



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>



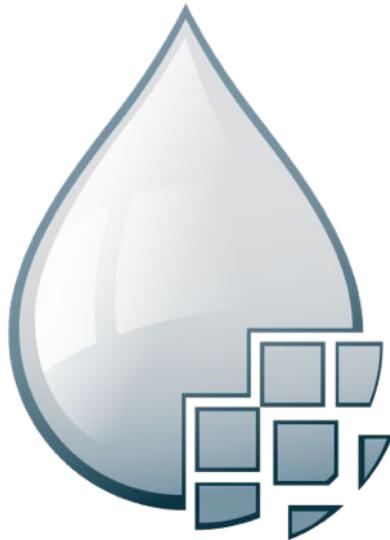
AI + Data Weekly by Tim Spann



<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

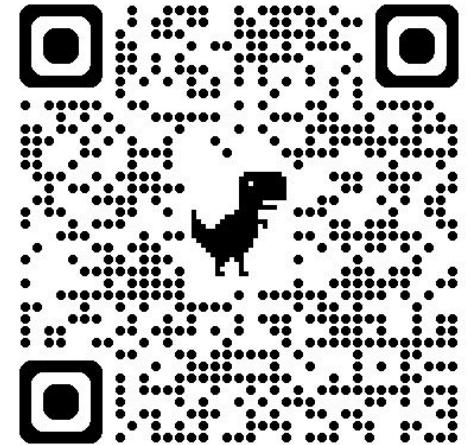
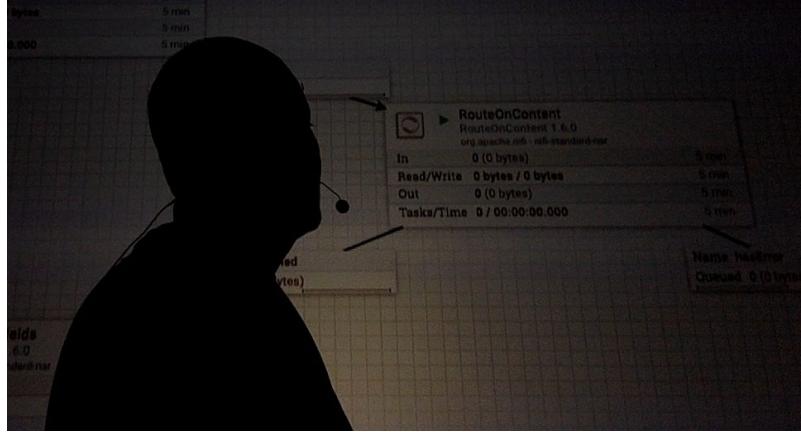


- Guaranteed delivery
- Data buffering
 - Backpressure
 - Pressure release
- Prioritized queuing
- Flow specific QoS
 - Latency vs. throughput
 - Loss tolerance
- Data provenance
- Supports push and pull models
- Hundreds of processors
- Visual command and control
- Hundreds of sources
- Flow templates
- 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

Add Processor

Source	Type	Version	Tags
all groups	Type	2.0.0-SNAPSHOT	python, test, generativeai, ai, W...
amazon attributes	CallWatsonXAI	2.0.0-SNAPSHOT	python, test, generativeai, ai, W...
avro aws azure	ChunkDocument	2.0.0-SNAPSHOT	embeddings, python, test, docu...
cloud csv fetch	ExtractCompanyName	2.0.0-SNAPSHOT	python, test, ai, company name,...
get google ingest	GetProcessSysMonitoring	2.0.0-SNAPSHOT	server, python, process, system...
json listen logs	ParseDocument	2.0.0-SNAPSHOT	embeddings, python, test, docu...
machine learning	PromptChatGPT	2.0.0-SNAPSHOT	python, test, langchain, docume...
message	PutChroma	2.0.0-SNAPSHOT	embeddings, python, test, vecto...
microsoft ml put	PutPinecone	2.0.0-SNAPSHOT	embeddings, python, test, vecto...
record restricted	QueryChroma	2.0.0-SNAPSHOT	embeddings, python, test, vecto...
source storage	QueryPinecone	2.0.0-SNAPSHOT	embeddings, python, test, vecto...
text update			

Displaying 10 of 303

python

GetProcessSysMonitoring 2.0.0-SNAPSHOT org.apache.nifi - python-extensions
Cross-platform process and system monitoring

CANCEL ADD





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

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

Filter matched 7 of 7

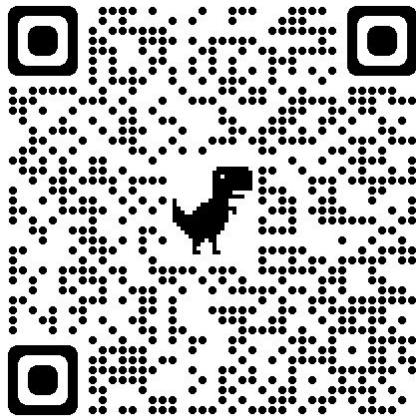
Event Time ↓	Type	FlowFile UUID	File Size	Component Name
09/05/2025 13:43:47.344 EDT	CONTENT_MODIFIED	2537c0eb-e678-43fe-b85a-2b08f6829abb	700.54 KB	RSSToCSV

<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



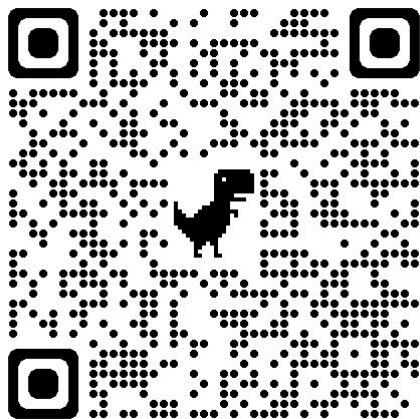
<https://github.com/tspannhw/FLaNK-python-processors>





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

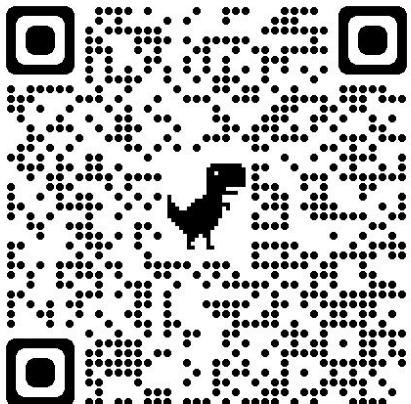


<https://github.com/tspannhw/FLaNK-python-processors>



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

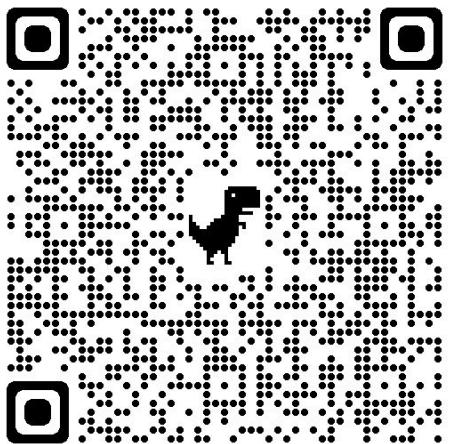


<https://github.com/tspannhw/FLaNK-python-processors>



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



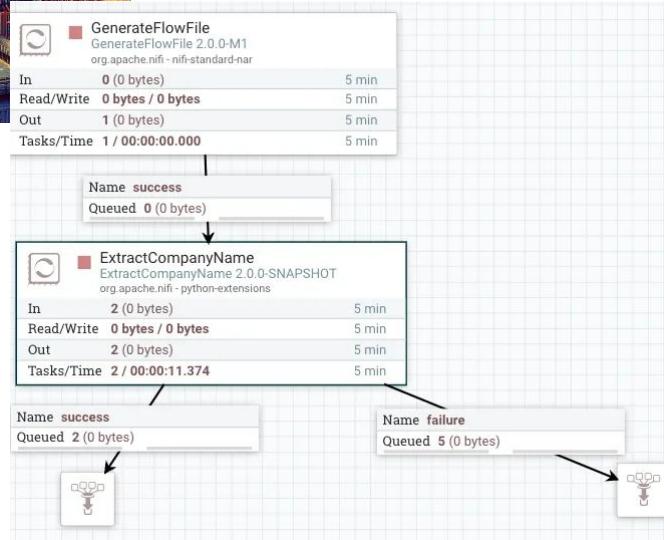
<https://github.com/tspannhw/FLaNK-python-processors>





Extract Company Names

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



Attribute Values

companylist

[**"Amazon", "Microsoft", "Cloudera", "DataSQLR", "Google", "IBM"**]

filename

36fb4ae6-701a-4e1d-b890-c93b44f2200b

parsedcompany

Amazon

path

./

uuid

6366a2c9-3dd4-4e8f-8825-83189d403b92



<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

Configure Processor | GetWikiData 2.0.0-M1

⚠ Invalid

✖ +

SETTINGS **SCHEDULING** **PROPERTIES** **RELATIONSHIPS** **COMMENTS**

Required field

Property	Value
Plain Text Wiki format or HTML	text
Wiki Page	\${{company}}

Source Displaying 12 of 305

Type python

Tags

all groups

Processor Name	Version	Description
CallWatsonXAI	2.0.0-SNAPSHOT	python, test, generative, ai, W...
ChunkDocument	2.0.0-SNAPSHOT	embeddings, python, test, docu...
ExtractCompanyName	2.0.0-SNAPSHOT	python, test, ai, company name...
GetFakeRecord	2.0.0-M1	testdata, python, test, faker, f...
GetProcessSysMonitoring	2.0.0-M1	serve, python, process, system...
GetWikiData	2.0.0-M1	python, test, wiki, wikipedia, tex...
GetDocument	2.0.0-SNAPSHOT	embeddings, python, test, docu...
PromptChatGPT	2.0.0-SNAPSHOT	python, test, langchain, docume...
PutChroma	2.0.0-SNAPSHOT	embeddings, python, test, vecto...
PutInference	2.0.0-SNAPSHOT	embeddings, python, test, vecto...
QueryChroma	2.0.0-SNAPSHOT	embeddings, python, test, vecto...
QueryInference	2.0.0-SNAPSHOT	embeddings, python, test, vecto...

GetWikiData 2.0.0-M1 org.apache.nifi - python-extensions

Get a Wiki Article

CANCEL APPLY

Add Processor

Source Displaying 12 of 305

Type python

Tags

all groups

Processor Name	Version	Description
CallWatsonXAI	2.0.0-SNAPSHOT	python, test, generative, ai, W...
ChunkDocument	2.0.0-SNAPSHOT	embeddings, python, test, docu...
ExtractCompanyName	2.0.0-SNAPSHOT	python, test, ai, company name...
GetFakeRecord	2.0.0-M1	testdata, python, test, faker, f...
GetProcessSysMonitoring	2.0.0-M1	serve, python, process, system...
GetWikiData	2.0.0-M1	python, test, wiki, wikipedia, tex...
GetDocument	2.0.0-SNAPSHOT	embeddings, python, test, docu...
PromptChatGPT	2.0.0-SNAPSHOT	python, test, langchain, docume...
PutChroma	2.0.0-SNAPSHOT	embeddings, python, test, vecto...
PutInference	2.0.0-SNAPSHOT	embeddings, python, test, vecto...
QueryChroma	2.0.0-SNAPSHOT	embeddings, python, test, vecto...
QueryInference	2.0.0-SNAPSHOT	embeddings, python, test, vecto...

GetWikiData 2.0.0-M1 org.apache.nifi - python-extensions

Get a Wiki Article

CANCEL ADD

FlowFile

DETAILS ATTRIBUTES

Attribute Values

Results

The Apache (ə-PÄTCH-ee) are several Southern Athabaskan language-speaking peoples of the Southwest and the Southern Plains. They are linguistically related to the Navajo. They migrated from the Athabaskan homelands in the north into the Southwest between 1000 and 1500 CE. Apache bands include the Chiricahua, Jicarilla, Lipan, Mescalero, Mimbreño, Sáñíne, Plains, and Western Apache (Aravaipa, Piñonero, Coyotero, and Tonto). Today, Apache tribes and reservations are located in Arizona, New Mexico, Texas, and Oklahoma. Historically, the Apache homelands have consisted of high mountains, sheltered arid waterways, deep canyons, deserts, and the southern Great Plains, including areas in what is now Eastern Arizona, Northern Mexico (Sonora and Chihuahua) and New Mexico, West Texas, and Southern Colorado. These areas are collectively known as Apacheria. The Apache tribes fought the invading Spanish and Mexican peoples for centuries. The first Apache raids on Sonora appear to have taken place in the 1600s. In the 1700s, they participated in the Comanche Wars, and in the 1800s, the U.S. Army forced the Apache to be fierce warriors and skillful strategists. Contemporary tribes Federally recognized Apache tribes are: Apache Tribe of Oklahoma Fort Sill Apache Tribe of Oklahoma, Oklahoma Fort McDowell Yavapai Nation, Arizona Jicarilla Apache Nation, New Mexico Mescalero Apache Tribe of the Mescalero Reservation, New Mexico San Carlos Apache Tribe of the San Carlos Reservation, Arizona Tohono O'odham Nation, and the Mountain Apache Tribe of the Fort Verde Reservation, Arizona Yavapai-Apache Nation of the Coconino Verde Indian Reservation, Arizona. The Jicarilla are headquartered in Dulce, New Mexico, while the Mescalero are headquartered in Mescalero, New Mexico. The Western Apache, located in Arizona, is divided into several reservations, which crosscut cultural divisions. The Western Apache reservations include the Fort Apache Indian Reservation, San Carlos Apache Indian Reservation, Camp Verde

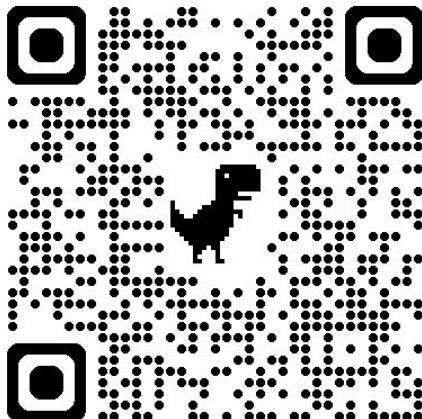
OK

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



Address To Lat/Long

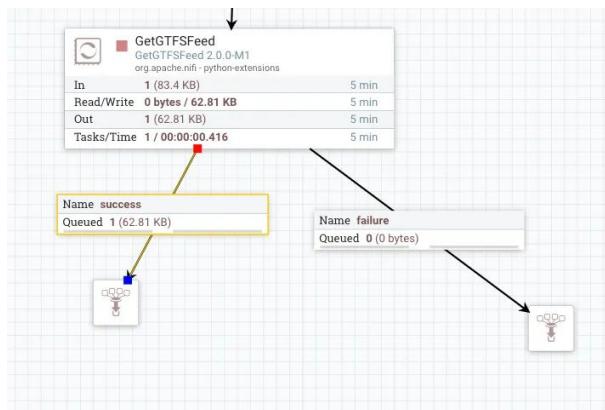
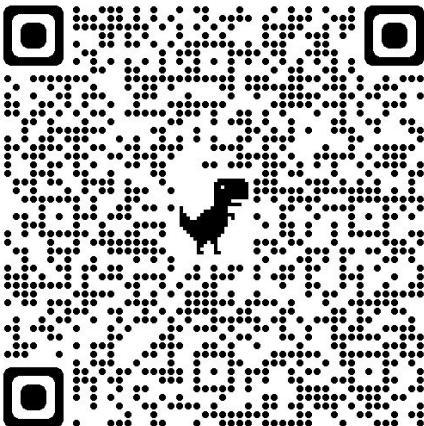
- Python 3.10+
- geopy Library
- Nominatim
- OpenStreetMaps (OSM)
- openstreetmap.org/copyright
- Returns as attributes and JSON file
- Works with partial addresses
- Categorizes location
- Bounding Box





Get GTFS Data

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



Add Processor				
Source	Displaying 13 of 306			python
all groups	Type	Version	Tags	
amazon attributes	CallWatsonXAI	2.0.0-SNAPSHOT	python, test, generativeai, ai, W...	
avro aws azure	ChunkDocument	2.0.0-SNAPSHOT	embeddings, python, test, docu...	
cloud csv fetch	ExtractCompanyName	2.0.0-SNAPSHOT	python, test, ai, company name...	
get google ingest	GetFileRecord	2.0.0-M1	testdata, python, test, faker, fak...	
json listen logs	GetGTFSFeed	2.0.0-M1	protobuf, python, test, transit, js...	
machine learning	GetProcessSystemMonitoring	2.0.0-M1	server, python, process, system...	
message	GetWikiData	2.0.0-M1	python, test, wikitext, wikipedia, tex...	
microsoft ml put	ParseDocument	2.0.0-SNAPSHOT	embeddings, python, test, docu...	
restricted storage test	PromptChatGPT	2.0.0-SNAPSHOT	python, test, langchain, document...	
update	PutChroma	2.0.0-SNAPSHOT	embeddings, python, test, vector...	
	PutPinecone	2.0.0-SNAPSHOT	embeddings, python, test, vecto...	
	QueryChroma	2.0.0-SNAPSHOT	embeddings, python, test, vector...	



Get Compound GTFS Data

- Python 3.10+
- GTFS to JSON



trip_update

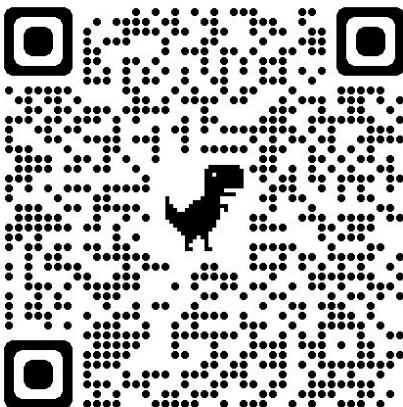
trip_update

vehicle

vehicle

alert

Reference parameter...



Processor Details | GetGTFSCompoundFeed 2.0.0-M2

Running

SETTINGS SCHEDULING PROPERTIES RELATIONSHIPS COMM

Required field

Property	Value
URL for GTFS Feed	Sensitive value set
API Key for header (MTA)	Sensitive value set
API Key for header name ex: (MTA)	x-api-key
Type for GTFS Feed	vehicle

<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



Get Fake Record

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

Configure Processor | GetFakeRecord 2.0.0-M1

○ Validating

SETTINGS SCHEDULING PROPERTIES RELATIONSHIPS COMMENTS

Required field

Property	Value
Include UUID	✓ true
Include CREATED_DT	✓ true
Include EMAIL	✓ true
Include IP_V4	✓ true
Include USER_NAME	✓ true
Include CLUSTER_NAME	✓ true
Include CITY	✓ true
INCLUDE COUNTRY	✓ true
Include POSTCODE	✓ true
Include STREET_ADDRESS	✓ true
Include LICENSE_PLATE	✓ true
Include EAN13	✓ true

CANCEL APPLY

Add Processor

Source: all groups Displaying 11 of 304 Type: python

Source	Type	Version	Tags
amazon attributes	CallWatsonXAI	2.0.0-SNAPSHOT	python, test, generativeai, ai, W...
avro aws azure	ChunkDocument	2.0.0-SNAPSHOT	embeddings, python, test, docu...
cloud csv fetch	ExtractCompanyName	2.0.0-SNAPSHOT	python, test, ai, company name...
get google ingest	GetFakeRecord	2.0.0-M1	testdata, python, test, faker, fak...
json listen logs	GetProcessSystemMonitoring	2.0.0-M1	server, python, process, system...
machine learning	ParseDocument	2.0.0-SNAPSHOT	embeddings, python, test, docu...
message	PromptChatGPT	2.0.0-SNAPSHOT	python, test, langchain, docume...
microsoft ml put	PutChroma	2.0.0-SNAPSHOT	embeddings, python, test, vecto...
record restricted	PutPinecone	2.0.0-SNAPSHOT	embeddings, python, test, vecto...
storage test text	QueryChroma	2.0.0-SNAPSHOT	embeddings, python, test, vecto...
update	QueryPinecone	2.0.0-SNAPSHOT	embeddings, python, test, vecto...

GetFakeRecord 2.0.0-M1 org.apache.nifi - python-extensions
Generate Faker data, synthetic data, test data, fake data

CANCEL ADD

FlowFile

DETAILS ATTRIBUTES

Attribute Values

catchphrase
Ameliorated needs-based matrix

city
West Ashleyshire

clustername
benefit-rate-ask

comment
orchestrate proactive technologies

company
Cruz, Martinez and Edwards

country
Faroe Islands

createddt
2021-01-01

**LET'S
SEE**

IS THIS ENOUGH DATA?





imgflip.com **TIME TO REBOOT THE CAT**

RESOURCES AND WRAP-UP

REFERENCES

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

<https://medium.com/@tspann/automating-transit-gtfs-flows-7a5c4d3afe8>

<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-mini-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-45aaaae6fe406>

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

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

<https://github.com/tspannhw/CortexAISeachForAirQuality>

