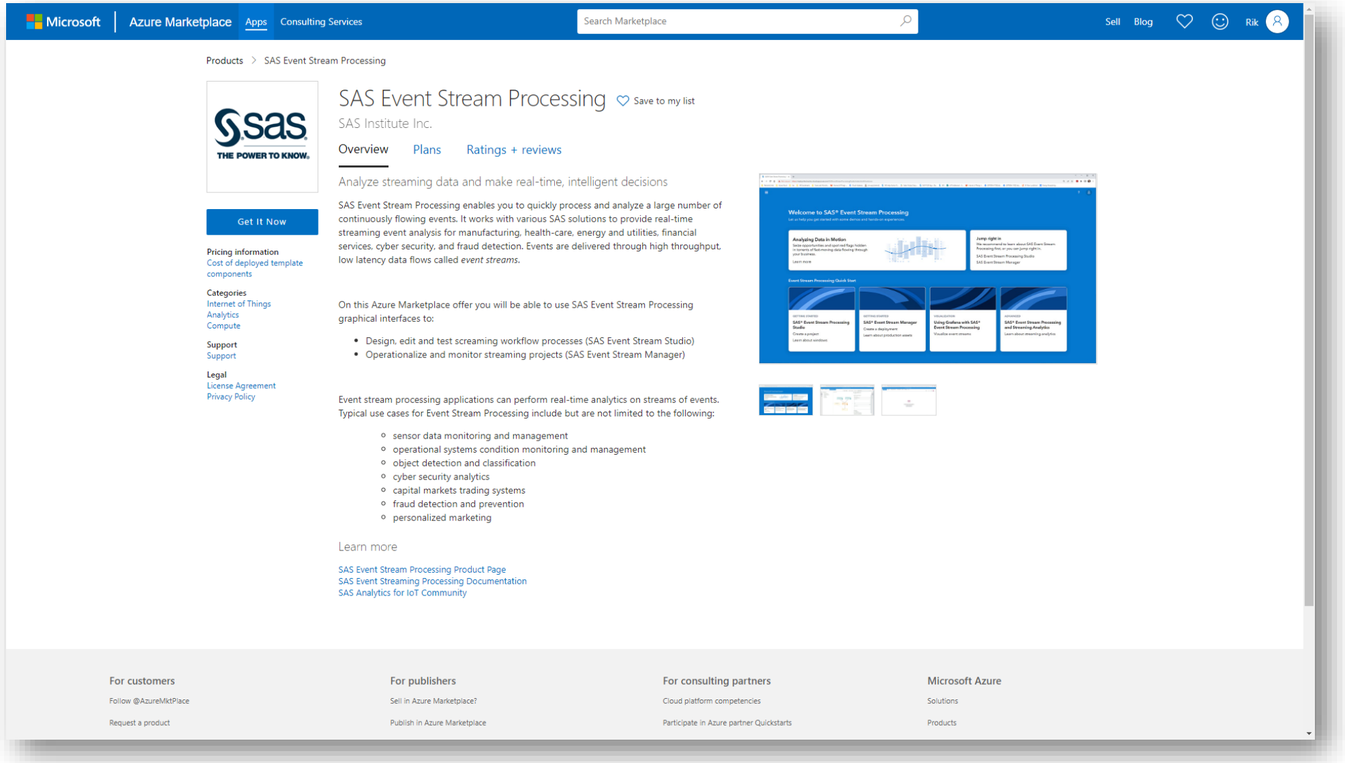
Company Confidential – For Internal Use Only
Copyright © SAS Institute Inc. All rights reserved.

SAS Event Stream Processing

Available on Azure Marketplace

You could deploy it with your existing license now!

For a trial license or for any question please write to iotcontact@sas.com



Company Confidential – For Internal Use Only
Copyright © SAS Institute Inc. All rights reserved.

References

Additional Materials

Company Confidential – For Internal Use Only
Copyright © SAS Institute Inc. All rights reserved.

Enablement & Support

ONNX example:

https://github.com/sassoftware/esp-studio-examples/tree/main/Advanced/onnx_object_detection

ONNX Model Zoo:

https://github.com/onnx/models

ESP Documentation:

https://go.documentation.sas.com/doc/en/espcdc/v_047/espan/p0b1zsgwrsirbln1typkfoz428y9.htm

ESP on Azure Marketplace:

https://azuremarketplace.microsoft.com/en/marketplace/apps/sas-institute-560503.sas_esp?tab=Overview

Microsoft Azure Documentation:

https://learn.microsoft.com/en-us/azure/?product=popular

Company Confidential – For Internal Use Only
Copyright © SAS Institute Inc. All rights reserved.

THANK YOU

For more information, visit: www.sas.com/esp

sas.com

Company Confidential – For Internal Use Only
Copyright © SAS Institute Inc. All rights reserved.