

Polyglot Pipelines with





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/>

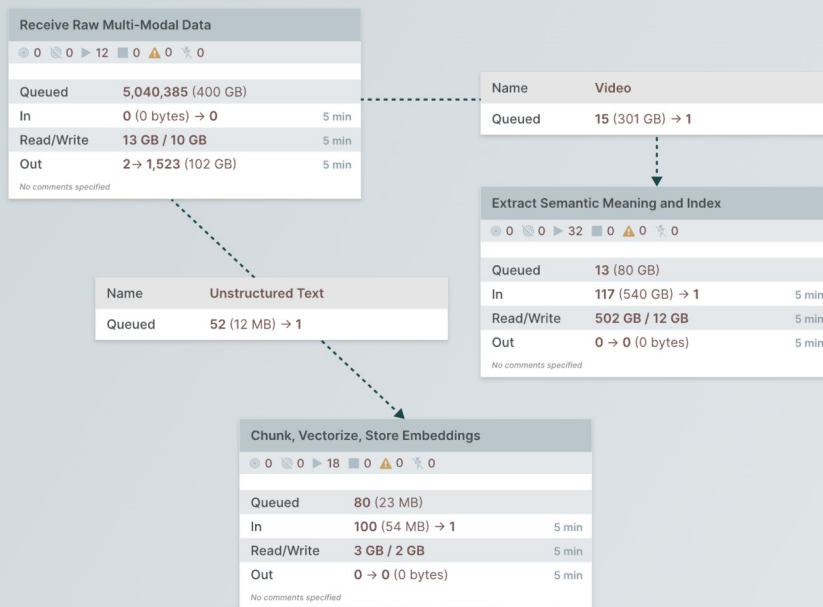


Apache NiFi

[Documentation](#)[Development](#)[Community](#)[Projects](#)[Apache](#)[Download](#)

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.

[Download](#)[View Documentation](#)

Apache NiFi

Building a Simple OpenSearch Index



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

Apache NiFi Loves Java

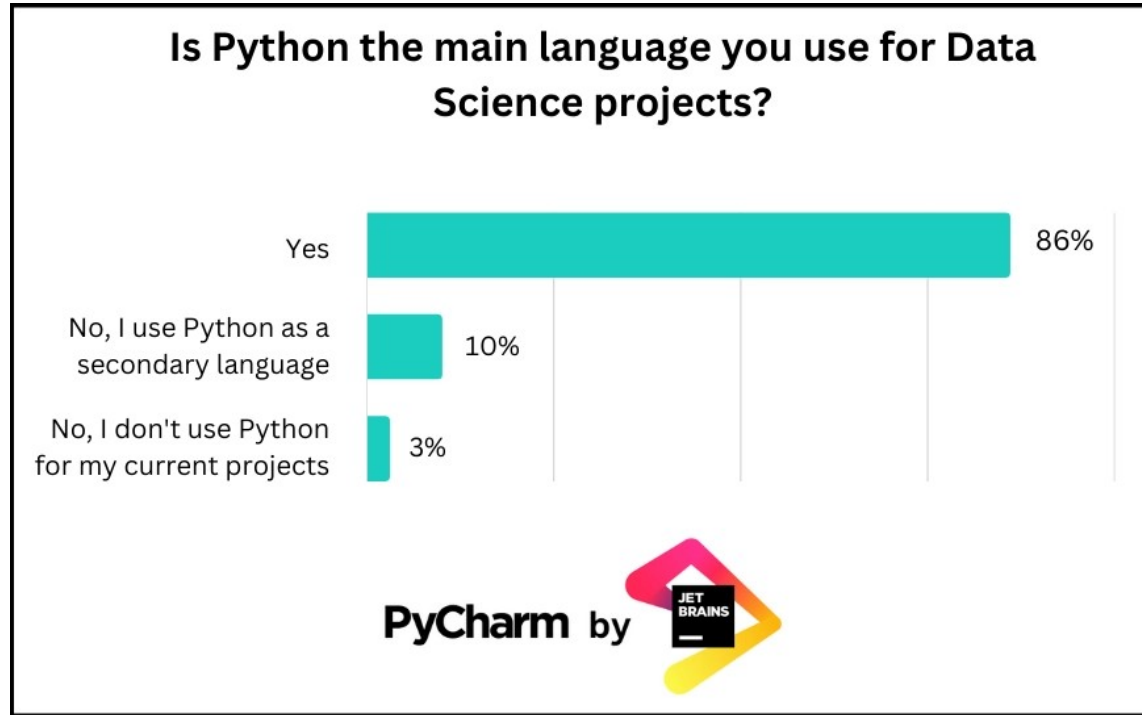


Is there anything that I might not
prefer to write in Java?



AI

ML and Data Science Community extensively use Python



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

Python ML Libraries



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



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

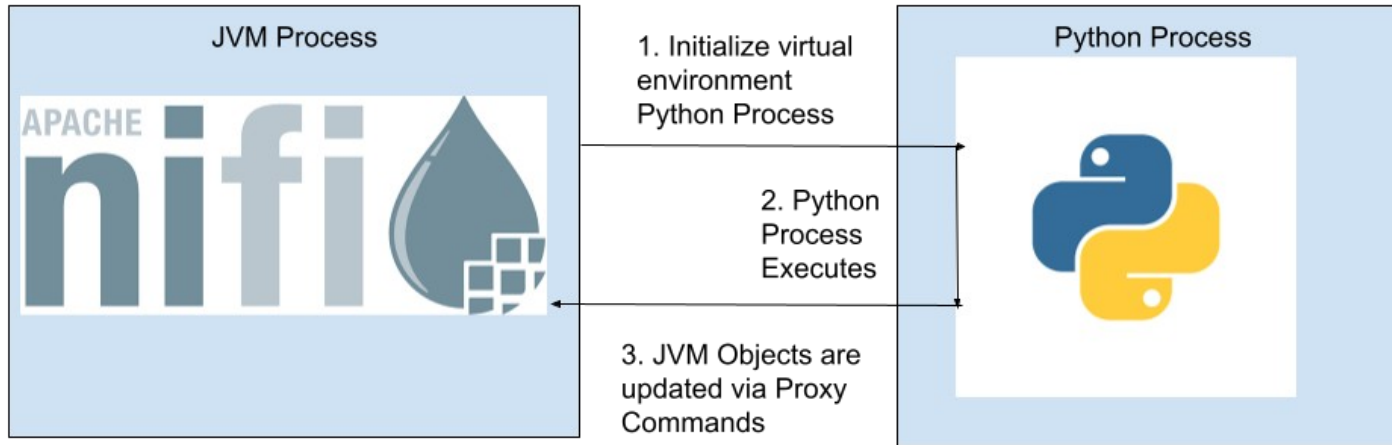
How Does Py4J Work in NiFi?

- 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?



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://arxiv.org/abs/2012.04268)".

Table Detection

Figure 1: A screenshot of a PDF document showing a table. The table is highlighted with a red dashed box. The table has 5 columns and 10 rows. The first row is the header row. The table is titled "Table 1: Prevalence of dental hard tissue anomalies in children with developmental dental anomalies (n=100)".

Table 1: Prevalence of dental hard tissue anomalies in children with developmental dental anomalies (n=100)

| Category | Prevalence (%) | 95% CI |
|----------------------------|----------------|-------------|
| Good oral hygiene status | 1.00 | 0.00 - 1.00 |
| Fair oral hygiene status | 0.02 | 0.00 - 0.05 |
| Poor oral hygiene status | 0.00 | 0.00 - 0.00 |
| Gender | | |
| Male | 1.00 | 0.00 - 1.00 |
| Female | 0.00 | 0.00 - 0.00 |
| Socioeconomic status | | |
| High socioeconomic class | 1.00 | 0.00 - 1.00 |
| Middle socioeconomic class | 0.00 | 0.00 - 0.00 |
| Low socioeconomic class | 0.00 | 0.00 - 0.00 |

and even lower than the rates prevalence in many other developing and developed countries. The risk and gender are factors for the rates in the study environment are also well understood [20]. This study provides evidence that the presence of developmental dental hard tissue anomalies does not increase the probability of children having caries in the study population.

(Of importance) is the significant association between developmental dental hard tissue anomalies and poor oral hygiene. The presence of dental hard tissue anomalies increases difficulty in tooth cleaning [21], also increases malocclusion, which also increases the risk for plaque retention and poor oral hygiene [22, 23]. The finding in this study is therefore consistent with prior observations [24, 25], and has programmatic implications for managing adolescents. Adolescents with developmental dental hard tissue anomalies should be treated as having high risk for poor oral hygiene and should therefore be monitored more frequently for dental visits with particular emphasis on brushing, flossing and eating habits.

The study found a non-significant association between caries and presence of enamel hypoplasia, unlike the findings of some previous studies [26-28]. While Vargas-Fernandez et al. [26] meta-analysis strongly indicates that developmental defects of the enamel are caused by pathogen in a risk factor for caries, this study finding indicates that enamel hypoplasia is not a risk factor for caries in the study population. There is a need for more research in countries where the caries prevalence and severity is low [29]. However, the non-significant association between developmental dental hard tissue anomalies and caries

and the significant association between developmental dental hard tissue anomalies and poor oral hygiene may highlight the probable pathophysiology of caries associated with developmental dental hard tissue anomalies caries results as a secondary outcome of poor hygiene and not through a direct pathway. This population would need further studies, as there are multiple other related factors that may increase the susceptibility of tooth with developmental dental hard tissue anomalies to caries.

The study finding on gender and socioeconomic class differences in the prevalence of enamel hypoplasia differed from the findings of Bollen et al. [30] in Spain who showed increased prevalence increased prevalence of developmental defects of the enamel (includes of enamel hypoplasia) in males and in children from middle and low socioeconomic status. The increasing risk for developmental defects of the enamel with decreasing socioeconomic status had been established, with risk association linked to poor nutritional status [31]. However, the differences in the prevalence of developmental defects of the enamel by gender remains unclear with others showing an association with a greater risk [32, 33], while others show no association [34, 35]. Also, these studies assessed enamel defects, regardless of whether it was enamel or hypoplasia.

This study was a school based study implying that children in full-time school, who do not attend school have been left out of this survey as reports state that a high proportion of children in higher age are not in school. This may have generalized the results of the study. However, within the limits of the design of the study, the data still provide useful information highlighting the prevalence of developmental dental hard tissue

Table Structure Recognition

Column

| Variables | Adjusted Prevalence Ratio (APR) | 95% CI | P-value | 95% CI (P-value) |
|----------------------------|---------------------------------|--------|---------|------------------|
| Good oral hygiene status | 1.00 | | | |
| Fair oral hygiene status | 0.02 | 0.00 | 0.04 | <0.001 - 0.05 |
| Poor oral hygiene status | 0.00 | 0.00 | 0.00 | 0.00 - 0.00 |
| Gender status | | | | |
| Male | 1.00 | | | |
| Female | 0.00 | 0.00 | 0.00 | <0.001 - 0.00 |
| Socioeconomic status | | | | |
| High socioeconomic class | 1.00 | | | |
| Middle socioeconomic class | <0.001 | 0.00 | 0.00 | <0.001 - 0.00 |
| Low socioeconomic class | <0.001 | 0.00 | 0.00 | <0.001 - 0.00 |

Row

| Variables | Adjusted Prevalence Ratio (APR) | 95% CI | P-value | 95% CI (P-value) |
|----------------------------|---------------------------------|--------|---------|------------------|
| Good oral hygiene status | 1.00 | | | |
| Fair oral hygiene status | 0.02 | 0.00 | 0.04 | <0.001 - 0.05 |
| Poor oral hygiene status | 0.00 | 0.00 | 0.00 | 0.00 - 0.00 |
| Gender status | | | | |
| Male | 1.00 | | | |
| Female | 0.00 | 0.00 | 0.00 | <0.001 - 0.00 |
| Socioeconomic status | | | | |
| High socioeconomic class | 1.00 | | | |
| Middle socioeconomic class | <0.001 | 0.00 | 0.00 | <0.001 - 0.00 |
| Low socioeconomic class | <0.001 | 0.00 | 0.00 | <0.001 - 0.00 |

Text Cell

| Variables | Adjusted Prevalence Ratio (APR) | 95% CI | P-value | 95% CI (P-value) |
|----------------------------|---------------------------------|--------|---------|------------------|
| Good oral hygiene status | 1.00 | | | |
| Fair oral hygiene status | 0.02 | 0.00 | 0.04 | <0.001 - 0.05 |
| Poor oral hygiene status | 0.00 | 0.00 | 0.00 | 0.00 - 0.00 |
| Gender status | | | | |
| Male | 1.00 | | | |
| Female | 0.00 | 0.00 | 0.00 | <0.001 - 0.00 |
| Socioeconomic status | | | | |
| High socioeconomic class | 1.00 | | | |
| Middle socioeconomic class | <0.001 | 0.00 | 0.00 | <0.001 - 0.00 |
| Low socioeconomic class | <0.001 | 0.00 | 0.00 | <0.001 - 0.00 |

Spanning Cell

Grid Cell

Table Functional Analysis

Column Header Cell

| Variables | Adjusted Prevalence Ratio (APR) | 95% CI | P-value | 95% CI (P-value) |
|----------------------------|---------------------------------|--------|---------|------------------|
| Good oral hygiene status | 1.00 | | | |
| Fair oral hygiene status | 0.02 | 0.00 | 0.04 | <0.001 - 0.05 |
| Poor oral hygiene status | 0.00 | 0.00 | 0.00 | 0.00 - 0.00 |
| Gender status | | | | |
| Male | 1.00 | | | |
| Female | 0.00 | 0.00 | 0.00 | <0.001 - 0.00 |
| Socioeconomic status | | | | |
| High socioeconomic class | 1.00 | | | |
| Middle socioeconomic class | <0.001 | 0.00 | 0.00 | <0.001 - 0.00 |
| Low socioeconomic class | <0.001 | 0.00 | 0.00 | <0.001 - 0.00 |

Projected Row Header Cell

| Variables | Adjusted Prevalence Ratio (APR) | 95% CI | P-value | 95% CI (P-value) |
|----------------------------|---------------------------------|--------|---------|------------------|
| Good oral hygiene status | 1.00 | | | |
| Fair oral hygiene status | 0.02 | 0.00 | 0.04 | <0.001 - 0.05 |
| Poor oral hygiene status | 0.00 | 0.00 | 0.00 | 0.00 - 0.00 |
| Gender status | | | | |
| Male | 1.00 | | | |
| Female | 0.00 | 0.00 | 0.00 | <0.001 - 0.00 |
| Socioeconomic status | | | | |
| High socioeconomic class | 1.00 | | | |
| Middle socioeconomic class | <0.001 | 0.00 | 0.00 | <0.001 - 0.00 |
| Low socioeconomic class | <0.001 | 0.00 | 0.00 | <0.001 - 0.00 |

<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 |
| 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 |

```
{
  "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"
    }
  ]
}
```

Building a Nifi Python Plugin



Dependencies



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



Building a NiFi Python Plugin

Processor Properties

Processor Details

TableDetectionProcessor 0.0.1-SNAPSHOT

Settings

Scheduling

Properties

Relationships

Comments

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

Settings

Scheduling

Properties

Relationships

Comments

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?



Thank You!

Code



Bob Paulin
@bobbpaulin



<https://datavolo.io>

Me

