COMMUNITY
THE ASF CONFERENCE
CODE

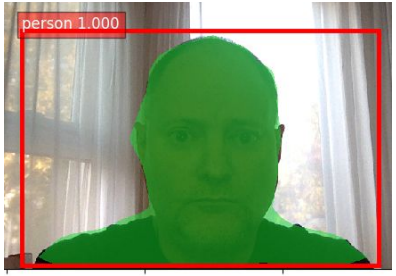
Enhancing Apache NiFi 2.x with Python Processors

Tim Spann, Senior Solutions Engineer, Snowflake

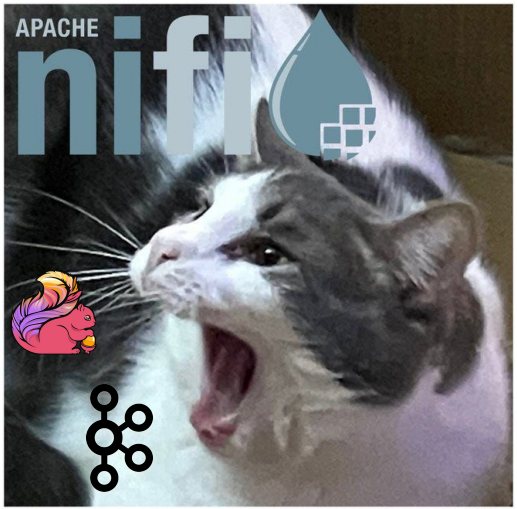
Tim Spann

paasdev.bsky.social
@PaasDev // Blog: **datainmotion.dev**
Senior Solutions Engineer, Snowflake
NY/NJ/Philly - Data + AI Meetups
ex-Zilliz, ex-Pivotal, ex-Cloudera, ex-HPE, ex-StreamNative, ex-EY, ex-Hortonworks.

https://medium.com/@tspann
https://github.com/tspannhw



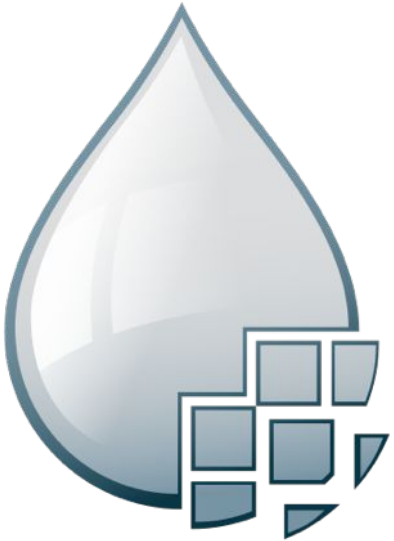
AI + Data Weekly by Tim Spann



https://bit.ly/32dAJft

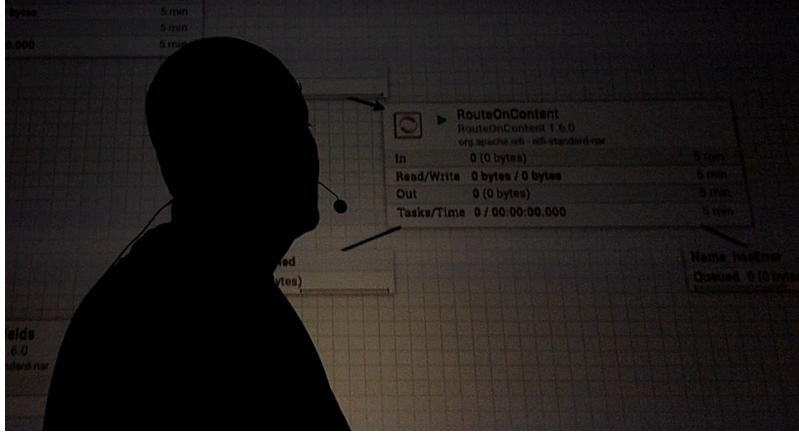
This week in Apache NiFi, Apache Polaris, Apache Flink, Apache Kafka, ML, AI, Streamlit, Jupyter, Apache Iceberg, Python, Java, LLM, GenAI, Snowflake, Unstructured Data and Open Source friends.

Apache NiFi for Data Ingest, Movement and Routing



- Guaranteed delivery
- Data buffering
  - Backpressure
  - Pressure release
- Prioritized queuing
- Flow specific QoS
  - Latency vs. throughput
  - Loss tolerance
- Data provenance
- Supports push and pull models
- Hundreds of processors
- Visual command and control
- Hundreds of sources
- Flow templates
- Pluggable/multi-role security
- Designed for extension
- Clustering
- Version Control

APACHE NIFI 2.5 FEATURES



Major Updates:

- Python Integration
- Parameterization
- JDK 21+
- Provenance / Data Lineage
- Amazon Glue Schema Reference Reader
- Additional Azure Processors
- Integration with Zendesk, Slack,
- Database Tables as Schemas
- Amazon Glue Schema Registry
- Open Telemetry Support

Unstructured Data

- **Archives** - tar, gzipped, zipped, …
- **Images** - PNG, JPG, GIF, BMP, …
- **Documents** - HTML, Markdown, RSS, PDF, Doc, RTF, Plain Text, …
- **Videos** - MP4, Clips, Mov, Youtube URL…
- **Sound** - MP3, …
- **Social / Chat** - Slack, Discord, Twitter, REST, Email, …
- **Identify Mime Types, Chunk Documents, Store to Vector Database**
- **Parse Documents -** HTML, Markdown, PDF, Word, Excel, Powerpoint

Simple Processor

```
from nifiapi.flowfiletransform import FlowFileTransform,
FlowFileTransformResult


class MyProcessorIsCool(FlowFileTransform):
    class Java:
        implements =
['org.apache.nifi.python.processor.FlowFileTransform']
    class ProcessorDetails:
        version = '1.0.0'
        dependencies = ['feedparser']
        Description = 'Description of processor'
```

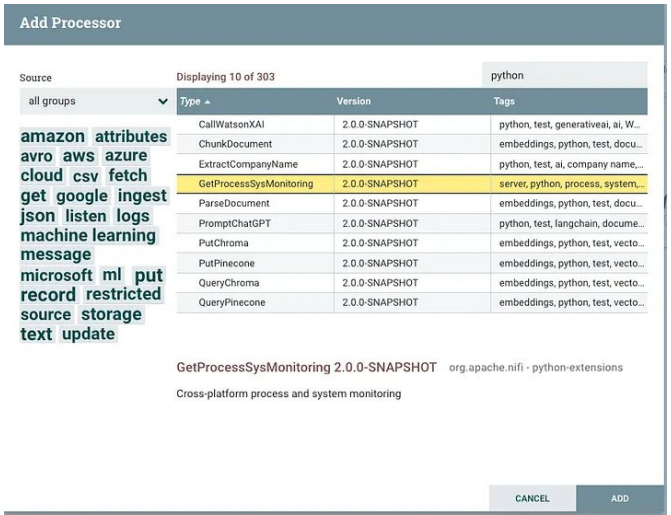
Simple Processor

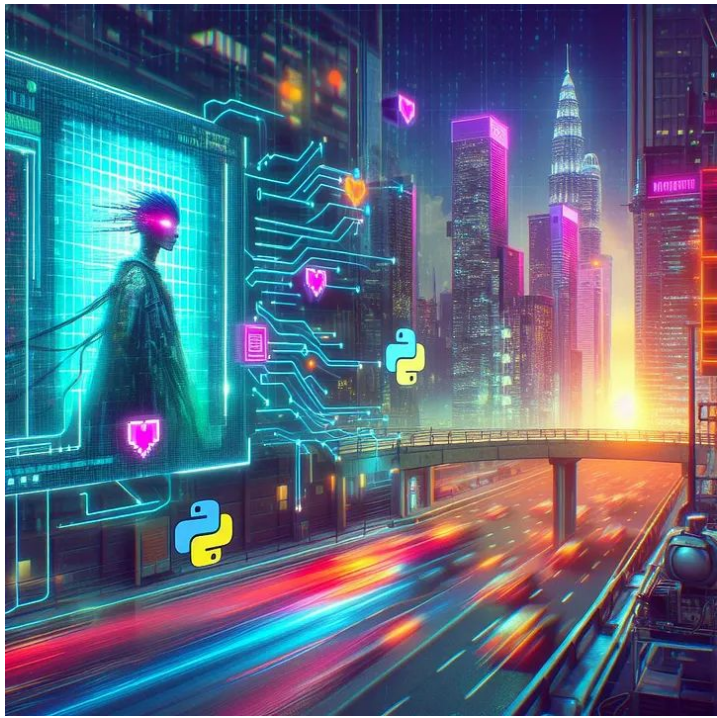
```
  def __init__(self, **kwargs):
        pass


    def transform(self, context, flowfile):
        return FlowFileTransformResult(relationship = "success",
contents = "Hello World", attributes = {"greeting": "hello"})
```

Custom Processor Examples

- Take existing Python code or libraries
- Wrap in a Python processor
- Build
- Deploy
- Test





RSS to CSV

- Python 3.10+
- Feedparser, pandas
- Reads RSS, outputs CSV file
- Augmented by cursor ai
- Example URL: https://travel.state.gov/_res/rss/TAsTWs.xml

Attribute Values

Show modified attributes only

filename
rss_feed_travel.state.gov:_Travel_Advisories.csv
ab5cb6f1-4fa1-495d-9eca-a6583b1d1559

mime.type
text/csv
No value set

rss.feed.description
As a first step in planning any trip abroad, check the Travel Advisories for your intended destination.
No value set

rss.feed.title
travel.state.gov: Travel Advisories
No value set

rss.item.count
211
No value set

Provenance

Oldest event available: 09/02/2025 17:46:20 EDT

Filter | Filter By: component name

Filter matched 7 of 7

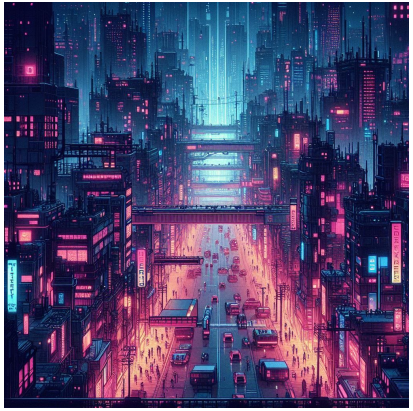
Event Time ↓	Type	FlowFile UUID	File Size	Component Name
09/05/2025 13:43:47.344 EDT	CONTENT_MODIFIED	2537c0eb-e678-43fe-b85a-2b08f6829abb	700.54 KB	RSSToCSV

https://github.com/tspannhw/Example-RSSToCSV

CaptionImage

- Python 3.10+
- Hugging Face
- Salesforce/blip-image-captioning-large
- Generate Captions for Images
- Adds captions to FlowFile Attributes
- Does not require download or copies of your images

https://github.com/tspannhw/FLaNK-python-processors



RESNetImageClassification

- Python 3.10+
- Hugging Face
- Transformers
- Pytorch
- Datasets
- microsoft/resnet-50
- Adds classification label to FlowFile Attributes
- Does not require download or copies of your images

https://github.com/tspannhw/FLaNK-python-processors

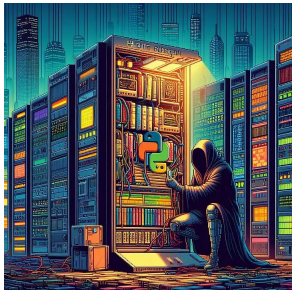
NSFWImageDetection

- Python 3.10+
- Hugging Face
- Transformers
- Falconsai/nsfw_image_detection
- Adds normal and nsfw to FlowFile Attributes
- Gives score on safety of image
- Does not require download or copies of your images

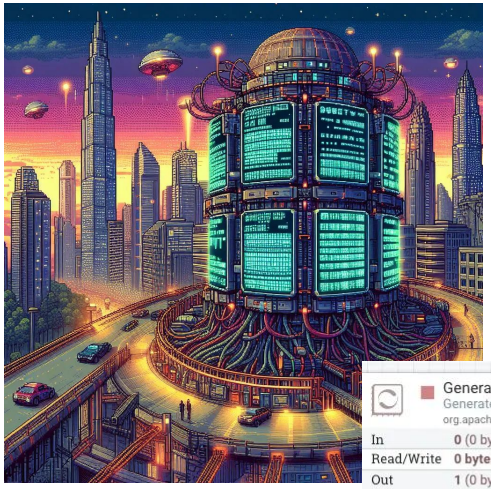
https://github.com/tspannhw/FLaNK-python-processors

FacialEmotionsImageDetection

- Python 3.10+
- Hugging Face
- Transformers
- facial_emotions_image_detection
- Image Classification
- Adds labels/scores to FlowFile Attributes
- Does not require download or copies of your images

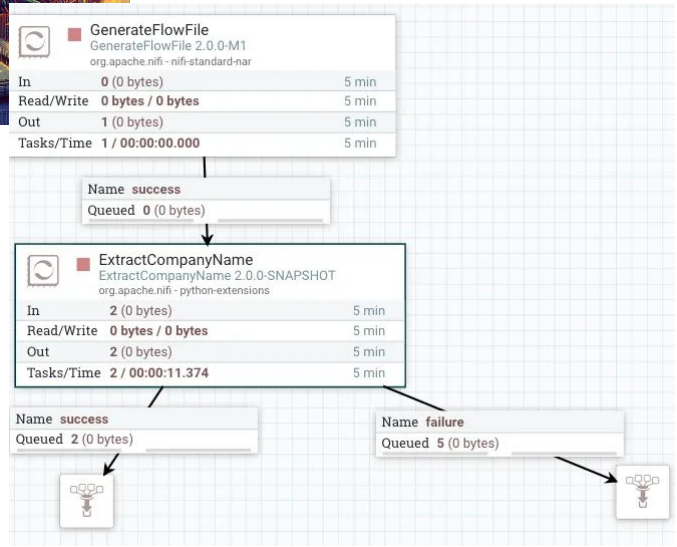


https://github.com/tspannhw/FLaNK-python-processors



Extract Company Names

- Python 3.10+
- Hugging Face, NLP, SpaCY, PyTorch



https://github.com/tspannhw/FLaNK-python-ExtractCompanyName-processor

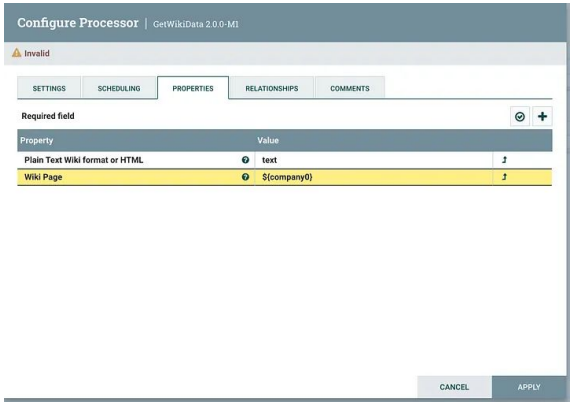
Extract Entities

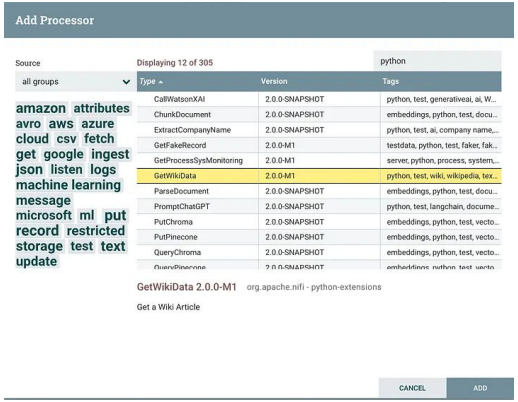
- Python 3.10+
- SpaCY
- GPE, ORG, DATE, PERSON, LOC, MONEY
- TIME, PRODUCT, QUANTITY, EVENT, FAC
- Extract Natural Language Entities

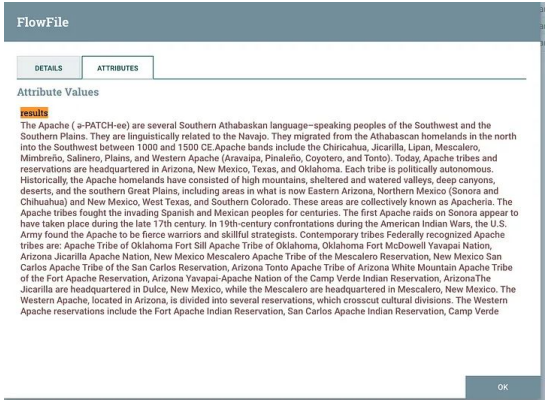
https://github.com/tspannhw/FLaNK-python-processors/blob/main/ExtractEntities.py

Get Wiki Data

- Python 3.10+
- Wikipediaapi
- Text or HTML output
- Input: Wikipage Name

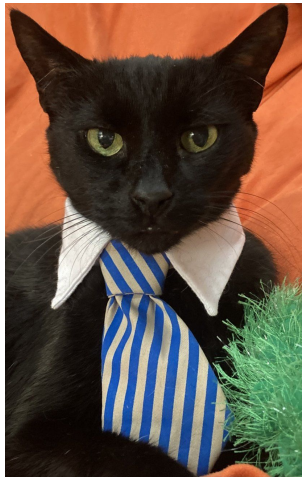






https://github.com/tspannhw/FLaNK-python-processors/blob/main/GetWikiData.py

Address To Lat/Long

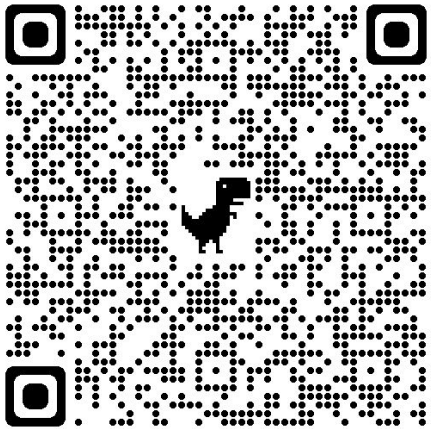


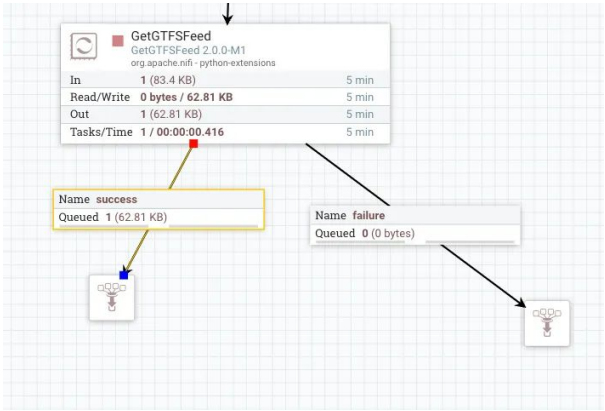
- Python 3.10+
- geopy Library
- Nominatim
- OpenStreetMaps (OSM)
- openstreetmap.org/copyright
- Returns as attributes and JSON file
- Works with partial addresses
- Categorizes location
- Bounding Box

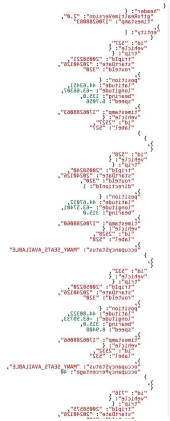
https://github.com/tspannhw/FLaNKAI-Boston

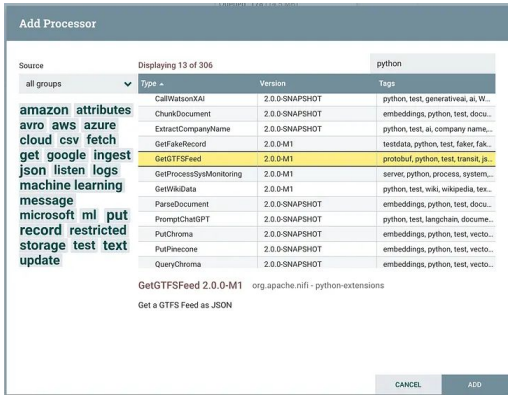
Get GTFS Data

- Python 3.10+
- GTFS from Transit URL
- Alerts, Trip Updates or Vehicle Positions
- Returns JSON
- google.transit and google.protobuf



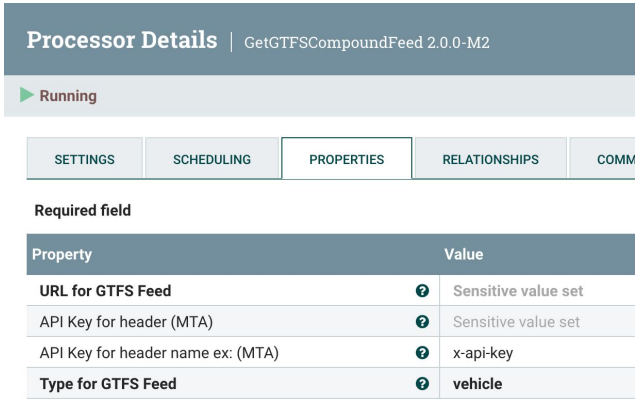






Get Compound GTFS Data

- Python 3.10+
- GTFS to JSON



https://github.com/tspannhw/FLaNK-python-processors/blob/main/GetGTFSCompoundFeed.py

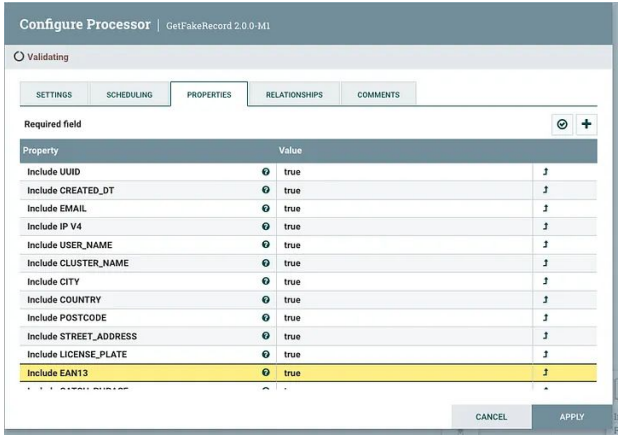
Get Process Sys Monitoring

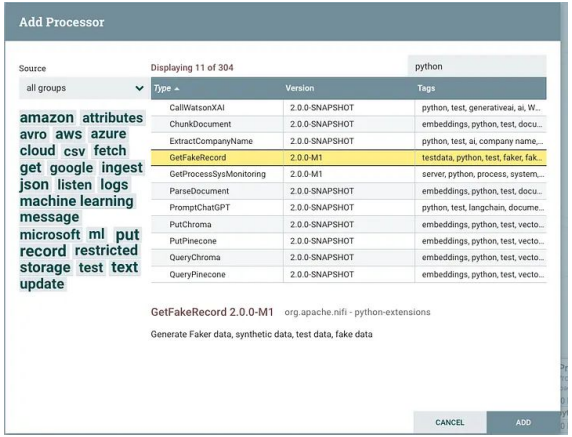
- Python 3.10+
- Local machine information
- psutil
- Users
- CPU, Swap Memory, Memory and Disk usage
- Network information

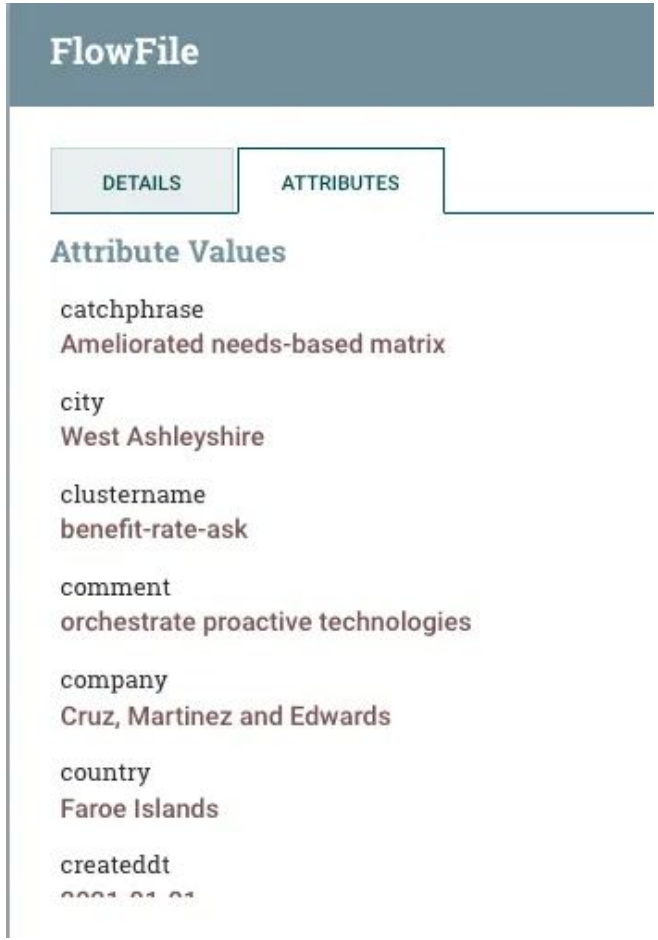
https://github.com/tspannhw/FLaNK-python-processors/blob/main/GetProcessSysMonitoring.py

Get Fake Record

- Python 3.10+
- Faker
- User Agent, Lat/Lng
- Phone, Address







https://github.com/tspannhw/FLaNK-python-processors/blob/main/GetFakeRecord.py

LET'S SEE

RESOURCES AND WRAP-UP

REFERENCES

https://nifi.apache.org/nifi-docs/python-developer-guide.html

https://medium.com/@tspann/automating-transit-gtfs-flows-7a5c4d3dafe8

https://medium.com/@tspann/populating-an-open-lakehouse-with-codeless-data-streams-04292375ddaf

https://medium.com/@tspann/real-time-enrichment-of-air-quality-data-3ce670e4fc5b

https://medium.com/@tspann/yes-apache-nifi-can-do-that-c7fcaca5a177

https://medium.com/cloudera-inc/real-time-in-boston-part-1-0f92d7da3496

https://medium.com/cloudera-inc/boston-wheres-my-bus-llm-streaming-to-the-rescue-586dfd019237

https://medium.com/@tspann/real-time-irish-transit-analytics-ea76164c9595

https://medium.com/cloudera-inc/streaming-street-cams-to-yolo-v8-with-python-and-nifi-to-minio-s3-3277e73723ce

https://medium.com/cloudera-inc/nyc-traffic-are-you-kidding-me-6d3fa853903b

REFERENCES

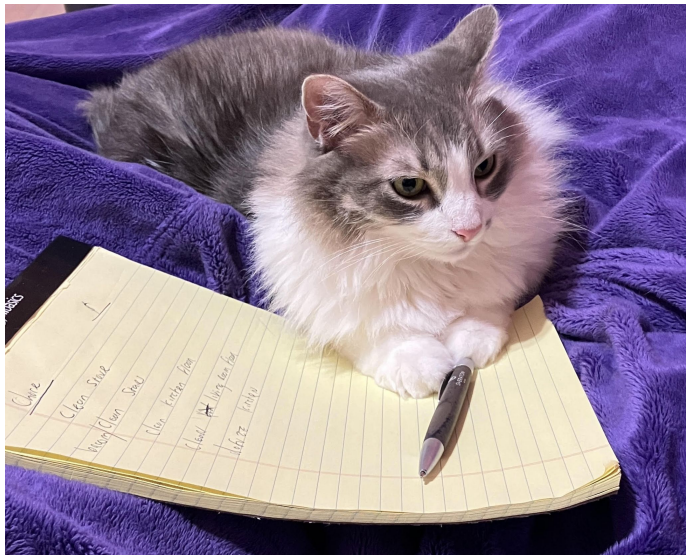
https://medium.com/cloudera-inc/searching-slack-from-apache-nifi-9ed562aa2397

https://medium.com/@tspann/yet-another-python-processor-45aaae6fe406

https://medium.com/cloudera-inc/building-a-library-of-python-processors-6b5517404a58

https://github.com/tspannhw/FLaNK-python-processors

https://github.com/tspannhw/CortexAISearchForAirQuality




SPIRIT
Business
Cat
ADULT
Size Costume
ONE SIZE FITS MOST


AUTOMATE EVERYTHING
imgflip.com