# Polyglot Pipelines with

APACHE nifi

# Bob Paulin

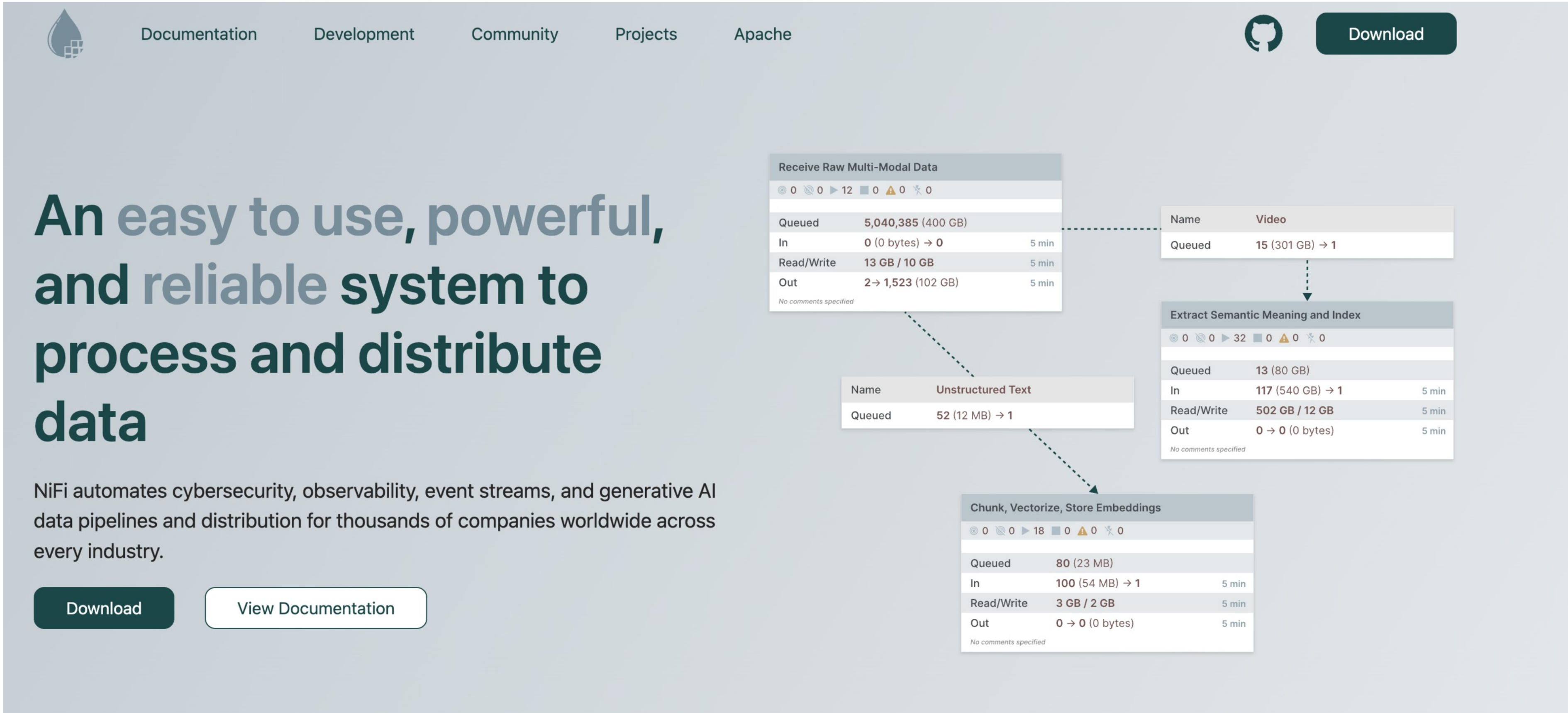
Datavolo – Software Engineer
ASF Member
Java Champion
CJUG Board Chair

Podcasts

Java Off Heap - https://www.javaoffheap.com/

Java Pub House - https://www.javapubhouse.com/

# Apache NiFi



## An easy to use, powerful, and reliable system to process and distribute data

NiFi automates cybersecurity, observability, event streams, and generative AI data pipelines and distribution for thousands of companies worldwide across every industry.

**Receive Raw Multi-Modal Data**

◎ 0 ⊘ 0 ▶ 12 ■ 0 ⚠ 0 ✗ 0

| | | |
|---|---|---|
| Queued | 5,040,385 (400 GB) | |
| In | 0 (0 bytes) → 0 | 5 min |
| Read/Write | 13 GB / 10 GB | 5 min |
| Out | 2→ 1,523 (102 GB) | 5 min |

*No comments specified*

| Name | Video |
|---|---|
| Queued | 15 (301 GB) → 1 |

**Extract Semantic Meaning and Index**

◎ 0 ⊘ 0 ▶ 32 ■ 0 ⚠ 0 ✗ 0

| | | |
|---|---|---|
| Queued | 13 (80 GB) | |
| In | 117 (540 GB) → 1 | 5 min |
| Read/Write | 502 GB / 12 GB | 5 min |
| Out | 0 → 0 (0 bytes) | 5 min |

*No comments specified*

| Name | Unstructured Text |
|---|---|
| Queued | 52 (12 MB) → 1 |

**Chunk, Vectorize, Store Embeddings**

◎ 0 ⊘ 0 ▶ 18 ■ 0 ⚠ 0 ✗ 0

| | | |
|---|---|---|
| Queued | 80 (23 MB) | |
| In | 100 (54 MB) → 1 | 5 min |
| Read/Write | 3 GB / 2 GB | 5 min |
| Out | 0 → 0 (0 bytes) | 5 min |

*No comments specified*

# Apache NiFi

## Building a Simple OpenSearch Index

https://data.cityofchicago.org/Administration-Finance/Current-Employee-Names-Salaries-and-Position-Title/xzkq-xp2w/about_data

https://data.cityofchicago.org/resource/xzkq-xp2w.csv

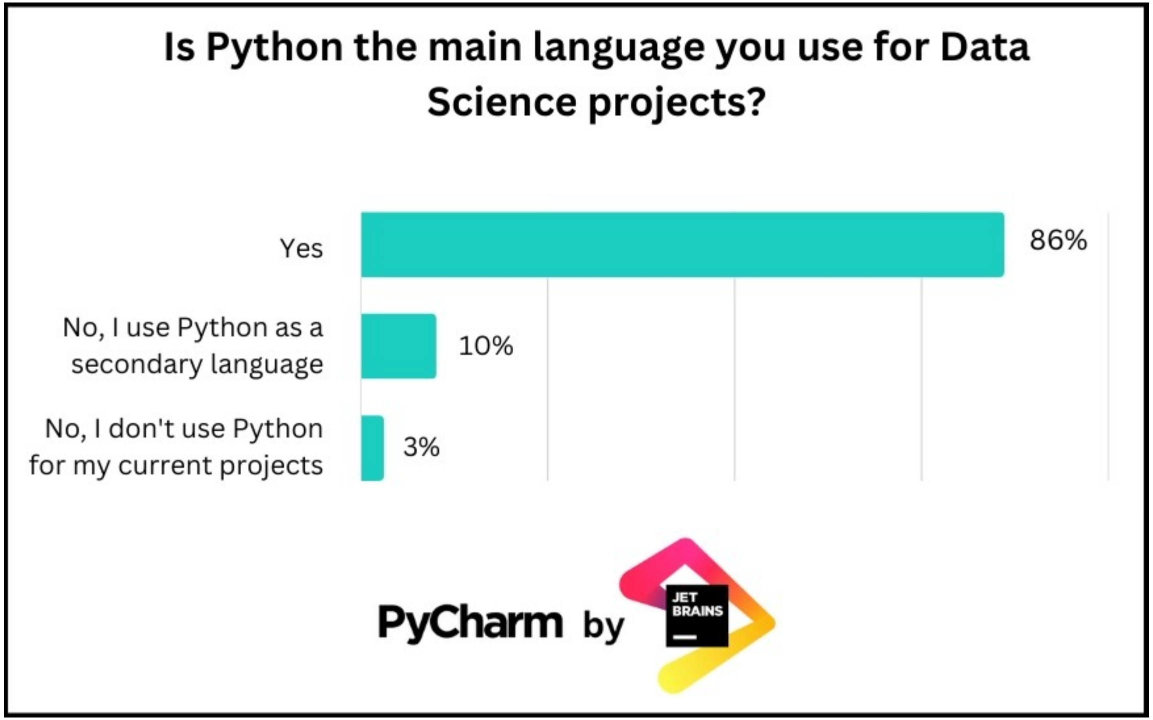
# Apache NiFi Loves Java

OpenSearch
APACHE kafka®
Groovy
Apache™ Tika
AVRO™
APACHE POI
hadoop
ICEBERG
Parquet

# Is there anything that I might not prefer to write in Java?

# AI

# ML and Data Science Community extensively use Python



https://blog.jetbrains.com/pycharm/2023/10/future-of-data-science/

# Python ML Libraries

NumPy

PyTorch

pandas

TensorFlow

ONNX

OpenCV

# Languages and Libraries evolve based on usage

”The next best thing to having good ideas is recognizing good ideas from your users. Sometimes the latter is better.” - Eric Raymond

See also

Growing a Language, by Guy Steele

https://youtu.be/_ahvzDzKdB0

How do I enable my developers to use the best tool for the job?

# Apache NiFi 2.x Py4J Plugin

# How Does Py4J Work in NiFi?

- Separate Python Process with Socket Connection

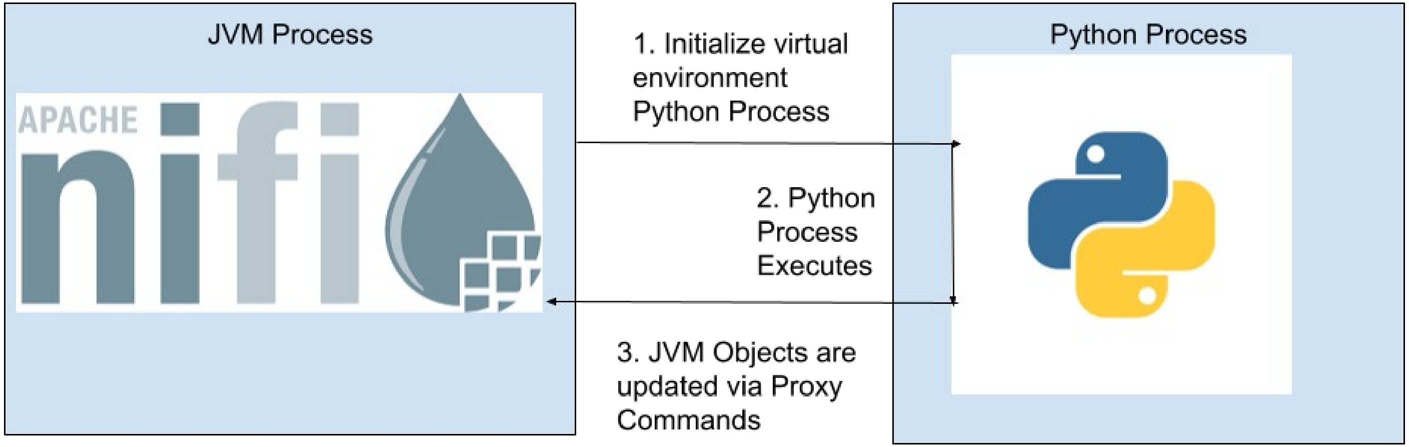
# How Does Py4J Work in NiFi?

- Separate Python Process with Socket Connection
- Python Virtual Environment for dependency isolation

# How Does Py4J Work in NiFi?

- Separate Python Process with Socket Connection
- Python Virtual Environment for dependency isolation
- Object proxies
  - Updates on the Python side mutates the objects on on the JVM

# How Does Py4J Work in NiFi?



JVM Process

Python Process

1. Initialize virtual environment Python Process

2. Python Process Executes

3. JVM Objects are updated via Proxy Commands

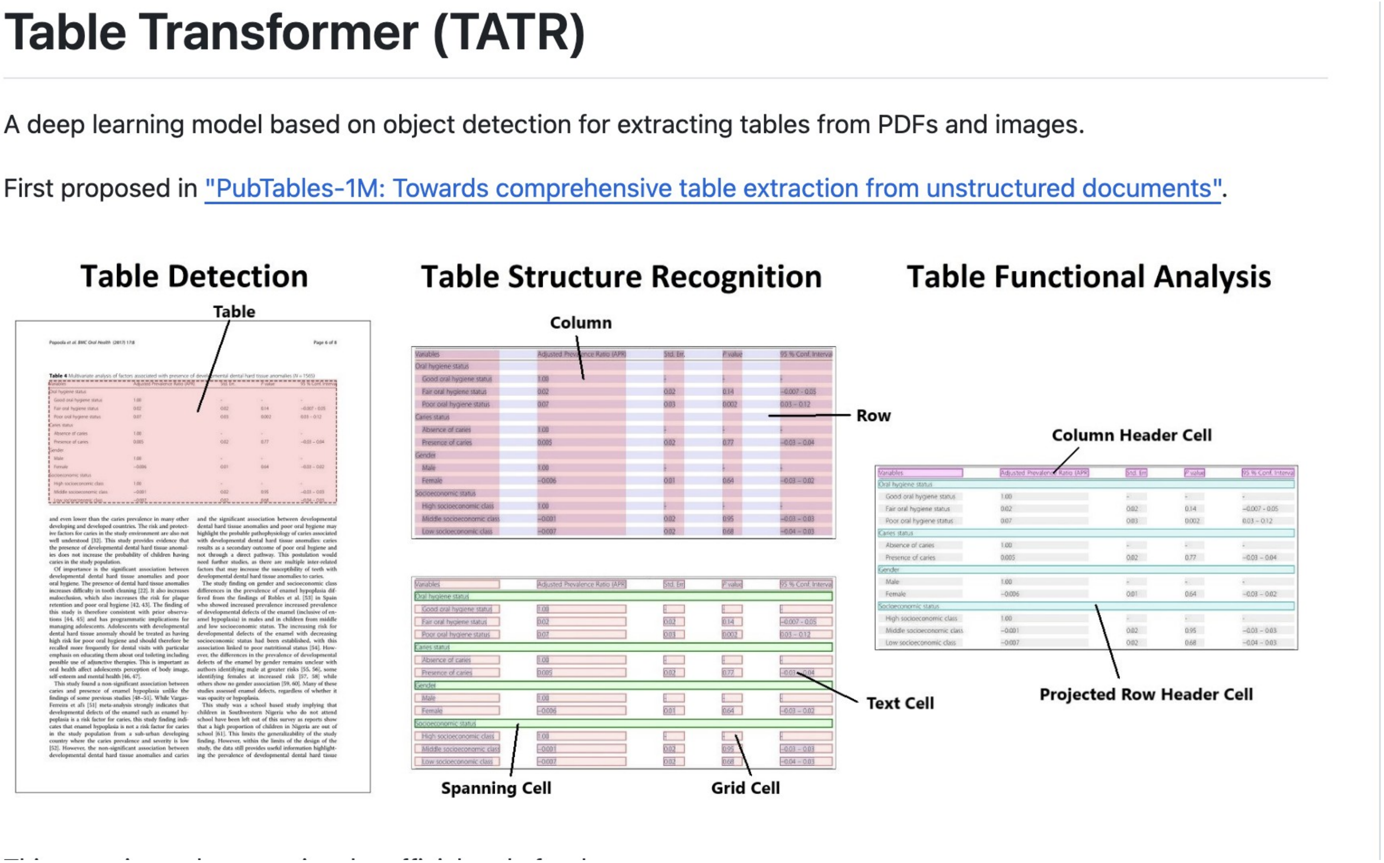
# Integrating Python Libraries

- Processor Member Field
- requirements.txt
- Bundled in NAR file

# What can I do in NiFi with Python?

- Transform Flow Files
- Create Flow Files
- Transform Records

# Lets Build a NiFi Python Plugin



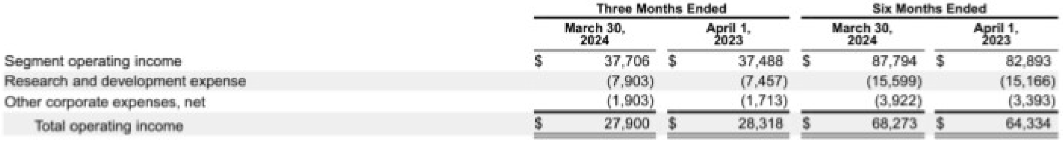
## Table Transformer (TATR)

A deep learning model based on object detection for extracting tables from PDFs and images.

First proposed in "PubTables-1M: Towards comprehensive table extraction from unstructured documents".

https://github.com/microsoft/table-transformer

# Lets Build a NiFi Python Plugin



| | Three Months Ended | | Six Months Ended | |
|---|---|---|---|---|
| | March 30, 2024 | April 1, 2023 | March 30, 2024 | April 1, 2023 |
| Segment operating income | $ 37,706 | $ 37,488 | $ 87,794 | $ 82,893 |
| Research and development expense | (7,903) | (7,457) | (15,599) | (15,166) |
| Other corporate expenses, net | (1,903) | (1,713) | (3,922) | (3,393) |
| Total operating income | $ 27,900 | $ 28,318 | $ 68,273 | $ 64,334 |

Encode

```
{'pixel_values': tensor([[[[2.2489, 2.2489, 2.2489,  ..., 2.2489, 2.2489, 2.2489]]]]), 'pixel_mask': tensor([[[1, 1,
1,  ..., 1, 1, 1],])}
```

Model
Inference

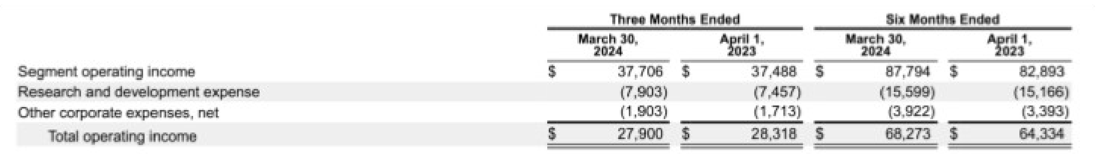
```
TableTransformerObjectDetectionOutput(loss=None, loss_dict=None, logits=tensor([[[-13.3932,  -7.7760,
-7.5944, -10.9697, -11.2009,  -7.4343,   4.2611],...]]], grad_fn=<ViewBackward0>),
pred_boxes=tensor([[[0.5433, 0.6224, 0.3335, 0.0981], [0.4947, 0.2442, 0.9166, 0.0779], ]]],
grad_fn=<SigmoidBackward0>), ……)
```

Post
Process

```
{'scores': tensor([1.0000, 0.9999, 0.9995, 0.9997, 0.9999, 0.9996, 0.9999, 0.9994, 0.9997,
        0.9995, 0.9999, 0.9998, 0.9999, 0.9999, 0.9997],  grad_fn=<IndexBackward0>),
'labels': tensor([1, 1, 3, 5, 1, 2, 1, 2, 2, 5, 0, 2, 1, 2, 2]), 'boxes': tensor([[1740.9777,
126.5295, 2830.4287,  659.9639], ...]]grad_fn=<IndexBackward0>)}
```

# Building a Nifi Python Plugin

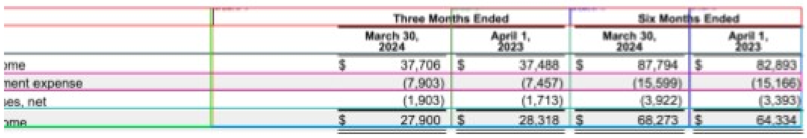
## Type of Processor



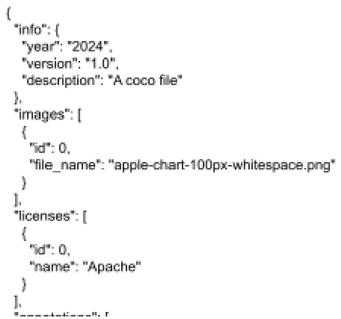
| | Three Months Ended | | Six Months Ended | |
|---|---|---|---|---|
| | March 30, 2024 | April 1, 2023 | March 30, 2024 | April 1, 2023 |
| Segment operating income | $ 37,706 | $ 37,488 | $ 87,794 | $ 82,893 |
| Research and development expense | (7,903) | (7,457) | (15,599) | (15,166) |
| Other corporate expenses, net | (1,903) | (1,713) | (3,922) | (3,393) |
| Total operating income | $ 27,900 | $ 28,318 | $ 68,273 | $ 64,334 |

annotated

coco



| | Three Months Ended | | Six Months Ended | |
|---|---|---|---|---|
| | March 30, 2024 | April 1, 2023 | March 30, 2024 | April 1, 2023 |
| ome | $ 37,706 | $ 37,488 | $ 87,794 | $ 82,893 |
| ment expense | (7,903) | (7,457) | (15,599) | (15,166) |
| ses, net | (1,903) | (1,713) | (3,922) | (3,393) |
| ome | $ 27,900 | $ 28,318 | $ 68,273 | $ 64,334 |



```
{
  "info": {
    "year": "2024",
    "version": "1.0",
    "description": "A coco file"
  },
  "images": [
    {
      "id": 0,
      "file_name": "apple-chart-100px-whitespace.png"
    }
  ],
  "licenses": [
    {
      "id": 0,
      "name": "Apache"
    }
  ],
  "annotations": [
```

# Building a Nifi Python Plugin

PyTorch

## Dependencies

NumPy

OpenCV

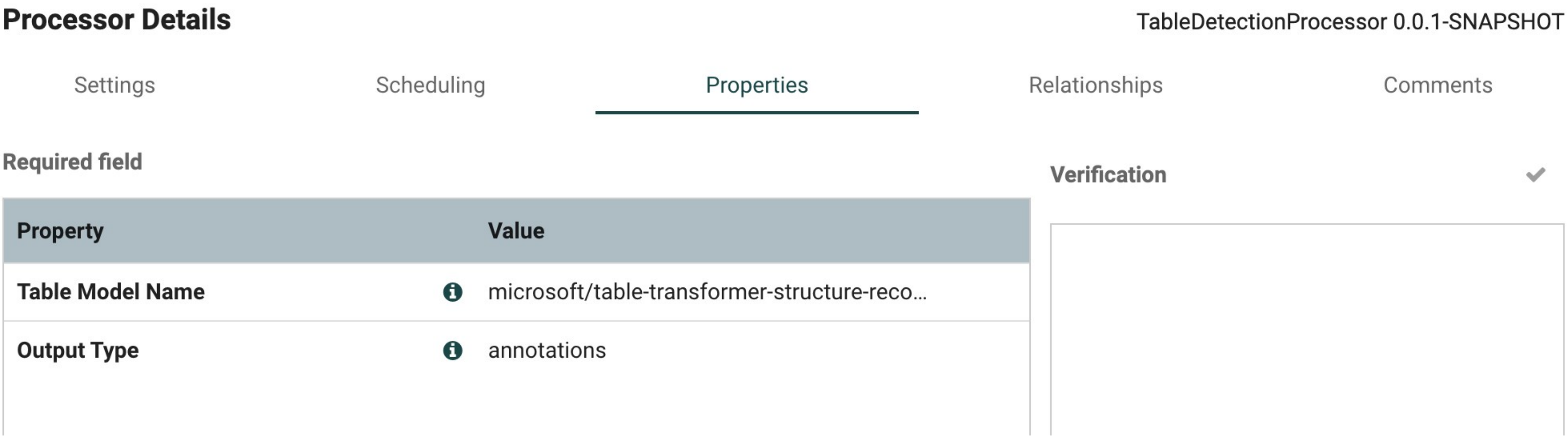
# Hatch Datavolo Nar Plugin

- Allows Python dependencies to be packaged with processor source code.
- Formats project for simple deployment as a NAR in nifi/extensions directory
- Open Source

DATAVOLO

# Building a NiFi Python Plugin

## Processor Properties



**Processor Details**

TableDetectionProcessor 0.0.1-SNAPSHOT

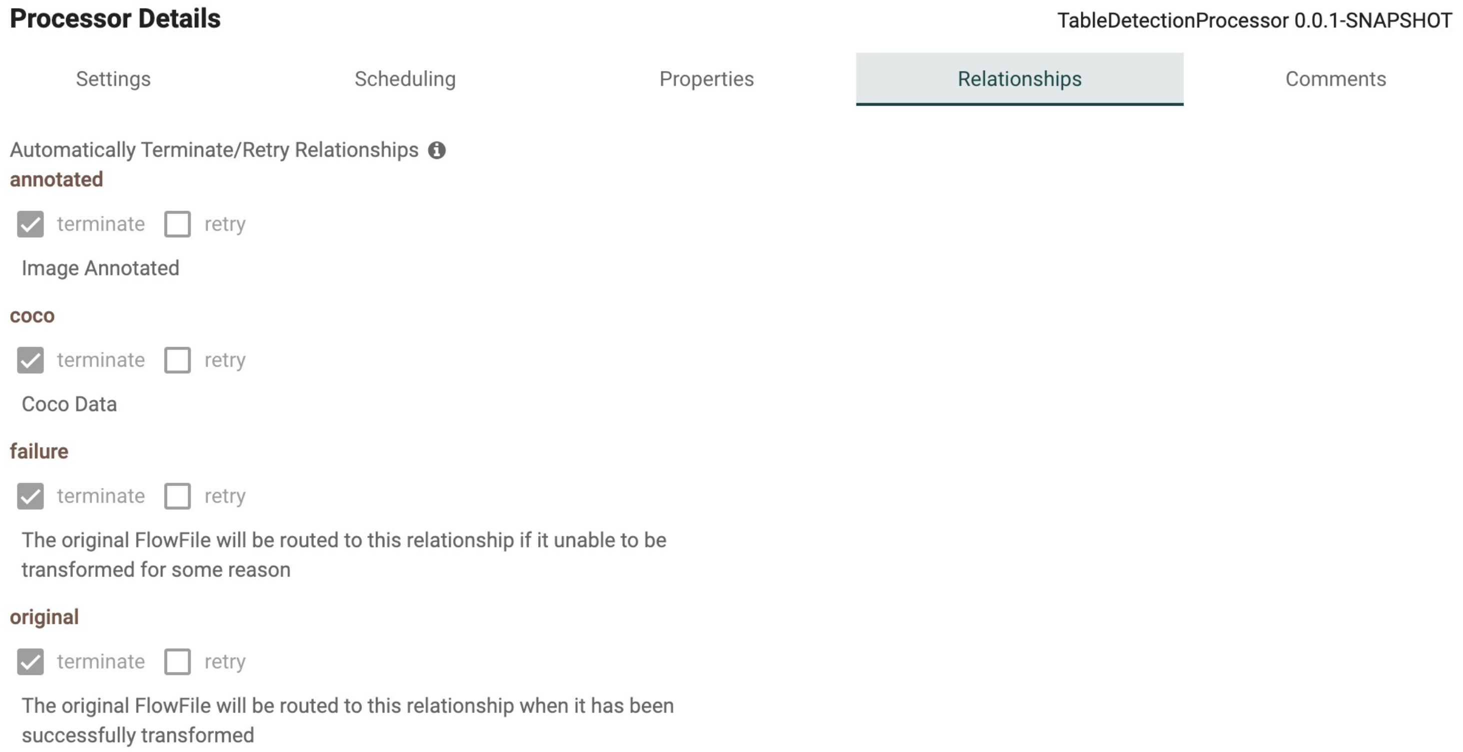
**Required field**

**Verification** ✓

| Property | Value |
|---|---|
| **Table Model Name** ⓘ | microsoft/table-transformer-structure-reco... |
| **Output Type** ⓘ | annotations |

# Building a Nifi Python Plugin

## Relationships



**Processor Details**

TableDetectionProcessor 0.0.1-SNAPSHOT

Automatically Terminate/Retry Relationships

**annotated**

☑ terminate □ retry

Image Annotated

**coco**

☑ terminate □ retry

Coco Data

**failure**

☑ terminate □ retry

The original FlowFile will be routed to this relationship if it unable to be transformed for some reason

**original**

☑ terminate □ retry

The original FlowFile will be routed to this relationship when it has been successfully transformed

# Building a Nifi Python Plugin

## NiFi Lifecycle Behavior

```
def onScheduled(self, context)
def onStopped(self, context)
```

# Building a Nifi Python Plugin

## Triggering Behavior

```
def transform(self, context, flowFile)
```

# Building a NiFi Python Plugin

Ship IT! ✅

# Challenges with Apache NiFi Python in Production

- Security Vulnerabilities

# Challenges with Apache NiFi Python in Production

- Security Vulnerabilities
- GPU Acceleration

# Challenges with Apache NiFi Python in Production

- Security Vulnerabilities
- GPU Acceleration
- Scaling

# Looking to do this in Production?

DATAVOLO

# Thank You!

Bob Paulin
@bobpaulin
https://www.linkedin.com/in/bobpaulin/

DATAVOLO

https://datavolo.io/